

APR 2 1 2003



Public Library San Mateo, CA

The New Encyclopædia Britannica

Volume 25

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15TH EDITION

Encyclopædia Britannica, Inc.
Jacob E. Safra, Chairman of the Board
Ilan Yeshua, Chief Executive Officer

Chicago
London/New Delhi/Paris/Seoul
Sydney/Taipei/Tokyo

Public Library San Mateo, CA

First Edition 1768–1771
Second Edition 1777–1784
Third Edition 1788–1797
Supplement 1801
Fourth Edition 1801–1809
Fifth Edition 1815
Sixth Edition 1820–1823
Supplement 1815–1824
Seventh Edition 1830–1842
Eighth Edition 1852–1860
Ninth Edition 1875–1889
Tenth Edition 1902–1903

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

Fourteenth Edition
© 1929, 1930, 1932, 1933, 1936, 1937, 1938, 1939, 1940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950, 1951, 1952, 1953, 1954, 1955, 1956, 1957, 1958, 1959, 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 1970, 1971, 1972, 1973
By Encyclopædia Britannica, Inc.

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1997, 1998, 2002, 2003
By Encyclopædia Britannica, Inc.

© 2003
By Encyclopædia Britannica, Inc.

Copyright under International Copyright Union
All rights reserved under Pan American, Berne and Universal Copyright Conventions by Encyclopædia Britannica, Inc.

No part of this work may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Printed in U.S.A.

Library of Congress Control Number: 2002113989
International Standard Book Number: 0-85229-961-3

Britannica may be accessed at http://www.britannica.com on the Internet.

Public Library San Mateo, CA

CONTENTS

1 Number Games and Other Mathematical Recreations
14 Number Theory
38 Numerical Analysis
50 Nutrition
76 Occultism
99 Occupational Diseases and Disorders
104 Oceanic Arts
134 Oceans
197 Olympic Games
202 Principles of Optics
217 The Mathematical Theory of Optimization
227 Ōsaka-Kōbe Metropolitan Area
231 Pacific Islands
302 The Pacific Ocean
318 The Art of Painting
336 The History of Western Painting
393 Pakistan
407 Palestine
425 Paraguay
434 Paris
449 Particle Accelerators
458 Pascal
461 Pasteur
463 The Apostle Paul
468 Peking (Beijing)
481 Human Perception
503 Personality
515 Peru
528 Peter I the Great, of Russia
532 Philadelphia
537 Philippines
550 Philosophical Anthropology
562 Western Philosophical Schools and Doctrines
652 Philosophies of the Branches of Knowledge
733 The History of Western Philosophy
761 Photography
799 Photosynthesis
807 Principles of Physical Science
828 The Physical Sciences
860 Picasso
865 Planck
867 Plants
883 Plate Tectonics
893 Plato and Platonism
908 Poisons and Poisoning
930 Poland
957 Police
973 Polish Literature
979 Political Parties and Interest Groups
991 The History of Western Political Philosophy
1004 Political Systems
1032 Jackson Pollock
1034 Population
1044 Portugal
1063 Portuguese Literature
1067 Postal Systems
1077 Prague

CONTENTS

Number Games
and Other Mathematical Recreations

Mathematical recreations comprise puzzles and games that vary from naive amusements to sophisticated problems, some of which have never been solved. They may involve arithmetic, algebra, geometry, theory of numbers, graph theory, topology, matrices, group theory, combinatorics (dealing with problems of arrangements or designs), set theory, symbolic logic, or probability theory. Any attempt to classify this colourful assortment of material is at best arbitrary. Included in this article are the history and the main types of number games and mathematical recreations and the principles on which they are based. Details, including descriptions of puzzles, games, and recreations mentioned in the article, will be found in the references listed in the bibliography.

At times it becomes difficult to tell where pastime ends and serious mathematics begins. An innocent puzzle requiring the traverse of a path may lead to technicalities of graph theory; a simple problem of counting parts of a geometric figure may involve combinatorial theory; dissecting a polygon may involve transformation geometry and group theory; logical inference problems may involve matrices. A problem regarded in medieval times—or before electronic computers became commonplace—as very difficult may prove to be quite simple when attacked by the mathematical methods of today.

Mathematical recreations have a universal appeal. The urge to solve a puzzle is manifested alike by young and old, by the unsophisticated as well as the sophisticated. An outstanding English mathematician, G.H. Hardy, observed that professional puzzle makers, aware of this propensity, exploit it diligently, knowing full well that the general public gets an intellectual kick out of such activities.

The relevant literature has become extensive, particularly since the beginning of the 20th century. Some of it is repetitious, but surprisingly enough, successive generations have found the older chestnuts to be quite delightful, whether dressed in new clothes or not. Much newly created material is continually being added.

This article is divided into the following sections:

History 1
  Early history 1
    Kinds of problems
    Some examples
  Pioneers and imitators 2
  18th and 19th centuries 2
  20th century 2
Types of games and recreations 3
  Arithmetic and algebraic recreations 3
    Number patterns and curiosities
    Digital problems
    Cryptarithms
    Paradoxes and fallacies
    Polygonal and other figurate numbers
    Pythagorean triples
    Perfect numbers and Mersenne numbers
    Fibonacci numbers
  Geometric and topological recreations 6
    Optical illusions
    Geometric fallacies and paradoxes
    Impossible figures
    Pathological curves
    Mazes
    Geometric dissections
    Graphs and networks
    Map-colouring problems
    Flexagons
  Manipulative recreations 9
    Puzzles involving configurations
    Chessboard problems
    The Fifteen Puzzle
    The Tower of Hanoi
    Polyominoes
    Soma Cubes
    Coloured squares and cubes
    Nim and similar games
  Problems of logical inference 12
    Logical puzzles
    Logical paradoxes
Bibliography 13

History

EARLY HISTORY

People have always taken delight in devising "problems" for the purpose of posing a challenge or providing intellectual pleasure. Thus, many mathematical recreations of early origin that have reappeared from time to time in new dress seem to have survived chiefly because they appeal to man's sense of curiosity or mystery. A few survived from the ancient Greeks and Romans: little was known about them during the Dark Ages, but a strong interest in such problems arose during the Middle Ages, stimulated partly by the invention of printing, partly by enthusiastic writers of arithmetic texts, and partly by the rivalry and disputations among early algebraists and scholars. Such activities were most prominent on the Continent, particularly in Italy and Germany. Notable contributors included Rabbi ben Ezra (1140), Fibonacci (Leonardo of Pisa; 1202), Robert Recorde (1542), and Gerolamo Cardano (1545).

Kinds of problems. The problems in general were of two kinds: those involving the manipulation of objects, and those requiring computation. The first required little or no mathematical skill, merely general intelligence and ingenuity, as for example, so-called decanting and difficult crossings problems. A typical example of the former is how to measure out one quart of a liquid if only an eight-, a five-, and a three-quart measure are available. Difficult crossings problems are exemplified by the dilemma of three couples trying to cross a stream in a boat that will hold only two persons, with each husband too jealous to leave his wife in the company of either of the other men. Many variants of both types of problems have appeared over the years.

Manipulation of objects and computation

Some examples. Problems involving computation also took on a variety of forms; some were as follows:

Finding a number. Think of a number, triple it, and take half the product; triple this and take half the result; then divide by 9. The quotient will be one-fourth the original number.

"God-Greet-You" problems. For example, in "God greet you, all you 30 companions," someone says: "If there were as many of us again and half as many more, then there would be 30 of us." How many were there?

The chessboard problem. How many grains of wheat are required in order to place one grain on the first square, 2 on the second, 4 on the third, and so on for the 64 squares?

The lion in the well. This is typical of many problems dealing with the time required to cover a certain distance at a constant rate while at the same time progress is hindered by a constant retrograde motion. There is a lion in a well whose depth is 50 palms. He climbs $1/7$ of a palm

daily and slips back 1/9 of a palm. In how many days will he get out of the well?

Courier problems. These are typified by the movements of bodies at given rates in which some position of these bodies is given and the time required for them to arrive at some other specified position is demanded.

PIONEERS AND IMITATORS

The 17th century produced books devoted solely to recreational problems not only in mathematics but frequently in mechanics and natural philosophy as well. The first important contribution was that of the Frenchman Claude-Gaspar Bachet de Méziriac, one of the earliest pioneers in this field, who is remembered for two mathematical works: his *Diophanti,* the first edition of a Greek text on the theory of numbers (1621), and his *Problèmes plaisans et delectables qui se font par les nombres* (1612). The latter passed through five editions, the last as late as 1959; it was the forerunner of similar collections of recreations to follow. The emphasis was placed on arithmetic rather than geometric puzzles. Among the outstanding problems given by Bachet were questions involving number bases other than 10; card tricks; watch-dial puzzles depending on numbering schemes; the determination of the smallest set of weights that would enable one to weigh any integral number of pounds from one pound to 40, inclusive; and difficult crossings or ferry problems.

Books and collections

In 1624 a French Jesuit, Jean Leurechon, writing under the pen name of van Etten, published *Récréations mathématiques.* This volume struck the popular fancy, passing through at least 30 editions before 1700, despite the fact that it was based largely on the work of Bachet, from whom he took the simpler problems, disregarding the more significant portions. Yet it did contain some original work, and it served as a model for others, including Mydorge and Schwenter. The first English edition (1633) bore the title: *Mathematicall Recreations, or a Collection of Sundrie Problemes, extracted out of the Ancient and Moderne Philosophers, as Secrets in Nature, and Experiments in Arithmeticke, Geometrie, Cosmographie, Horologographie, Astronomie, Navigation, Musicke, Opticks, Architecture, Staticke, Machanicks, Chimestrie, Waterworkes, Fireworks, etc. Not vulgarly made manifest until this Time . . . . Most of which were written first in Greeke and Latine, lately compiled in French, by HENRY VAN ETTEN Gent. And now delivered in the English Tongue with the Examinations, Corrections, and Augmentations [translated by William Oughtred].*

The rising tide of interest was exploited by French mathematicians Claude Mydorge, whose *Examen du livre des récréations mathématiques* was published in 1630, and Denis Henrion, whose *Les Récréations mathématiques avec l'examen de ses problèmes en arithmétique, géométrie, méchanique, cosmographie, optique, catoptrique,* etc., based largely upon Mydorge's book, appeared in 1659. Leurechon's book, meanwhile, had found its way into Germany: Daniel Schwenter, a professor of Hebrew, Oriental languages, and mathematics, assiduously compiled a comprehensive collection of recreational problems based on a translation of Leurechon's book, together with many other problems that he himself had previously collected. This work appeared posthumously in 1636 under the title *Deliciae Physico-mathematicae oder Mathematische und Philosophische Erquickstunden.* Immensely popular, Schwenter's book was enlarged by two supplementary editions in 1651–53. For some years thereafter Schwenter's enlarged edition was the most comprehensive treatise of its kind, although in 1641–42 the Italian Jesuit Mario Bettini had issued a two-volume work called *Apiaria Universae Philosophiae Mathematicae in Quibus Paradoxa et Nova Pleraque Machinamenta Exhibentur,* which was followed in 1660 by a third volume entitled *Recreationum Mathematicarum Apiaria Novissima Duodecim . . . .* And in 1665 one Johann Mohr in Schleswig published an imitation of Schwenter under the title of *Arithmetische Lustgarten.*

In England, somewhat belatedly, William Leybourn, a mathematics teacher, textbook writer, and surveyor, in 1694, published his *Pleasure with Profit: Consisting of Recreations of Divers Kinds, viz., Numerical, Geometrical, Mechanical, Statical. Astronomical, Horometrical, Cryptographical, Magnetical, Automatical, Chymical, and Historical.* The title page further states that the purpose of the book was to "recreate ingenious spirits and to induce them to make farther scrutiny into these sublime sciences, and to divert them from following such vices, to which Youth (in this Age) are so much inclined." Much of the volume is conventional textbook material, for most of Leybourn's published works grew out of his teaching.

18TH AND 19TH CENTURIES

The 18th century saw a continuation of this interest. Published in England were volumes by Edward Hatton, Thomas Gent, Samuel Clark, and William Hooper. In 1775 Charles Hutton published five volumes of extracts from the *Ladies' Diary* dealing with "entertaining mathematical and poetical parts." On the Continent there appeared several writers, including: Christian Pescheck, Abat Bonaventura, the Dutch writer Paul Halcken, and A.H. Guyot's four volumes of *Nouvelles Récréations physiques et mathématiques,* etc. (1769, 1786). But by far the outstanding work was that of Jacques Ozanam, the precursor of books to follow for the next 200 years. First published in four volumes in 1694, his *Récréations mathématique et physiques* went through many editions; based on the works of Bachet, Mydorge, Leurechon, and Schwenter, it was later revised and enlarged by Montucla, then translated into English by Charles Hutton (1803, 1814) and again revised by Edward Riddle (1840, 1844).

The work of Ozanam

The first half of the 19th century produced only a moderate number of lesser writers on mathematical recreations, but the second half of the 19th century witnessed a crescendo of interest, culminating in the outstanding contributions of Édouard Lucas, C.L. Dodgson (Lewis Carroll), and others at the turn of the century. Lucas' four-volume *Récréations mathématiques* (1882–94) became a classic. The mathematical recreations of Dodgson included *Symbolic Logic* and *The Game of Logic; Pillow Problems* and *A Tangled Tale,* 2 vol. (1885–95).

20TH CENTURY

Among the more colourful figures at the turn of the 20th century were two Americans named Sam Loyd, father and son. Tremendously successful in making puzzles, the elder Loyd sold his weekly puzzle column to a national syndicate for years, and, in addition, created or adapted hundreds of mechanical puzzles fashioned of cardboard, wood, and metal that were also financially rewarding. When Loyd II died in 1934 at the age of 60, it was estimated that he had produced at least 10,000 puzzles.

In Germany, Hermann Schubert published *Zwölf Geduldspiele* in 1899 and the *Mathematische Mussestunden* (3rd ed., 3 vol.) in 1907–09. Between 1904 and 1920 Wilhelm Ahrens published several works, the most significant being his *Mathematische Unterhaltungen und Spiele* (2 vol., 1910) with an extensive bibliography.

Among British contributors, Henry Dudeney, a contributor to the *Strand Magazine,* published several very popular collections of puzzles that have been reprinted from time to time (1917–67). The first edition of W.W. Rouse Ball's *Mathematical Recreations and Essays* appeared in 1892; it soon became a classic, largely because of its scholarly approach. After passing through 10 editions it was revised by the British professor H.S.M. Coxeter in 1938; it is still a standard reference.

Outstanding work was that of Maurice Kraitchik, editor of the periodical *Sphinx* and author of several well-known works published between 1900 and 1942.

About the middle third of the 20th century, there was a gradual shift in emphasis on various topics. Up to that time interest had focussed largely on such amusements as numerical curiosities; simple geometric puzzles; arithmetical story problems; paper folding and string figures; geometric dissections; manipulative puzzles; tricks with numbers and with cards; magic squares; those venerable diversions concerning angle trisection, duplication of the cube, squaring the circle, as well as the elusive fourth dimension. By the middle of the century, interest be-

Modern shift in emphasis

gan to swing toward more mathematically sophisticated topics: cryptograms; recreations involving modular arithmetic, numeration bases, and number theory; graphs and networks; lattices, group theory; topological curiosities; packing and covering; flexagons; manipulation of geometric shapes and forms; combinatorial problems; probability theory; inferential problems; logical paradoxes; fallacies of logic; and paradoxes of the infinite.

Types of games and recreations

ARITHMETIC AND ALGEBRAIC RECREATIONS

Number patterns and curiosities. Some groupings of natural numbers, when operated upon by the ordinary processes of arithmetic, reveal rather remarkable patterns, affording pleasant pastimes. For example:

$1 \times 8 + 1 = 9$	$3 \times 37 = 111$	$(1)^2 = 1$
$12 \times 8 + 2 = 98$	$6 \times 37 = 222$	$(11)^2 = 121$
$123 \times 8 + 3 = 987$	$9 \times 37 = 333$	$(111)^2 = 12321$
$1234 \times 8 + 4 = 9876$	$12 \times 37 = 444$	$(1111)^2 = 1234321$
etc.	etc.	etc.

Another type of number pleasantry concerns multigrades; *i.e.,* identities between the sums of two sets of numbers and the sums of their squares or higher powers—*e.g.,*

$$1^n + 6^n + 8^n = 2^n + 4^n + 9^n \text{ (for } n = 1 \text{ or } 2).$$

An easy method of forming a multigrade is to start with a simple equality—*e.g.,* $1 + 5 = 2 + 4$—then add, for example, 5 to each term: $6 + 10 = 7 + 9$. A second-order multigrade is obtained by "switching sides" and combining, as shown below:

$$1^n + 5^n + 7^n + 9^n = 2^n + 4^n + 6^n + 10^n \ (n = 1 \text{ or } 2).$$

On each side the sum of the first powers (S_1) is 22 and of the second powers (S_2) is 156.

Ten may be added to each term to derive a third-order multigrade:

$$11^n + 15^n + 17^n + 19^n = 12^n + 14^n + 16^n + 20^n \ (n = 1 \text{ or } 2).$$

Switching sides and combining, as before:

$$1^n + 5^n + 7^n + 9^n + 12^n + 14^n + 16^n + 20^n = 2^n + 4^n + 6^n + 10^n + 11^n + 15^n + 17^n + 19^n \quad (n = 1, 2, \text{ or } 3).$$

In this example $S_1 = 84$, $S_2 = 1{,}152$, and $S_3 = 17{,}766$.

This process can be continued indefinitely to build multigrades of successively higher orders. Similarly, all terms in a multigrade may be multiplied or divided by the same number without affecting the equality. Many variations are possible: for example, palindromic multigrades that read the same backward and forward, and multigrades composed of prime numbers.

Narcissistic numbers

Other number curiosities and oddities are to be found. Thus, narcissistic numbers are numbers that can be represented by some kind of mathematical manipulation of their digits. A whole number, or integer, that is the sum of the nth powers of its digits (*e.g.,* $153 = 1^3 + 5^3 + 3^3$) is called a perfect digital invariant. On the other hand, a recurring digital invariant is illustrated by:

$$55: \ 5^3 + 5^3 = 250;$$
$$250: \ 2^3 + 5^3 + 0^3 = 133;$$
$$133: \ 1^3 + 3^3 + 3^3 = 55.$$

(From *Mathematics on Vacation,* Joseph Madachy; Charles Scribner's Sons.)

A variation of such digital invariants is

$$165{,}033 = 1^6 3^3 + 50^3 + 33^3.$$

Another curiosity is exemplified by a number that is equal to the nth power of the sum of its digits:

$$81 = (8 + 1)^2 = 9^2;$$
$$4913 = (4 + 9 + 1 + 3)^3 = 17^3.$$

An automorphic number is an integer whose square ends with the given integer, as $(25)^2 = 625$, and $(76)^2 = 5776$. Strobogrammatic numbers read the same after having been rotated through 180°; *e.g.,* 69, 96, 1001.

It is not improbable that such curiosities should have suggested intrinsic properties of numbers bordering on mysticism.

Problem of the four n's

Digital problems. The problem of the four n's calls for the expression of as large a sequence of integers as possible, beginning with 1, representing each integer in turn by a given digit used exactly four times. The answer depends upon the rules of operation that are admitted. Two partial examples are shown.

For four 1s:

$$1 = 1 + \tfrac{1}{1} - 1$$
$$2 = 1 + 1 + 1 - 1$$
$$3 = 1 + 1 + \tfrac{1}{1}$$
$$4 = 1 + 1 + 1 + 1$$
$$5 = (1 + 1 + 1)! - 1$$
etc.

For four 4s:

$$1 = (\tfrac{4}{4}) \cdot (\tfrac{4}{4})$$
$$2 = \tfrac{4}{4} + \tfrac{4}{4}$$
$$3 = \tfrac{4}{4} + \sqrt{4/4}$$
$$4 = \sqrt{(4)(4)} \cdot (\tfrac{4}{4})$$
$$5 = \sqrt{4} + \sqrt{4} + \tfrac{4}{4}$$
etc.

(In M. Bicknell & V. Hoggatt, "64 Ways to Write 64 Using Four 4's," *Recreational Mathematics Magazine,* No. 14, Jan.–Feb. 1964, p. 13.)

Obviously, many alternatives are possible; *e.g.,* $7 = 4 + \sqrt{4} + 4/4$ could also be expressed as $4!/4 + 4/4$, or as $44/4 - 4$. The factorial of a positive integer is the product of all the positive integers less than or equal to the given integer; *e.g.,* "factorial 4," or $4! = 4 \times 3 \times 2 \times 1$. If the use of factorial notation is not allowed, it is still possible to express the numbers from 1 to 22 inclusive with four "4s"; thus $22 = (4 + 4)/.4 + \sqrt{4}$. But if the rules are extended, many additional combinations are possible.

A similar problem requires that the integers be expressed by using the first m positive integers, $m > 3$ ("m is greater than three") and the operational symbols used in elementary algebra. For example, using the digits 1, 2, 3, and 4:

$$2 = 4 - 3 + 2 - 1$$
$$4 = \sqrt{4} + 3 - 2 + 1$$
$$6 = \sqrt{4} + 3 + 2 - 1$$
$$8 = \sqrt{4} + 3 + 2 + 1$$
$$9 = 4 + 3 + (2 \cdot 1).$$

Such problems have many variations; for example, more than 100 ways of arranging the digits 1 to 9, *in order,* to give a value of 100 have been demonstrated.

All of these digital problems require considerable ingenuity but involve little significant mathematics.

Cryptarithms. The term "crypt-arithmetic" was introduced in 1931, when the following multiplication problem appeared in the Belgian journal *Sphinx:*

$$\begin{array}{r} \text{ABC} \\ \text{DE} \\ \hline \text{FEC} \\ \text{DEC} \\ \hline \text{HGBC} \end{array}$$

The shortened word cryptarithm now denotes mathematical problems usually calling for addition, subtraction, multiplication, or division and replacement of the digits by letters of the alphabet or some other symbols.

An analysis of the original puzzle suggested the general method of solving a relatively simple cryptarithm:

1. In the second partial product $D \times A = D$, hence $A = 1$.
2. $D \times C$ and $E \times C$ both end in C; since for any two digits 1–9 the only multiple that will produce this result is 5 (zero if both digits are even, 5 if both are odd), $C = 5$.
3. D and E must be odd. Since both partial products have only three digits, neither D nor E can be 9. This leaves only 3 and 7. In the first partial product $E \times B$ is a number of two digits, while in the second partial product $D \times B$ is a number of only one digit. Thus E is larger than D, so $E = 7$ and $D = 3$.
4. Since $D \times B$ has only one digit, B must be 3 or less. The only two possibilities are 0 and 2. B cannot be zero because 7B is a two digit number. Thus $B = 2$.
5. By completing the multiplication, $F = 8$, $G = 6$, and $H = 4$.
6. Answer: $125 \times 37 = 4{,}625$.

(From *150 Puzzles in Crypt-Arithmetic* by Maxey Brooke; Dover Publications, Inc., New York, 1963. Reprinted through the permission of the publisher.)

Such puzzles had apparently appeared, on occasion, even earlier. Alphametics refers specifically to cryptarithms in which the combinations of letters make sense, as in one of the oldest and probably best known of all alphametics:

$$\begin{array}{r} \text{SEND} \\ +\ \text{MORE} \\ \hline \text{MONEY} \end{array}$$

Unless otherwise indicated, convention requires that the initial letters of an alphametic cannot represent zero, and that two or more letters may not represent the same digit. If these conventions are disregarded, the alphametic must be accompanied by an appropriate clue to that effect. Some cryptarithms are quite complex and elaborate and have multiple solutions. Electronic computers have been used for the solution of such problems.

Paradoxes and fallacies. Mathematical paradoxes and fallacies have long intrigued mathematicians. A mathematical paradox is a mathematical conclusion so unexpected that it is difficult to accept even though every step in the reasoning is valid. A mathematical fallacy, on the other hand, is an instance of improper reasoning leading to an unexpected result that is patently false or absurd. The error in a fallacy generally violates some principle of logic or mathematics, often unwittingly. Such fallacies are quite puzzling to the tyro, who, unless he is aware of the principle involved, may well overlook the subtly concealed error. A sophism is a fallacy in which the error has been knowingly committed, for whatever purpose. If the error introduced into a calculation or a proof leads innocently to a *correct* result, the result is a "howler," often said to depend on "making the right mistake."

Many paradoxes arise from the concepts of infinity and limiting processes. For example, the infinite series

$$1 + \tfrac{1}{2} + \tfrac{1}{4} + \tfrac{1}{8} + \cdots$$

has a continually greater sum the more terms are included, but the sum always remains less than 2, although it approaches nearer and nearer to 2 as more terms are included. On the other hand, the series

$$1 + \tfrac{1}{2} + \tfrac{1}{3} + \tfrac{1}{4} + \tfrac{1}{5} + \tfrac{1}{6} + \cdots$$

is called divergent: it has no limit, the sum becoming larger than any chosen value if sufficient terms are taken. Another paradox is the fact that there are just as many even natural numbers as there are even and odd numbers altogether, thus contradicting the notion that "the whole is greater than any of its parts." This seeming contradiction arises from the properties of collections containing an infinite number of objects. Since both are infinite, they are for both practical and mathematical purposes equal.

Zeno's paradoxes

The so-called paradoxes of Zeno (*c.* 450 BC) are, strictly speaking, sophisms. In the race between Achilles and the tortoise, the two start moving at the same moment, but, if the tortoise is initially given a lead and continues to move ahead, Achilles can run at any speed and never catch up. Zeno's argument rests on the presumption that Achilles must first reach the point where the tortoise started, by which time the tortoise will have moved ahead to another point, and so on. Obviously, Zeno did not believe what he claimed; his interest lay in locating the error in his argument. The same observation is true of the three remaining paradoxes of Zeno, the *Dichotomy,* "motion is impossible"; the *Arrow,* "motionless even while in flight"; and the *Stadium,* or "a given time interval is equivalent to an interval twice as long." Beneath the sophistry of these contradictions lie subtle and elusive concepts of limits and infinity, only completely explained in the 19th century when the foundations of analysis became more rigorous and the theory of transfinite numbers had been formulated.

Common algebraic fallacies usually involve a violation of one or another of the following assumptions:

1. If $a = b$, then $a/k = b/k$, provided $k \neq 0$.
2. If $a > b$, then $ka > kb$, provided k is positive.
3. If a is nonnegative, then $\sqrt{a^2} = +a$.

Three examples of such violations follow:

A. Solve: $6x - 18 = 4x - 12$
Factoring: $3(2x - 6) = 2(2x - 6)$
Dividing by $(2x - 6)$: $3 = 2$

B. Since $+1/-1 = -1/+1$, then $\sqrt{+1}/\sqrt{-1} = \sqrt{-1}/\sqrt{+1}$, and so $(\sqrt{+1})(\sqrt{+1}) = (\sqrt{-1})(\sqrt{-1})$, hence $+1 = -1$

C. Given two positive numbers, a and b:

then,	also,
$a > -b$	$b > -a$
$b > -b$	$a > -a$
Multiplying: $ab > b^2$	$ab > a^2$
Or $a > b$	$b > a$

Thus a is both greater than b and less than b.

An example of an illegal operation or "lucky boner" is:

$$\tfrac{16}{64} = \tfrac{19}{64} = \tfrac{1}{4}.$$

Polygonal and other figurate numbers. Among the many relationships of numbers that have fascinated man are those that suggest (or were derived from) the arrangement of points representing numbers into series of geometrical figures. Such numbers, known as figurate or polygonal numbers, appeared in 15th-century arithmetic books and were probably known to the ancient Chinese; but they were of especial interest to the ancient Greek mathematicians. To the Pythagoreans (*c.* 500 BC), numbers were of paramount significance; everything could be explained by numbers, and numbers were invested with specific characteristics and personalities. Among other properties of numbers, the Pythagoreans recognized that numbers had "shapes." Thus, the triangular numbers, 1, 3, 6, 10, 15, 21, etc., were visualized as points or dots arranged in the shape of a triangle.

The "shapes" of numbers

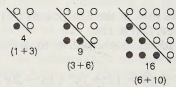


Figure 1: Square numbers shown formed from consecutive triangular numbers.

From *The Number of Things: Pythagoras, Geometry and Humming Strings* by Evans G. Valens, copyright © 1964 by Evans G. Valens; published by E.P. Dutton & Co., Inc., and used with their permission

Square numbers are the squares of natural numbers, such as 1, 4, 9, 16, 25, etc., and can be represented by square arrays of dots, as shown in Figure 1. Inspection reveals that the sum of any two adjacent triangular numbers is always a square number.

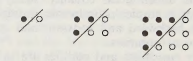


Figure 2: Oblong numbers formed by doubling triangular numbers.

From *The Number of Things: Pythagoras, Geometry and Humming Strings* by Evans G. Valens, copyright © 1964 by Evans G. Valens; published by E.P. Dutton & Co., Inc., and used with their permission

Oblong numbers are the numbers of dots that can be placed in rows and columns in a rectangular array, each row containing one more dot than each column. The first few oblong numbers are 2, 6, 12, 20, and 30. This series of numbers is the successive sums of the series of even numbers or the products of two consecutive numbers: $2 = 1 \cdot 2$; $6 = 2 \cdot 3 = 2 + 4$; $12 = 3 \cdot 4 = 2 + 4 + 6$; $20 = 4 \cdot 5 = 2 + 4 + 6 + 8$; etc. An oblong number also is formed by doubling any triangular number (see Figure 2).

The gnomons include all of the odd numbers; these can be represented by a right angle, or a carpenter's square, as illustrated in Figure 3. Gnomons were extremely use-

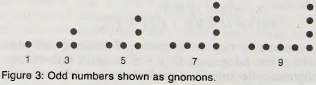


Figure 3: Odd numbers shown as gnomons.

From *The Number of Things: Pythagoras, Geometry and Humming Strings* by Evans G. Valens, copyright © 1964 by Evans G. Valens; published by E.P. Dutton & Co., Inc., and used with their permission



Figure 4: Addition of gnomons to squares to form larger squares.

From *The Number of Things: Pythagoras, Geometry and Humming Strings* by Evans G. Valens, copyright © 1964 by Evans G. Valens; published by E.P. Dutton & Co., Inc., and used with their permission

ful to the Pythagoreans. They could build up squares by adding gnomons to smaller squares and from such a figure could deduce many interrelationships: thus $1^2 + 3 = 2^2$, $2^2 + 5 = 3^2$, etc.; or $1 + 3 + 5 = 3^2$, $1 + 3 + 5 + 7 = 4^2$, $1 + 3 + 5 + 7 + 9 = 5^2$, etc. (see Figure 4). Indeed, it is quite likely that Pythagoras first realized the famous relationship between the sides of a right triangle, represented by $a^2 + b^2 = c^2$, by contemplating the properties of gnomons and square numbers, observing that any odd square can be added to some even square to form a third square. Thus

$$3^2 + 4^2 = 5^2, \text{ where } 3^2 = 4 + 5;$$

$$5^2 + 12^2 = 13^2, \text{ where } 5^2 = 12 + 13;$$

and, in general, $a^2 + b^2 = c^2$, where $a^2 = b + c$. This is a special class of Pythagorean triples (see below).

Besides these, the Greeks also studied numbers having pentagonal, hexagonal, and other shapes. Many relationships can be shown to exist between these geometric patterns and algebraic expressions.

Figurate numbers

Polygonal numbers constitute a subdivision of a class of numbers known as figurate numbers. Examples include the arithmetic sequences

$$1, 2, 3, 4, \ldots, r$$

and

$$1, 3, 5, 7, \ldots, (2r - 1).$$

When new series are formed from the sums of the terms of these series, the results are, respectively,

$$1, 3, 6, 10, \ldots$$

and

$$1, 4, 9, 16, \ldots.$$

These series are not arithmetic sequences but are seen to be the polygonal triangular and square numbers. Polygonal number series can also be added to form three-dimensional figurate numbers; these sequences are called pyramidal numbers.

The significance of polygonal and figurate numbers lies in their relation to the modern theory of numbers. Even the simple, elementary properties and relations of numbers often demand sophisticated mathematical tools. Thus, it has been shown that every integer is either a triangular number, the sum of two triangular numbers, or the sum of three triangular numbers: *e.g.*, $8 = 1 + 1 + 6$, $42 = 6 + 36$, $43 = 15 + 28$, $44 = 6 + 10 + 28$.

Pythagorean triples. The study of Pythagorean triples as well as the general theorem of Pythagoras leads to many unexpected byways in mathematics. A Pythagorean triple is formed by the measures of the sides of an integral right triangle—*i.e.*, any set of three positive integers such that $a^2 + b^2 = c^2$. If a, b, and c are relatively prime—*i.e.*, if no two of them have a common factor—the set is a primitive Pythagorean triple.

A formula for generating all primitive Pythagorean triples is

$$a = p^2 - q^2,\ b = 2pq,\ c = p^2 + q^2,$$

in which p and q are relatively prime, p and q are neither both even nor both odd, and $p > q$. By choosing p and q appropriately, for example, primitive Pythagorean triples such as the following are obtained:

p	q	(*a*) $p^2 - q^2$	(*b*) $2pq$	(*c*) $p^2 + q^2$
2	1	3	4	5
3	2	5	12	13

The only primitive triple that consists of consecutive integers is 3, 4, 5.

Certain characteristic properties are of interest:

1. Either a or b is divisible by 3.
2. Either a or b is divisible by 4.
3. Either a or b or c is divisible by 5.
4. The product of a, b, and c is divisible by 60.
5. One of the quantities a, b, $a + b$, $a - b$ is divisible by 7.

It is also true that if n is any integer, then $2n + 1$, $2n^2 + 2n$, and $2n^2 + 2n + 1$ form a Pythagorean triple.

Certain properties of Pythagorean triples were known to the ancient Greeks—*e.g.*, that the hypotenuse of a primitive triple is always an odd integer. It is now known that an odd integer R is the hypotenuse of such a triple if and only if every prime factor of R is of the form $4k + 1$, where k is a positive integer.

Perfect numbers and Mersenne numbers. Most numbers are either "abundant" or "deficient." In an abundant number, the sum of its proper divisors (*i.e.*, including 1 but excluding the number itself) is greater than the number; in a deficient number, the sum of its proper divisors is less than the number. A perfect number is an integer that equals the sum of its proper divisors. For example, 24 is abundant, its divisors giving a sum of 36; 32 is deficient, giving a sum of 31. The number 6 is a perfect number, since $1 + 2 + 3 = 6$; so is 28, since $1 + 2 + 4 + 7 + 14 = 28$. The next two perfect numbers are 496 and 8,128. The first four perfect numbers were known to the ancients. Indeed, Euclid suggested that any number of the form $2^{n-1}(2^n - 1)$ is a perfect number whenever $2^n - 1$ is prime, but it was not until the 18th century that the Swiss mathematician Leonhard Euler proved that every even perfect number must be of the form $2^{n-1}(2^n - 1)$, where $2^n - 1$ is a prime.

A number of the form $2^n - 1$ is called a Mersenne number after the French mathematician Marin Mersenne; it may be prime (*i.e.*, having no factor except itself or 1) or composite (composed of two or more prime factors). A necessary though not sufficient condition that $2^n - 1$ be a prime is that n be a prime. Thus, all even perfect numbers have the form $2^{n-1}(2^n - 1)$ where both n and $2^n - 1$ are prime numbers. Until comparatively recently, only 12 perfect numbers were known. In 1876 the French mathematician Édouard Lucas found a way to test the primality of Mersenne numbers. By 1952 the U.S. mathematician Raphael M. Robinson had applied Lucas' test and, by means of electronic digital computers, had found the Mersenne primes for $n = 521$; 607; 1,279; 2,203; and 2,281, thus adding five more perfect numbers to the list. By 1989 there were more than 30 known perfect numbers.

It is known that to every Mersenne prime there corresponds an even perfect number and vice versa. But two questions are still unanswered: the first is whether there are any odd perfect numbers, and the second is whether there are infinitely many perfect numbers.

Properties of perfect numbers

Many remarkable properties are revealed by perfect numbers. All perfect numbers, for example, are triangular. Also, the sum of the reciprocals of the divisors of a perfect number (including the reciprocal of the number itself) is always equal to 2. Thus

for 6: $\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{6} = 2$

and for 28: $\frac{1}{1} + \frac{1}{2} + \frac{1}{4} + \frac{1}{7} + \frac{1}{14} + \frac{1}{28} = 2.$

Fibonacci numbers. In 1202 the mathematician Leonardo of Pisa, also called Fibonacci, published an influential treatise, *Liber abaci.* It contained the following recreational problem: "How many pairs of rabbits can be produced from a single pair in one year if it is assumed that every month each pair begets a new pair which from the second month becomes productive?" Straightforward calculation generates the following sequence:

Month:	1	2	3	4	5	6	7	8	9	10	11	12
No. of pairs:	1	1	2	3	5	8	13	21	34	55	89	144

The second row represents the first 12 terms of the sequence now known by Fibonacci's name, in which each term (except the first two) is found by adding the two terms immediately preceding; in general, $x_n = x_{n-1} + x_{n-2}$, a relation that was not recognized until about 1600.

Over the years, especially in the middle decades of the 20th century, the properties of the Fibonacci numbers have been extensively studied, resulting in a considerable literature. Their properties seem inexhaustible; for example, $x_{n+1} \cdot x_{n-1} = x_n^2 + (-1)^n$. Another formula for generating the Fibonacci numbers is attributed to Édouard Lucas:

$$x_n = \frac{1}{\sqrt{5}} \left\{ \left(\frac{1+\sqrt{5}}{2} \right)^n - \left(\frac{1-\sqrt{5}}{2} \right)^n \right\}.$$

The ratio $(\sqrt{5}+1) : 2 = 1.618\ldots$, designated as Φ, is known as the golden number; the ratio $(\sqrt{5}-1) : 2$, the reciprocal of Φ, is equal to $0.618\ldots$. Both these ratios are related to the roots of $x^2 - x - 1 = 0$, an equation derived from the Divine Proportion of the 15th-century Italian mathematician Luca Pacioli, namely, $a/b = b/(a+b)$, when $a < b$, by setting $x = b/a$. In short, dividing a segment into two parts in mean and extreme proportion, so that the smaller part is to the larger part as the larger is to the entire segment, yields the so-called Golden Section, an important concept in both ancient and modern artistic and architectural design. Thus, a rectangle the sides of which are in the approximate ratio of $3 : 5$ ($\Phi^{-1} = 0.618\ldots$), or $8 : 5$ ($\Phi = 1.618\ldots$), is presumed to have the most pleasing proportions, aesthetically speaking.

The Golden Section

Raising the golden number to successive powers generates the sequence that begins as follows:

$$\Phi = (\sqrt{5}+1)/2 \qquad \Phi^4 = (3\sqrt{5}+7)/2$$
$$\Phi^2 = (\sqrt{5}+3)/2 \qquad \Phi^5 = (5\sqrt{5}+11)/2$$
$$\Phi^3 = (2\sqrt{5}+4)/2 \qquad \Phi^6 = (8\sqrt{5}+18)/2$$

In this sequence the successive coefficients of the radical $\sqrt{5}$ are Fibonacci's 1, 1, 2, 3, 5, 8, while the successive second terms within the parentheses are the so-called Lucas sequence: 1, 3, 4, 7, 11, 18. The Lucas sequence shares the recursive relation of the Fibonacci sequence; that is, $x_n = x_{n-1} + x_{n-2}$.

If a golden rectangle ABCD is drawn and a square ABEF is removed, the remaining rectangle ECDF is also a golden rectangle. If this process is continued and circular arcs are drawn, the curve formed approximates the logarithmic spiral, a form found in nature (see Figure 5). The

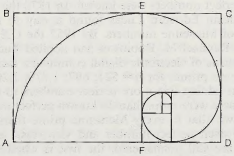


Figure 5: Golden rectangles and the logarithmic spiral.

logarithmic spiral is the graph of the equation $r = k^{\Phi\theta}$, in polar coordinates, where $k = \Phi^{2/\pi}$. The Fibonacci numbers are also exemplified by the botanical phenomenon known as phyllotaxis. Thus, the arrangement of the whorls on a pinecone or pineapple, of petals on a sunflower, and of branches from some stems follows a sequence of Fibonacci numbers or the series of fractions

$$\tfrac{1}{1}, \tfrac{1}{2}, \tfrac{2}{3}, \tfrac{3}{5}, \tfrac{5}{8}, \tfrac{8}{13}, \text{ etc.}$$

GEOMETRIC AND TOPOLOGICAL RECREATIONS

Optical illusions. The creation and analysis of optical illusions may involve mathematical and geometric principles such as the proportionality between the areas of similar figures and the squares of their linear dimensions. Some involve physiological or psychological considerations, such as the fact that, when making visual comparisons, relative lengths are more accurately perceived than relative areas.

For treatment of optical illusions and their illusory effects, including unorthodox use of perspective, distorted angles, deceptive shading, unusual juxtaposition, equivocal contours or contrasts, colour effects, chromatic aberration, and afterimages, see the article PERCEPTION.

Geometric fallacies and paradoxes. Some geometric fallacies include "proofs": (1) that every triangle is isosceles (*i.e.*, has two equal sides); (2) that every angle is a right angle; (3) that if ABCD is a quadrilateral in which AB = CD, then AD must be parallel to BC; and (4) that every point in the interior of a circle lies on the circle.

The explanations of fallacious proofs in geometry usually include one or another of the following: faulty construction; violation of a logical principle, such as assuming the truth of a converse, or confusing partial inverses or converses; misinterpretation of a definition, or failing to take note of "necessary and sufficient" conditions; too great dependence upon diagrams and intuition; being trapped by limiting processes and deceptive appearances.

Impossible figures. At first glance, drawings such as those in Figure 6 appear to represent plausible three-dimensional objects, but closer inspection reveals that they

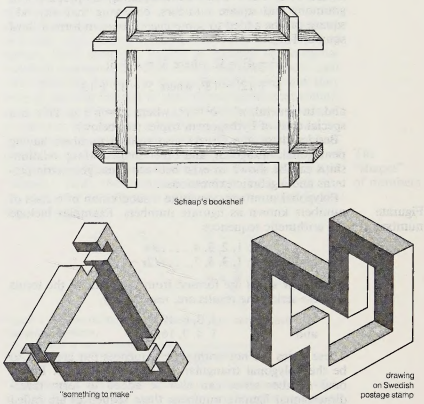


Figure 6: Impossible figures.

cannot; the representation is flawed by faulty perspective, false juxtaposition, or psychological distortion. Among the first to produce these drawings—also called undecidable figures—was Oscar Reutersvard of Sweden, who made them the central features of a set of Swedish postage stamps.

In 1958 L.S. Penrose, a British geneticist, and his son Roger Penrose, a mathematical physicist, introduced the undecidable figures called strange loops. One of these is the Penrose square stairway (Figure 7), which one could apparently traverse in either direction forever without get-

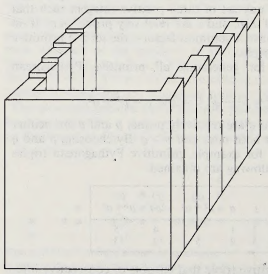


Figure 7: The endless stair.

ting higher or lower. Strange loops are important features of some of M.C. Escher's lithographs, including "Ascending and Descending" (1960) and "Waterfall" (1961). The concept of the strange loop is related to the idea of infinity and also to logical paradoxes involving self-referential statements, such as that of Epimenides (see below *Logical paradoxes*).

Pathological curves. A mathematical curve is said to be pathological if it lacks certain properties of continuous curves. For example, its tangent may be undefined at some—or indeed any—point; the curve may enclose a finite area but be infinite in length; or its curvature may be undefinable. Some of these curves may be regarded as the limit of a series of geometrical constructions; their lengths or the areas they enclose appear to be the limits of sequences of numbers. Their idiosyncrasies constitute paradoxes rather than optical illusions or fallacies.

Von Koch's snowflake curve, for example, is the figure obtained by trisecting each side of an equilateral triangle and replacing the centre segment by two sides of a smaller equilateral triangle projecting outward, then treating the resulting figure the same way, and so on. The first two stages of this process are shown in Figure 8. As the construction proceeds, the perimeter of the curve increases without limit, but the area it encloses does approach an upper bound, which is $^8/_5$ the area of the original triangle.



Figure 8: Von Koch's snowflake curve.

In seeming defiance of the fact that a curve is "one-dimensional" and thus cannot fill a given space, it can be shown that the curve produced by continuing the stages in Figure 9, when completed, will pass through *every* point in the square. In fact, by similar reasoning, the curve can be made to fill completely an entire cube.

From E. Kasner and J. Newman, *Mathematics and the Imagination* (copyright © 1940 by Edward Kasner and James R. Newman); reprinted by permission of Simon and Schuster, Inc.

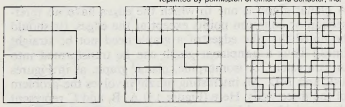


Figure 9: A space-filling curve (see text).

The Sierpinski curve, the first few stages of which are shown in Figure 10, contains every point interior to a square, and it describes a closed path. As the process of forming the curve is continued indefinitely, the length of the curve approaches infinity, while the area enclosed by it approaches $^5/_{12}$ that of the square.

A fractal curve, loosely speaking, is one that retains the same general pattern of irregularity regardless of how much it is magnified; von Koch's snowflake is such a curve. At each stage in its construction, the length of its perimeter increases in the ratio of 4 to 3. The mathematician Benoit Mandelbrot has generalized the term dimension, symbolized D, to denote the power to which 3 must be raised to produce 4; that is, $3^D = 4$. The dimension that characterizes von Koch's snowflake is therefore log 4/log 3, or approximately 1.26.

Beginning in the 1950s Mandelbrot and others have intensively studied the self-similarity of pathological curves, and they have applied the theory of fractals in modelling natural phenomena. Random fluctuations induce a statistical self-similarity in natural patterns; analysis of these patterns by Mandelbrot's techniques has been found useful in such diverse fields as fluid mechanics, geomorphology, human physiology, economics, and linguistics. Specifically, for example, characteristic "landscapes" revealed by microscopic views of surfaces in connection with Brownian movement, vascular networks, and the shapes of polymer molecules are all related to fractals.

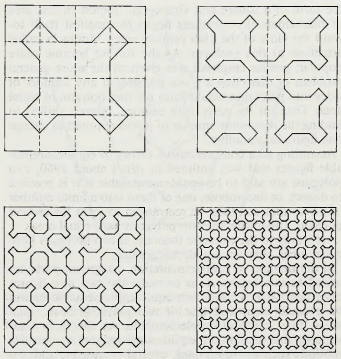


Figure 10: The Sierpinski curve.

Mazes. A maze having only one entrance and one exit can be solved by placing one hand against either wall and keeping it there while traversing it; the exit can always be reached in this manner, although not necessarily by the shortest path. If the goal is within the labyrinth, the "hand-on-wall" method will also succeed, provided that there is no closed circuit; *i.e.*, a route that admits of complete traverse back to the beginning (Figure 11).

From Martin Gardner, *The Second Scientific American Book of Mathematical Puzzles and Diversions* (copyright © 1961 by Martin Gardner); reprinted by permission of Simon and Schuster, Inc.

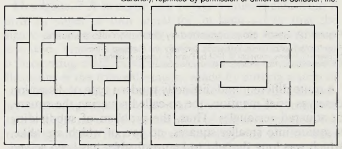


Figure 11: *Examples of mazes.*
(Left) "Simply connected" maze. (Right) "Multiply connected" maze (see text).

If there are no closed circuits—*i.e.*, no detached walls—the maze is "simply connected"; otherwise the maze is "multiply connected." A classic general method of "threading a maze" is to designate a place where there is a choice of turning as a node; a path or node that has not yet been entered as a "new" path or node; and one that has already been entered as an "old" path or node.

The procedure is as follows:

1. Never traverse a path more than twice.
2. When arriving at a new node, select either path.
3. When arriving at an old node or at a dead end by a new path, return by the same path.
4. When arriving at an old node by an old path, select a new path, if possible; otherwise, an old path.

Although recreational interest in mazes has diminished, two areas of modern science have found them to be of value: psychology and communications technology. The former is concerned with learning behaviour, the latter with improved design of computers.

Geometric dissections. Geometric dissection problems involve the cutting of geometric figures into pieces that can be arranged to form other geometric figures; for example, cutting a rectangle into parts that can be put together in

the form of a square and vice versa. Interest in this area of mathematical recreations began to manifest itself toward the close of the 18th century when Montucla called attention to this problem. As the subject became more popular, greater emphasis was given to the more general problem of dissecting a given polygon of any number of sides into parts that would form another polygon of equal area. Then, in the early 20th century, interest shifted to finding the *minimum* number of pieces required to change one figure into another.

Theory of equidecomposable figures

According to a comprehensive theory of equidecomposable figures that was outlined in detail about 1960, two polygons are said to be equidecomposable if it is possible to dissect, or decompose, one of them into a finite number of pieces that can then be rearranged to form the second polygon. Obviously, the two polygons have equal areas.

According to the converse theorem, if two polygons have equal areas, they are equidecomposable.

In the method of complementation, congruent parts are added to two figures so as to make the two new figures congruent. It is known that equicomplementable figures have equal areas and that, if two polygons have equal areas, they are equicomplementable. As the theory advanced, the relation of equidecomposability to various motions such as translations, central symmetry, and, indeed, to groups of motions in general, was explored. Studies were also extended to the more difficult questions of dissecting polyhedra.

On the "practical" side, the execution of a dissection, such as converting the Greek cross into a square (Figure 12), may require the use of ingenious procedures, some of which have been described by H. Lindgren (see *Bibliography*).

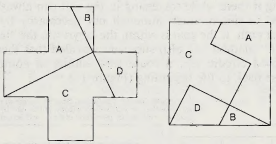


Figure 12: Greek cross converted by dissection into a square.
From *The Number of Things: Pythagoras, Geometry and Humming Strings* by Evans G. Valens, copyright © 1964 by Evans G. Valens; published by E.P. Dutton & Co., Inc., and used with their permission

A quite different and distinctly modern type of dissection deserves brief mention, the so-called squaring the square, or squared rectangles. Thus, the problem of subdividing a square into smaller squares, no two of which are alike, which was long thought to be unsolvable, has been solved by the means of network theory. In this connection, a squared rectangle is a rectangle that can be dissected into a finite number of squares; if no two of these squares are equal, the squared rectangle is said to be perfect. The order of a squared rectangle is the number of constituent squares. It is known that there are no perfect rectangles of orders less than 9, and that there are exactly two perfect rectangles of order 9. (One of these is shown as Figure 13.) The dissection of a square into unequal squares, deemed impossible as early as 1907, was first reported in 1939.

From Martin Gardner, *The Second Scientific American Book of Mathematical Puzzles and Diversions* (copyright © 1961 by Martin Gardner); reprinted by permission of Simon and Schuster, Inc.

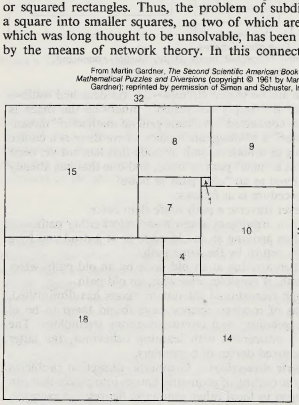


Figure 13: Squared rectangle (see text).

Graphs and networks. The word graph may refer to the familiar curves of analytic geometry and function theory, or it may refer to simple geometric figures consisting of points and lines connecting some of these points; the latter are sometimes called linear graphs, although there is little confusion within a given context. Such graphs have long been associated with puzzles.

If a finite number of points are connected by lines (Figure 14A), the resulting figure is a graph; the points, or corners, are called the vertices, and the lines are called the edges. If every pair of vertices is connected by an edge, the graph is called a complete graph (Figure 14B).

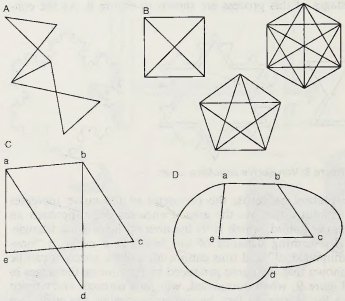


Figure 14: *Examples of linear graphs.*
(A) Graph. (B) Complete graphs. (C) Nonplanar graph. (D) Nonplanar graph of (C) changed to equivalent planar graph.

A planar graph is one in which the edges have no intersection or common points except at the edges. (It should be noted that the edges of a graph need not be straight lines.) Thus a nonplanar graph can be transformed into an equivalent, or isomorphic, planar graph, as in Figures 14C and 14D. An interesting puzzle involves the problem of the three wells. Here (Figure 15) A, B, and C represent three neighbours' houses, and R, S, and T three wells. It is desired to have paths leading from each house to each well, allowing no path to cross any other path. The proof that the problem is impossible depends on the so-called Jordan curve theorem that a continuous closed curve in a plane divides the plane into an interior and an exterior region in such a way that any continuous line connecting a point in the interior with a point in the exterior must intersect the curve. Planar graphs have proved useful in the design of electrical networks.

From *Graphs and Their Uses*, by Oystein Ore. Copyright © 1963 by Yale University. Reprinted by permission of Random House, Inc.

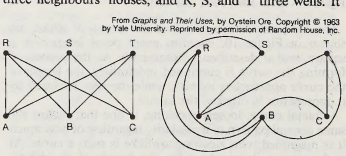


Figure 15: Three wells problem (see text).

Problem of the three wells

A connected graph is one in which every vertex, or point (or, in the case of a solid, a corner), is connected to every other point by an arc; an arc denotes an unbroken suc-

cession of edges. A route that never passes over an edge more than once, although it may pass through a point any number of times, is sometimes called a path.

Modern graph theory (in the sense of linear graphs) had its inception with the work of Euler in connection with the "Königsberg bridge problem" and was, for many years, associated with curves now called Eulerian paths—*i.e.,* figures that can be drawn without retracing edges or lifting the pencil from the paper. The city of Königsberg (now Kaliningrad) embraces the banks and an island of the forked Pregel (Pregolya) River; seven bridges span the different branches (see Figure 16A). The problem was: Could a person leave home, take a walk, and return, crossing each bridge just once? Euler showed why it is impossible.

Briefly stated, Euler's principles (which apply to any closed network) are as follows:

1. The number of even points—*i.e.,* those in which an even number of edges meet—is of no significance.
2. The number of odd points is always even; this includes the case of a network with only even points.
3. If there are *no odd points,* one can start at any point and finish at the same point.
4. If there are *exactly two odd points,* one can start at either of the odd points and finish at the other odd point.
5. If there are *more than two odd points,* the network cannot be traced in one continuous path; if there are $2n$ odd points and no more, it can be traced in n separate paths.

Thus Figures 16B and 16C can be traversed by Eulerian paths; Figures 16D and 16E cannot; Figure 16F shows a network corresponding to the Königsberg bridge problem, in which the points represent the land areas and the edges the seven bridges.

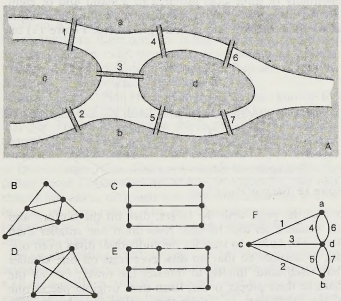


Figure 16: *Illustrations of Euler's principles.*
(A) Königsberg bridge problem. (B) and (C) Eulerian networks. (D) and (E) Non-Eulerian networks. (F) Network corresponding to Königsberg bridge problem.

Networks are related to a variety of recreational problems that involve combining or arranging points in a plane or in space. Among the earliest was a puzzle invented by an Irish mathematician, Sir William Rowan Hamilton (1859), which required finding a route along the edges of a regular dodecahedron that would pass once and only once through every point. In another version, the puzzle was made more convenient by replacing the dodecahedron by a graph isomorphic to the graph formed by the 30 edges of the dodecahedron (Figure 17). A Hamilton circuit is one that passes through each point exactly once but does not, in general, cover all the edges; actually, it covers only two of the three edges that intersect at each vertex. The route shown in heavy lines is one of several possible Hamilton circuits.

The Hamilton circuit

Graph theory lends itself to a variety of problems involving combinatorics: for example, designing a network to connect a set of cities by railroads or by telephone lines; planning city streets or traffic patterns; matching jobs with applicants; arranging round-robin tournaments such

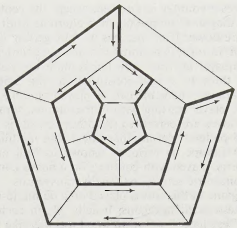


Figure 17: Hamilton circuit.
From Martin Gardner, *The Scientific American Book of Mathematical Puzzles and Diversions* (copyright © 1959 by Martin Gardner), reprinted by permission of Simon and Schuster, Inc.

that every team or individual meets every other team or individual. For a more formal treatment, see the section *Graph theory* in the article COMBINATORICS.

Map-colouring problems. Cartographers have long recognized that no more than four colours are needed to shade the regions on any map in such a way that adjoining regions are distinguished by colour. The corresponding mathematical question, framed in 1852, became the celebrated "four-colour map problem": Is it possible to construct a planar map for which five colours are necessary? Similar questions can be asked for other surfaces. For example, it was found by the end of the 19th century that seven colours, but no more, may be needed to colour a map on a torus. Meanwhile the classical four-colour question withstood mathematical assaults until 1976, when mathematicians at the University of Illinois announced that four colours suffice. Their published proof, including diagrams derived from more than 1,000 hours of calculations on a high-speed computer, was the first significant mathematical proof to rely heavily on artificial computation. Map-colouring problems are treated in greater detail in the article COMBINATORICS.

The four-colour problem

Flexagons. A flexagon is a polygon constructed from a strip of paper or thin metal foil in such a way that the figure possesses the property of changing its faces when it is flexed. First discussed in 1939, flexagons have become a fascinating mathematical recreation. One of the simplest flexagons is the trihexaflexagon, made by cutting a strip of suitable material and marking off 10 equilateral triangles. By folding appropriately several times and then gluing the last triangle onto the reverse side of the first triangle, the resulting model may be flexed so that one of the faces disappears and another face takes its place.

MANIPULATIVE RECREATIONS

Puzzles involving configurations. One of the earliest puzzles and games that require arranging counters into some specified alignment or configuration was Lucas' Puzzle: in a row of seven squares, each of the three squares at the left end is occupied by a black counter, each of the three squares at the right end is occupied by a white counter, and the centre square is vacant. The object is to move one counter at a time until the squares originally occupied by white counters are occupied by black, and vice versa; black counters can be moved only to the right and white only to the left. A counter may move to an adjacent vacant square or it may jump one counter of the other colour to occupy a vacant square. The puzzle may be enlarged to any number of counters of each colour. For n counters of each kind the number of required moves is $n(n + 2)$.

Lucas' Puzzle

A similar puzzle uses eight numbered counters placed on nine positions. The aim is to shift the counters so that they will appear in reverse numerical order; only single moves and jumps are permitted.

Well known, but by no means as trivial, are games for two players, such as Ticktacktoe and its more sophisticated variations, one of which calls for each player to begin with three counters (3 black, 3 white); the first

Ticktacktoe

player places a counter in any cell, except the center cell, of a 3×3 diagram; the players then alternate until all the counters are down. If neither has won by getting three in a row, each, in turn, is permitted to move a counter to an adjacent square, moving only horizontally or vertically. Achieving three in a row constitutes a win. There are many variations. The game can be played on a 4×4 diagram, each player starting with four counters; sometimes diagonal moves are permitted. Another version is played on a 5×5 pattern. Yet another interesting modification, popular in Europe, is variously known as Mill or Nine Men's Morris, played with counters on a board consisting of three concentric squares and eight transversals.

Another game of this sort is played on a diamond-shaped board of tessellated hexagons, usually 11 on each edge, where by "tessellated" we mean fitted together like tiles to cover the board completely. Two opposite edges of the diamond are designated "white"; the other two sides, "black." Each player has a supply of black or white counters. The players alternately place a piece on any vacant hexagon; the object of the game is for each player to complete an unbroken chain of his pieces between the sides designating his colour. Though the game does not end until one of the players has made a complete chain, it may meander across the board; it cannot end in a draw because the only way one player can block the other is by completing his own chain. The game was created by Piet Hein in 1942 in Denmark, where it quickly became popular under the name of Polygon. It was invented independently in the United States in 1948 by John Nash, and a few years later one version was marketed under the name of Hex.

In addition to the aforementioned varieties of a class of games that can be loosely described as "three in a row" or "specified alignment," many others also exist, such as three- and four-dimensional Ticktacktoe and even a computer Ticktacktoe. The game strategy in Ticktacktoe is by no means simple; an excellent mathematical analysis is given by F. Schuh (see *Bibliography*).

Chessboard problems. Recreational problems posed with regard to the conventional chessboard are legion. Among the most widely discussed is the problem of how to place eight queens on a chessboard in such a way that none of the queens is attacking any other queen; the problem interested the great German mathematician C.F. Gauss (*c.* 1850). Another group of problems has to do with the knight's tour; in particular, to find a closed knight's tour that ends at the starting point, that does not enter any square more than once, but that passes through all the squares in one tour. Problems of the knight's tour are intimately connected with the construction of magic squares. Other chessboard problems are concerned with determining the relative values of the various chess pieces; finding the maximum number of pieces of any one type that can be put on a board so that no one piece can take any other; finding the minimum number of pieces of any one type that can be put on a board so as to command all cells; and how to place 16 queens on a board so that no three of them are in a straight line.

The Fifteen Puzzle. One of the best known of all puzzles is the Fifteen Puzzle, which was invented by Sam Loyd the elder about 1878. It is also known as the Boss Puzzle, Jeu de Taquin, and Diablotin. It became popular all over Europe almost at once. It consists essentially of a shallow square tray that holds 15 small square counters numbered from 1 to 15, and one square blank space. With the 15 squares initially placed in random order and with the blank space in the lower right-hand corner, the puzzle is to rearrange them in numerical order by sliding only, with the blank space ending up back in the lower right-hand corner. It may overwhelm the reader to learn that there are more than 20,000,000,000,000 possible different arrangements that the pieces (including the blank space) can assume. But in 1879 two American mathematicians proved that only one-half of all possible initial arrangements, or about 10,000,000,000,000, admitted of a solution. The mathematical analysis is as follows. Basically, no matter what path it takes, as long as it ends its journey in the lower right-hand corner of the tray, any numeral must pass through an even number of boxes. In the normal position of the squares (Figure 18A), regarded row by row from left to right, each number is larger than all the preceding numbers; *i.e.*, no number *precedes* any number smaller than itself. In any other than the normal arrangement, one or more numbers will precede others smaller than themselves. Every such instance is called an inversion. For example, in the sequence 9, 5, 3, 4, the 9 precedes three numbers smaller than itself and the 5 precedes two numbers smaller than itself, making a total of five inversions. If the total number of *all* the inversions in a given arrangement is *even*, the puzzle can be solved by bringing the squares back to the normal arrangement; if the total number of inversions is *odd*, the puzzle cannot be solved. Thus, in Figure 18B there are two inversions,



Figure 18: (A) Fifteen Puzzle with no inversions. (B) With two inversions. (C) With five inversions.

and the puzzle can be solved; in Figure 18C there are five inversions, and the puzzle has no solution. Theoretically, the puzzle can be extended to a tray of $m \times n$ spaces with $(mn - 1)$ numbered counters.

The Tower of Hanoi. The puzzle of the Tower of Hanoi is believed to have been originated in 1883 by Lucas, under the name of M. Claus. Ever popular, made of wood or plastic, it still can be found in toy shops. It consists essentially of three pegs fastened to a stand and of eight circular disks, each having a hole in the centre. The disks, all of different radii, are initially placed (see Figure 19) on

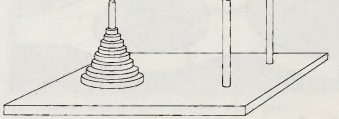


Figure 19: Tower of Hanoi.

one of the pegs, with the largest disk on the bottom and the smallest on top; no disk rests upon one smaller than itself. The task is to transfer the individual disks from one peg to another so that no disk ever rests on one smaller than itself, and, finally, to transfer the tower; *i.e.*, all the disks in their proper order, from their original peg to one of the other pegs. It can be shown that for a tower of n disks, there will be required $2^n - 1$ transfers of individual disks to shift the tower completely to another peg. Thus for 8 disks, the puzzle requires $2^8 - 1$, or 255 transfers. If the original "needle" (peg) was a tower with 64 disks, the number of transfers would be $2^{64} - 1$, or 18,446,744,073,709,551,615; this is exactly the same number required to fill an 8×8 checkerboard with grains of wheat, 1 on the first square, 2 on the second, 4 on the next, then 8, 16, 32, etc.

Polyominoes. The term polyomino was introduced in 1953 as a jocular extension of the word domino. A polyomino is a simply connected set of equal-sized squares, each joined to at least one other along an edge. The simpler polyomino shapes are shown in Figure 20A. Somewhat more fascinating are the pentominoes, of which there are exactly 12 forms (Figure 20B). Asymmetrical pieces, which have different shapes when they are flipped over, are counted as one.

The number of distinct polyominoes of any order is a function of the number of squares in each, but, as yet, no general formula has been found. It has been shown that there are 35 types of hexominoes and 108 types of heptominoes, if the dubious heptomino with an interior "hole" is included.

Solving the Fifteen Puzzle

Manipulating squares

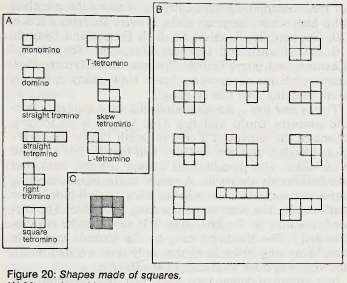


Figure 20: *Shapes made of squares.*
(A) Monomino with simple polyominoes. (B) Pentominoes.
(C) Heptomino with interior "hole."

From Martin Gardner, *The Scientific American Book of Mathematical Puzzles and Diversions* (copyright © 1959 by Martin Gardner); reprinted by permission of Simon and Schuster, Inc.

Recreations with polyominoes include a wide variety of problems in combinatorial geometry, such as forming desired shapes and specified designs, covering a chessboard with polyominoes in accordance with prescribed conditions, etc. Two illustrations may suffice.

The 35 hexominoes, having a total area of 210 squares, would seem to admit of arrangement into a rectangle 3×70, 5×42, 6×35, 7×30, 10×21, or 14×15; however, no such rectangle can be formed.

Can the 12 pentominoes, together with one square tetromino, form an 8×8 checkerboard? A solution of the problem was shown around 1935. It is not known how many solutions there are, but it has been estimated to be at least 1,000. In 1958, by use of a computer, it was shown that there are 65 solutions in which the square tetromino is exactly in the centre of the checkerboard.

Soma Cubes. Piet Hein of Denmark, also known for his invention of the mathematical games known as Hex and Tac Tix, stumbled upon the fact that all the irregular shapes that can be formed by combining three or four congruent cubes joined at their faces can be put together to form a larger cube. There are exactly seven such shapes, called Soma Cubes; they are shown in Figure 21. No two shapes are alike, although the fifth and sixth are mirror images of each other. The fact that these seven pieces (comprising 27 "unit" cubes) can be reassembled to form one large cube is indeed remarkable.

Many interesting solid shapes can be formed from the seven Soma Cubes, shapes resembling, for example, a sofa, a chair, a castle, a tunnel, a pyramid, and so on. Even the assembling of the seven basic pieces into a large cube can be done in more than 230 essentially different ways.

As a recreation, the Soma Cubes are fascinating. With experience, many persons find that they can solve Soma problems mentally. Psychologists who have used them find that the ability to solve Soma problems is roughly correlated with general intelligence, although there are some strange anomalies at both ends of the distribution of intelligence. In any event, people playing with the cubes do not appear to want to stop; the variety of interesting structures possible seems endless.

Coloured squares and cubes. There is a wide variety of puzzles involving coloured square tiles and coloured cubes. In one, the object is to arrange the 24 three-colour patterns, including repetitions, that can be obtained by subdividing square tiles diagonally, using three different colours, into a 4×6 rectangle so that each pair of touching edges is the same colour and the entire border of the rectangle is the same colour.

More widely known perhaps is the 30 Coloured Cubes Puzzle. If six colours are used to paint the faces there result 2,226 different combinations. If from this total only those cubes that bear all six colours on their faces are selected, a set of 30 different cubes is obtained; two cubes are regarded as "different" if they cannot be placed side by side so that all corresponding faces match. Many fascinating puzzles arise from these coloured squares and cubes; many more could be devised. Some of them have appeared commercially at various times under different names, such as the Mayblox Puzzle, the Tantalizer, and the Katzenjammer. Stimulating discussions may be found in books by MacMahon, Winter, and O'Beirne (see *Bibliography*).

A revival of interest in coloured-cube problems was aroused by the appearance of a puzzle known as Instant Insanity, consisting of four cubes, each of which has its faces painted white, red, green, and blue in a definite scheme. The puzzle is to assemble the cubes into a $1 \times 1 \times 4$ prism such that all four colours appear on each of the four long faces of the prism. Since each cube admits of 24 different orientations, there are 82,944 possible prismatic arrangements; of these only two are the required solutions.

This puzzle was soon superseded by Rubik's Cube, developed independently by Ernő Rubik (who obtained a Hungarian patent in 1975) and Terutoshi Ishigi (who obtained a Japanese patent in 1976). The cube appears to be composed of 27 smaller cubes, or cubelets; in its initial state, each of the six faces of the cube is made up of nine cubelet faces all of the same colour. In the commercial versions of the puzzle, an internal system of pivots allows any layer of nine cubelets to be rotated with respect to the rest, so that successive rotations about the three axes cause the cubelet faces to become scrambled. The challenge of restoring a scrambled cube to its original configuration

From M. Gardner, "A Game in Which Standard Pieces of Cubes Are Assembled into Larger Forms (Soma Cubes)." Copyright © 1958 by Scientific American, Inc. All rights reserved

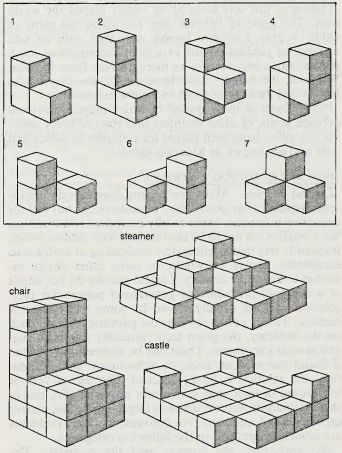


Figure 21: *Soma Cubes.*
(Top) The seven basic pieces. (Bottom) Examples of some of the shapes that can be built from Soma pieces.

is formidable, inasmuch as more than 10^{19} states can be reached from a given starting condition. A thriving literature quickly developed for the exposition of systematic solutions (based on group theory) of scrambled cubes.

Nim and similar games. A game so old that its origin is obscure, Nim lends itself nicely to mathematical analysis. In its generalized form, any number of objects (counters) are divided arbitrarily into several piles. Two people play alternately; each, in turn, selects any one of the piles and removes from it all the objects, or as many as he chooses, but at least one object. The player removing the last object wins. Every combination of the objects may be considered "safe" or "unsafe"; *i.e.*, if the position left by a player after his move assures a win for that player, the position

"Safe" and "unsafe" positions

is called safe. Every unsafe position can be made safe by an appropriate move, but every safe position is made unsafe by any move. To determine whether a position is safe or unsafe, the number of objects in each pile may be expressed in binary notation: if each column adds up to zero or an even number, the position is safe. For example, if at some stage of the game, three piles contain 4, 9, and 15 objects, the calculation is:

$$\begin{array}{rl} 4 \rightarrow & 1\,0\,0 \\ 9 \rightarrow & 1\,0\,0\,1 \\ 15 \rightarrow & 1\,1\,1\,1 \\ & 2\,2\,1\,2 \end{array}$$

Since the second column from the right adds up to 1, an odd number, the given combination is unsafe. A skillful player will always move so that every unsafe position left to him is changed to a safe position.

A similar game is played with just two piles; in each draw the player may take objects from either pile or from both piles, but in the latter event he must take the same number from each pile. The player taking the last counter is the winner.

Games such as Nim make considerable demands upon the player's ability to translate decimal numbers into binary numbers and vice versa. Since digital computers operate on the binary system, however, it is possible to program a computer (or build a special machine) that will play a perfect game. Such a machine was invented by E.U. Condon and an associate; their automatic Nimatron was exhibited at the New York World's Fair in 1940.

Games of this sort seem to be widely played the world over. The game of Pebbles, also known as the game of Odds, is played by two people who start with an odd number of pebbles placed in a pile. Taking turns, each player draws one, or two, or three pebbles from the pile. When all the pebbles have been drawn, the player who has an odd number of them in his possession wins.

Predecessors of these games, in which players distribute pebbles, seeds, or other counters into rows of holes under varying rules, have been played for centuries in Africa and Asia and are known as Mancala games.

PROBLEMS OF LOGICAL INFERENCE

Logical puzzles. Many challenging questions do not involve numerical or geometrical considerations but call for deductive inferences based chiefly on logical relationships. Such puzzles are not to be confounded with riddles, which frequently rely upon deliberately misleading or ambiguous statements, a play on words, or some other device intended to catch the unwary. Logical puzzles do not admit of a standard procedure or generalized pattern for their solution and are usually solved by some trial-and-error method. This is not to say that the guessing is haphazard; on the contrary, the given facts (generally minimal) suggest several hypotheses. These can be successively rejected if found inconsistent, until, by substitution and elimination, the solution is finally reached. The use of various techniques of logic may sometimes prove helpful, but in the last analysis, success depends largely upon that elusive capacity called ingenuity. For convenience, logic problems are arbitrarily grouped in the following categories.

Solution of logical problems

The brakeman, the fireman, and the engineer. The brakeman-fireman-engineer puzzle has become a classic. The following version of it appeared in O. Jacoby and W.H. Benson's *Mathematics for Pleasure* (1962).

The names, not necessarily respectively, of the brakeman, fireman, and engineer of a certain train were Smith, Jones, and Robinson. Three passengers on the train happened to have the same names and, in order to distinguish them from the railway employees, will be referred to hereafter as Mr. Smith, Mr. Jones, and Mr. Robinson. Mr. Robinson lived in Detroit; the brakeman lived halfway between Chicago and Detroit; Mr. Jones earned exactly $2,000 per year; Smith beat the fireman at billiards; the brakeman's next-door neighbour, one of the passengers, earned exactly three times as much as the brakeman; and the passenger who lived in Chicago had the same name as the brakeman. What was the name of the engineer?

Overlapping groups. The following problem is typical of the overlapping-groups category. Among the members of a high-school language club, 21 were studying French; 20, German; 26, Spanish; 12, both French and Spanish; 10, both French and German; nine, both Spanish and German; and three, French, Spanish, and German. How many club members were there? How many members were studying only one language?

Truths and lies. Another kind of logical inference puzzle concerns truths and lies. One variety is as follows: Some of the natives of a certain South Pacific island are pure-blooded and the rest are half-breeds, but they all look alike. The pure-blooded always tell the truth, but the half-breeds always lie. A visitor to the island, meeting three natives, asks them whether they are full-blooded or half-breeds. The first says something inaudible. The second, pointing to the first, says, "He says that he is pure-blooded." The third, pointing to the second, says, "He lies." Knowing beforehand that only one is a half-breed, the visitor decides what each of the three is.

In a slightly different type, four men, one of whom was known to have committed a certain crime, made the following statements when questioned by the police:

Archie: Dave did it.
Dave: Tony did it.
Gus: I didn't do it.
Tony: Dave lied when he said I did it.

If only one of these four statements is true, who was the guilty man? On the other hand, if only one of these four statements is false, who was the guilty man? (From *101 Puzzles in Thought and Logic* by C.R. Wylie, Jr.; Dover Publications, Inc., New York, 1957. Reprinted through the permission of the publisher.)

The smudged faces. The problem of the smudged faces is another instance of pure logical deduction. Three travellers were aboard a train that had just emerged from a tunnel, leaving a smudge of soot on the forehead of each. While they were laughing at each other, and before they could look into a mirror, a neighbouring passenger suggested that although no one of the three knew whether he himself was smudged, there was a way of finding out without using a mirror. He suggested: "Each of the three of you look at the other two; if you see at least one whose forehead is smudged, raise your hand." Each raised his hand at once. "Now," said the neighbour, "as soon as one of you knows for sure whether his own forehead is smudged or not, he should drop his hand, but not before." After a moment or two, one of the men dropped his hand with a smile of satisfaction, saying: "I know." How did that man know that his forehead was smudged?

The unexpected hanging. A final example might be the paradox of the unexpected hanging, a remarkable puzzle that first became known by word of mouth in the early 1940s. One form of the paradox is the following: A prisoner has been sentenced on Saturday. The judge announces that "the hanging will take place at noon on one of the seven days of next week, but you will not know which day it is until you are told on the morning of the day of the hanging." The prisoner, on mulling this over, decided that the judge's sentence could not possibly be carried out. "For example," said he, "I can't be hanged next Saturday, the last day of the week, because on Friday afternoon I'd still be alive and I'd know for sure that I'd be hanged on Saturday. But I'd known this *before* I was told about it on Saturday morning, and this would contradict the judge's statement." In the same way, he argued, they could not hang him on Friday, or Thursday, or Wednesday, Tuesday, or Monday. "And they can't hang me tomorrow," thought the prisoner, "because I know it today!"

Careful analysis reveals that this argument is false, and that the decree can be carried out. The paradox is a subtle one. The crucial point is that a statement about a future event can be known to be a true prediction by one person but not known to be true by another person until *after* the event has taken place.

Logical paradoxes. Highly amusing and often tantalizing, logical paradoxes generally lead to searching discussions of the foundations of mathematics. As early as the 6th century BC, the Cretan prophet Epimenides allegedly observed that "All Cretans are liars," which, in effect,

means that "All statements made by Cretans are false." Since Epimenides was a Cretan, the statement made by him is false. Thus the initial statement is self-contradictory. A similar dilemma was given by an English mathematician, P.E.B. Jourdain, in 1913, when he proposed the card paradox. This was a card on one side of which was printed:

"The sentence on the other side of this card is TRUE."

On the other side of the card the sentence read:

"The sentence on the other side of this card is FALSE."

The barber paradox

The barber paradox, offered by Bertrand Russell, was of the same sort: The only barber in the village declared that he shaved everyone in the village who did not shave himself. On the face of it, this is a perfectly innocent remark until it is asked "Who shaves the barber?" If he does not shave himself, then he is one of those in the village who does not shave himself and so is shaved by the barber, namely, himself. If he shaves himself, he is, of course, one of the people in the village who is not shaved by the barber. The self-contradiction lies in the fact that a statement is made about "all" the members of a certain class, when the statement or the object to which the statement refers is itself a member of the class. In short, the Russell paradox hinges on the distinction between those classes that are members of themselves and those that are not members of themselves. Russell attempted to resolve the paradox of the class of all classes by introducing the concept of a hierarchy of logical types but without much success. Indeed, the entire problem lies close to the philosophical foundations of mathematics.

BIBLIOGRAPHY

General works: W.W. ROUSE BALL and H.S.M. COXETER, *Mathematical Recreations and Essays*, 12th ed. (1974); JOHN H. CONWAY, *On Numbers and Games* (1976); HENRY E. DUDENEY, *536 Puzzles and Curious Problems*, ed. by MARTIN GARDNER (1967); KOBON FUJIMURA, *The Tokyo Puzzles*, trans. from the Japanese, ed. by MARTIN GARDNER (1978); MARTIN GARDNER, more than a dozen collections of mathematical recreations, including *Mathematical Circus* (1979, reissued 1981) and *Wheels, Life and Other Mathematical Amusements* (1983); DOUGLAS R. HOFSTADTER, *Gödel, Escher, Bach: An Eternal Golden Braid* (1979, reprinted 1980); J.A.H. HUNTER, *Challenging Mathematical Teasers* (1980); J.A.H. HUNTER and JOSEPH S. MADACHY, *Mathematical Diversions*, rev. ed. (1975); DAVID A. KLARNER, ed., *The Mathematical Gardner* (1981), a tribute to Martin Gardner; BORIS A. KORDEMSKY, *The Moscow Puzzles*, trans. from the Russian by ALBERT PARRY, ed. by MARTIN GARDNER (1972); MAURICE KRAITCHIK, *Mathematical Recreations*, 2nd rev. ed. (1953); JOSEPH S. MADACHY, *Madachy's Mathematical Recreations* (1979); T.H. O'BEIRNE, *Puzzles and Paradoxes* (1965, reprinted 1984); HUBERT PHILLIPS ("CALIBAN"), *Question Time: An Omnibus of Problems for a Brainy Day* (1938); FREDERIK SCHUH, *The Master Book of Mathematical Recreations*, trans. from the Dutch, ed. by T.H. O'BEIRNE (1968); HUGO STEINHAUS, *Mathematical Snapshots*, 3rd U.S. ed. (1983; originally published in Polish, 1954).

Books on special topics: (*Cube puzzles*): JOHN EWING and CZES KOSNIOWSKI, *Puzzle It Out: Cubes, Groups, and Puzzles* (1982); P.A. MACMAHON, *New Mathematical Pastimes* (1921); JAMES G. NOURSE, *The Simple Solution to Cubic Puzzles* (1981); DON TAYLOR and LEANNE RYLANDS, *Cube Games* (1981); FERDINAND WINTER, *Das Spiel der 30 Bunten Würfel* (1934). (*Dissections*): V.G. BOLTYANSKII, *Equivalent and Equidecomposable Figures* (1963; originally published in Russian, 1956); HARRY LINDGREN, *Geometric Dissections* (1964). (*Fallacies*): V.M. BRADIS, V.L. MINKOVSKII, and A.K. KHARCHEVA, *Lapses in Mathematical Reasoning* (1963; originally published in Russian, 2nd ed., 1959); EDWIN A. MAXWELL, *Fallacies in Mathematics* (1959, reprinted 1969). (*Fibonacci numbers*): VERNER E. HOGGATT, JR., *Fibonacci and Lucas Numbers* (1969). (*Fractals*): BENOIT B. MANDELBROT, *The Fractal Geometry of Nature*, rev. ed. (1983). (*Graphs*): OYSTEIN ORE, *Graphs and Their Uses* (1963). (*Logical inference*): MAXEY BROOKE, *150 Puzzles in Crypt-Arithmetic*, 2nd rev. ed. (1969); HUBERT PHILLIPS ("CALIBAN"), *My Best Puzzles in Logic and Reasoning* (1961); RAYMOND M. SMULLYAN, *What Is the Name of This Book? The Riddle of Dracula and Other Logical Puzzles* (1978), and *This Book Needs No Title* (1980); GEORGE J. SUMMERS, *Test Your Logic: 50 Puzzles in Deductive Reasoning* (1972); CLARENCE R. WYLIE, *101 Puzzles in Thought and Logic* (1957). (*Manipulative puzzles and games*): MAXEY BROOKE, *Fun for the Money* (1963); SOLOMON W. GOLOMB, *Polyominoes* (1965); RONALD C. READ, *Tangrams: 330 Puzzles* (1965); T. SUNDARA ROW, *Geometric Exercises in Paper Folding*, 2nd ed. (1905, reprinted 1966); SID SACKSON, *A Gamut of Games* (1969, reissued 1982). (*Mazes*): WALTER SHEPHERD, *Mazes and Labyrinths: A Book of Puzzles*, rev. ed. (1961). (*Polytopes*): H.S.M. COXETER, *Regular Polytopes*, 3rd ed. (1973); H. MARTYN CUNDY and A.P. ROLLETT, *Mathematical Models*, 2nd ed. (1961, reprinted 1977); L. FEJES TÓTH, *Regular Figures* (1964). (*Probability*): WARREN WEAVER, *Lady Luck: The Theory of Probability* (1963, reprinted 1982).

Periodicals: *Journal of Recreational Mathematics* (quarterly; United States); *Pythagoras* (4 issues a year; The Netherlands, England, and the United States).

(W.L.S.)

Number Theory

Number theory is a branch of mathematics that grew out of human curiosity concerning properties of the positive integers 1, 2, 3, 4, 5, . . . , also called whole numbers, natural numbers, or counting numbers. These numbers were the first mathematical creation of the human mind; interest in them is as old as civilization itself.

Number theory has always fascinated amateurs as well as professional mathematicians. In contrast to other branches of mathematics, many of the problems and theorems of number theory can be understood by laymen, although solutions to the problems and proofs of the theorems often require sophisticated mathematical background. The intellectual challenge of number theory's unsolved problems has attracted noted mathematicians from the time of Pythagoras (*c.* 540 BC). Work on these problems has often led to the creation of new branches of mathematics.

Until the mid-20th century, number theory was considered the purest branch of mathematics, with no direct applications to the real world. The advent of digital computers and digital communications revealed that number theory could provide unexpected answers to real-world problems. At the same time, improvements in computer technology enabled number theorists to make remarkable advances in factoring large numbers, determining primes, testing conjectures, and solving numerical problems once considered out of reach.

Modern number theory is a broad subject that encompasses topics only indirectly connected with the integers. It is classified into subheadings such as elementary number theory, algebraic number theory, analytic number theory, geometric number theory, and probabilistic number theory. These categories reflect the methods used to shed light on problems concerning the integers. For example, algebraic number theory makes extensive use of abstract algebra, while analytic number theory employs calculus and complex function theory. Most of this article deals with elementary number theory and can be read with profit by anyone with a good background in high school algebra. More advanced topics, such as algebraic and analytic number theory, are discussed in later sections.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 10/21 and 10/22, and the *Index.*

This article is divided into the following sections:

Elementary number theory 14
  Divisibility and prime numbers 14
  The greatest common divisor of two or more integers 15
  Perfect numbers and Mersenne primes 16
  Formulas that produce prime numbers 16
  Congruences and residue classes 17
  The Chinese remainder theorem 18
  Quadratic congruences and the quadratic reciprocity law 19
  Primitive roots 20
  Arithmetical functions 21
  Diophantine equations 22
  Applications of number theory 24
  Recreational number theory 24
  Approximating irrational numbers by rational numbers 26
  Algebraic and transcendental numbers 27
Algebraic number theory 28
  Algebraic integers and unique factorization 28
  Algebraic integers and Diophantine equations 28
  Ideals and class numbers of algebraic number fields 29
Analytic number theory 30
  The order of magnitude of arithmetical functions 30
  The prime number theorem 31
  Primes in arithmetic progressions 32
  Generating functions 32
  Additive number theory 33
  Lattice point problems 34
Unsolved problems 36
Bibliography 37

ELEMENTARY NUMBER THEORY

Divisibility and prime numbers. The integers consist of the natural numbers 1, 2, 3, 4, 5, . . . , together with their negatives $-1, -2, -3, -4, -5, \ldots$, and 0 (zero). Addition, subtraction, and multiplication of two integers always produces an integer. Division of two integers does not always produce an integer. For example, when 147 is divided by 10, the result is 14.7, which is not an integer. But, when 147 is divided by 7, the result is 21, an integer. In number theory 147 is said to be divisible by 7 but not by 10. More generally, an integer n is said to be divisible by an integer d if the quotient n/d is an integer c, so that $n = cd$. If so, n is called a multiple of d, and d is called a divisor, or factor, of n. This relation is expressed by the notation $d \mid n$, read "d divides n." The notation $d \nmid n$ means that d does not divide n. Thus, $7 \mid 147$, but $10 \nmid 147$.

Every positive integer is divisible by itself and 1. If an integer is greater than 1 and has only itself and 1 as positive divisors, it is called a prime.

The first few primes are 2, 3, 5, 7, 11, 13, 17, and 19. By mathematical convention, 1 is not a prime; this makes it possible to state propositions about primes without introducing qualifications. There are exactly 25 primes below 100. They can be found easily by the following "sifting" method, discovered by Eratosthenes of Cyrene around 200 BC: List the integers from 2 to 100 in increasing order. Strike out the even numbers greater than 2 (4, 6, 8, 10, . . .); they are divisible by 2 and therefore are not primes. Among the numbers that remain, strike out all multiples of 3 greater than 3 (9, 15, 21, 27, . . .); they are divisible by 3 and therefore are not primes. Continue the sifting process by crossing out multiples of 5 greater than 5, then multiples of 7 greater than 7. The numbers that remain are all the primes less than 100:

2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61, 67, 71, 73, 79, 83, 89, and 97.

Sieve of Eratosthenes

This sifting process, though laborious to do by hand, is highly efficient. It is not necessary to sift out multiples of 11 because the next such multiple not already crossed out is $11^2 = 121$, which is greater than 100. For the general case, when the sieve of Eratosthenes is used to find all the primes below a given integer n, only the multiples of primes smaller than $\sqrt{n}$ need be struck out.

All integers greater than 1 that are not primes are called composite numbers. A composite number is the product of two integers, each greater than 1. By repeated factorization, every composite number can be expressed as the product of primes, some of which may be repeated. For example, $147 = 7 \cdot 21 = 7 \cdot 7 \cdot 3$. Because every integer greater than 1 is a prime or a product of primes, the study of integers depends ultimately on properties of prime numbers.

Tables of primes reveal great irregularity in their distribution, from pairs of "twin" primes, differing only by 2, such as 3 and 5, 11 and 13, and 71 and 73, to primes separated by arbitrarily large gaps. A simple example shows why such gaps must occur. Let n be the product of all the positive integers from 1 to 100. The 99 consecutive integers $n + 2$, $n + 3$, . . . , $n + 100$ are composite because

Distribution of primes

they are divisible, respectively, by 2, 3, . . . , 100. The same type of argument, with 100 replaced by any positive integer, demonstrates that there are arbitrarily long strings of consecutive composite numbers and consequently arbitrarily large gaps between successive primes.

The distribution of primes appears to be more regular if they are counted in large blocks, say in groups of 1,000 consecutive integers. The first six blocks of length 1,000 contain, successively,

168, 135, 127, 120, 119, and 114

prime numbers. These figures suggest that the number of primes decreases in successive blocks of 1,000. But, as is often the case in number theory, the data can be misleading. The next four blocks contain

117, 107, 110, and 112

primes. Thus, the numbers of primes in successive blocks do not always decrease. But they do decrease on the average, with the 10,000th block (between 9,999,001 and 10,000,000) containing only 53 primes. The primes themselves become rarer and rarer among the large integers.

This observation raises the question of whether the primes terminate after some point. It has been known since antiquity that they do not. A proof is given in Euclid's *Elements* (Book IX, proposition 20) and is based on the idea that for any finite collection of primes, say

$$p_1, p_2, \ldots, p_k,$$

there is always another prime that is not one of these. To see why this is so, simply multiply the primes together and add 1 to form the number

$$N = 1 + p_1 p_2 \cdot \cdot \cdot p_k.$$

This number, being greater than 1, has at least one prime divisor; that divisor cannot be any of $p_1, p_2, \ldots, p_k$, because division of N by each of these primes leaves a remainder of 1. Therefore N must be divisible by a prime other than these (which may be N itself), so that the collection of all primes cannot be finite.

There are many other proofs of the infinitude of primes, but none is as elegant as the one in Euclid. For example, in 1845 the French mathematician Joseph Bertrand verified that for every integer n from 2 to 6,000,000 there is always at least one prime number between n and $2n$. The general proposition, without the 6,000,000 limitation on n, became known as Bertrand's postulate, and in 1851 the Russian mathematician Pafnuty Lvovich Chebyshev proved that it is true for all n. This result implies that there are infinitely many primes. It also implies that, if the primes are listed in increasing order as $p_1, p_2, \ldots$, then each prime is less than twice its predecessor: $p_{n+1} < 2p_n$.

A different type of proof makes use of the sum of the reciprocals of the primes. It can be shown that for every positive integer n there is a finite set of primes, the sum of whose reciprocals exceeds n. If there were only finitely many primes in all, say k of them, take n in the previous sentence to be k, and a contradiction follows at once. In the terminology of infinite series, the sum of the reciprocals of all the primes is a divergent series.

Evidence suggests that the number of twin primes is infinite, but this has not been proved or disproved. It is known that the sum of the reciprocals of all the twin primes cannot exceed 3. Whether their number is finite or infinite, twin primes are comparatively rare.

The ancient Greeks knew that every integer greater than 1 is a product of prime factors, but it was not until 1801 that Carl Friedrich Gauss stated and proved that this factorization can be done in only one way, apart from the order of the factors (*i.e.*, it does not matter whether the factorization of 21 is written as $7 \cdot 3$ or $3 \cdot 7$). Of course, some integers can be factored in different ways using composite factors. For example, $600 = 15 \cdot 40 = 12 \cdot 50 = 8 \cdot 75$. But the factorization into prime factors, $600 = 2 \cdot 2 \cdot 2 \cdot 3 \cdot 5 \cdot 5$, is unique (apart from the order of the factors). This basic result about unique factorization is known as the fundamental theorem of arithmetic. The primes are building blocks out of which every natural number greater than 1 can be obtained in one and only one way by multiplication.

Fundamental theorem of arithmetic

The factorization of a number into primes is often written as a product of powers of distinct primes. For example, $600 = 2^3 \cdot 3^1 \cdot 5^2$. In general, every integer n greater than 1 can be expressed as a product of the form

$$n = p^{\alpha} \cdot q^{\beta} \cdot \cdot \cdot,$$

where $p, q, \ldots$ are the distinct prime factors of n, and the exponents $A, B, \ldots$ indicate how many times the corresponding prime occurs in the factorization. Once this factorization is known, it is easy to determine all the positive divisors of n. Simply take all products of the form $p^{\alpha} \cdot q^{\beta} \cdot \cdot \cdot$, where the exponent α takes the values $0, 1, \ldots, A$, the exponent β takes the values $0, 1, \ldots, B$, and so on. For example, the positive divisors of 600 are all numbers of the form $2^{\alpha} \cdot 3^{\beta} \cdot 5^{\gamma}$, where α is 0, 1, 2, or 3, β is 0 or 1, and γ is 0, 1, or 2. In this case there are four choices for α, two choices for β, and three choices for γ, giving a total of $4 \cdot 2 \cdot 3 = 24$ positive divisors of 600 (including 1 and 600). This illustrates the general theorem that, if $n = p^{\alpha} \cdot q^{\beta} \cdot \cdot \cdot$, where $p, q, \ldots$ are distinct primes, then the total number of positive divisors of n is $(A + 1)(B + 1) \cdot \cdot \cdot$. In particular, if n is the product of k distinct primes, then the number of positive divisors of n is 2^k.

The greatest common divisor of two or more integers. Any two positive integers whose only common positive divisor is 1 are said to be relatively prime, or coprime. For example, the numbers $8 = 2^3$ and $9 = 3^2$ are relatively prime, as are

$$600 = 2^3 \cdot 3 \cdot 5^2 \text{ and } 77 = 7 \cdot 11,$$

because they have no prime factors in common. If two positive integers a and b have a common divisor greater than 1, then the prime factors of that common divisor must also divide both a and b. The largest positive common divisor of a and b is known as the greatest common divisor (GCD) of a and b and is denoted by the symbol (a, b). It can be obtained by multiplying together the common prime factors of a and b, each factor raised to the highest power to which it occurs in both a and b. For example, to find the GCD of 500 and 600, one can first determine that $500 = 2^2 \cdot 5^3$ and $600 = 2^3 \cdot 3 \cdot 5^2$, so that $(500, 600) = 2^2 \cdot 5^2 = 100$.

If the prime factorizations of a and b are known, their greatest common divisor is easily determined by the method just described. But there is a simpler method, called the Euclidean algorithm, that does not require prime factorization. If a is greater than b and b is not zero, division of a by b produces a quotient q and a remainder r, where r is nonnegative and less than b. This can be expressed in the form

The Euclidean algorithm

$$a = bq + r, \text{ where } 0 \le r < b.$$

Division of b by the remainder r produces a new quotient and remainder, and the process is repeated. Because the remainders form a decreasing sequence of nonnegative integers, they eventually reach zero. The last nonzero remainder is the GCD of a and b. For the case $a = 78{,}696$ and $b = 19{,}332$, there are six steps in the algorithm:

$$\begin{aligned} 78{,}696 &= 19{,}332 \cdot 4 + 1{,}368 \\ 19{,}332 &= 1{,}368 \cdot 14 + 180 \\ 1{,}368 &= 180 \cdot 7 + 108 \\ 180 &= 108 \cdot 1 + 72 \\ 108 &= 72 \cdot 1 + 36 \\ 72 &= 36 \cdot 2 + 0. \end{aligned}$$

The last nonzero remainder, 36, occurs in step five and is the GCD of 78,696 and 19,332. If the two integers a and b are relatively prime, the last nonzero remainder in the algorithm is 1.

The Euclidean algorithm also shows that the GCD can be written as a linear combination of a and b; that is,

$$(a, b) = ax + by,$$

where x and y are integers. For the above example, one choice of x and y can be obtained as follows. Write the next to last equation as

$$36 = 108 - 72 \cdot 1.$$

This expresses the GCD as a linear combination of the previous quotient, 108, and the previous remainder, 72. Use the next equation up to replace 72 by $180 - 108 \cdot 1$ and obtain

$$\begin{aligned} 36 &= 108 - (180 - 108 \cdot 1) \\ &= 2 \cdot 108 - 1 \cdot 180, \end{aligned}$$

a linear combination of the quotient and remainder in step three. Use the next equation up to eliminate 108 and obtain

$$36 = 2 \cdot 1{,}368 - 15 \cdot 180.$$

Continue eliminating remainders in this way, working back up the list of equations in the algorithm to obtain finally

$$36 = 212 \cdot 78{,}696 - 863 \cdot 19{,}332.$$

This has the form $36 = ax + by$, with $x = 212$ and $y = -863$ representing one choice of x and y.

The Euclidean algorithm is described in Euclid's *Elements* (Book VII, proposition 2). The process, which was probably known much earlier by the Pythagoreans, is efficient and easy to use and has become a basic tool in number theory and modern algebra. The French mathematician Gabriel Lamé (1795–1870) showed that the number of divisions in the Euclidean algorithm never exceeds five times the number of digits in the smaller number.

The GCD of more than two integers can be found by repeated use of the Euclidean algorithm. For three positive integers a, b, c, first find the GCD of two of them, say $D = (a, b)$, and then find the GCD (D, c). This will be the GCD of all three numbers a, b, c, denoted (a, b, c). The result is independent of which two integers are selected first and is a linear combination of the three integers:

$$(a, b, c) = ax + by + cz$$

for some integers x, y, z. The same process works for four or more integers. For example, to find the GCD of a, b, c, d, let

$$D_1 = (a, b),\ D_2 = (D_1, c), \text{ and } D_3 = (D_2, d).$$

Then D_3 is the GCD of a, b, c, d, denoted (a, b, c, d). Moreover, there exist integers x, y, z, and w, such that

$$(a, b, c, d) = ax + by + cz + dw.$$

Perfect numbers and Mersenne primes. The true origin of number theory is usually attributed to the Pythagoreans, who were the first to classify integers in various ways, such as even, odd, prime, and composite. The Pythagoreans called 6 a perfect number because $6 = 1 + 2 + 3$, the sum of all its proper divisors (that is, the sum of all divisors less than 6). The next perfect number is 28, whose proper divisors are 1, 2, 4, 7, and 14. The Greeks referred to the proper divisors of a number as its "parts," and they considered a number perfect if it was equal to the sum of all its parts. They also proposed the problem of determining all perfect numbers, which is the oldest unsolved problem in number theory. In Euclid's *Elements* (Book IX, proposition 36) it is shown that an even number is perfect if (in modern notation) it has the form $2^{p-1}(2^p - 1)$, where both p and $2^p - 1$ are primes. In 1757 the Swiss mathematician Leonhard Euler proved that, conversely, every even perfect number must be of Euclid's type. For example, when $p = 2, 3, 5$, and 7, the quantity $(2^p - 1)$ has the values 3, 7, 31, and 127, all of which are primes, and Euclid's formula gives the perfect numbers 6, 28, 496, and 8,128; these are the first four perfect numbers. For $p = 11$, the quantity $2^p - 1$ is composite and $2^{p-1}(2^p - 1)$ is not perfect. The next perfect number, corresponding to $p = 13$, is 33,550,336, discovered in the 5th century AD.

Numbers of the form $2^n - 1$ are called Mersenne numbers (denoted M_n) after the French mathematician Marin Mersenne, who studied them in 1644. If M_n is prime, it is easy to show that n itself must be prime. (The converse is not true, because $M_{11} = 2{,}047 = 23 \cdot 89$, a composite number.) If a Mersenne number is a prime, it is called a Mersenne prime. Mersenne numbers are of interest because each new Mersenne prime leads to a new perfect number. In 1750 Euler showed that M_{31} is prime, and in 1876 the French mathematician Édouard Lucas developed a method for testing whether M_p is prime and used it to prove that M_{127} is prime.

The Lucas primality test, as refined by the American mathematician Derrick Henry Lehmer in 1930, involves a sequence of numbers $S_2 = 4,\ S_3 = 14,\ S_4 = 194, \ldots$, where for $n > 2$ the S_n are defined successively by the formula $S_n = S_{n-1}^2 - 2$. The Lucas–Lehmer test states that a Mersenne number M_p with $p > 2$ is prime if M_p divides S_p but is composite otherwise. Lucas showed that S_{127} is divisible by M_{127} and deduced that M_{127} is prime. This was a formidable computational task because the numbers S_{127} and M_{127} are quite large, the latter having 39 digits. Nevertheless, the work was manageable because divisibility by M_{127} could be determined without actually computing S_{127}.

Lucas–Lehmer test for primes

No further Mersenne primes were found until the mid-20th century, when the advent of high-speed electronic computing machines working in the binary scale made divisibility by numbers of the form $2^p - 1$ relatively simple. In a series of investigations begun by D.H. Lehmer and Raphael M. Robinson and continued by others, the Lucas–Lehmer test was used to discover many new Mersenne primes. In 1978 two American high school students, Laura Nickel and Curt Noll, found the 25th and 26th Mersenne primes, requiring 440 hours on a large computer. By mid-1989, 31 perfect numbers were known, corresponding to the following values of p in Euclid's formula:

$p =$ 2, 3, 5, 7, 13, 17, 19, 31, 61, 89, 107, 127, 521, 607, 1,279, 2,203, 2,281, 3,217, 4,253, 4,423, 9,689, 9,941, 11,213, 19,937, 21,701, 23,209, 44,497, 86,243, 110,503, 132,049, and 216,091.

The largest known Mersenne prime was also the largest known prime number in mid-1989.

No odd perfect numbers are known, and it is not known if any exist. But if any do exist they must be larger than 10^{160} and must have at least eight distinct prime factors. It has been conjectured, but not proved, that the number of Mersenne primes is infinite. Thus, the Pythagorean problem of determining all perfect numbers remains unsolved.

Formulas that produce prime numbers. The irregularity in the distribution of primes suggests that there is no simple formula for producing all the primes. Some formulas do yield many primes. For example, the quadratic polynomial $x^2 - x + 41$, introduced by Euler in 1772, gives a prime for each of the 81 integer values of x from -40 to 40. However, in 1752 the German-Russian mathematician Christian Goldbach proved that no polynomial with integer coefficients can give a prime for all integer values of x, even if cubes and higher powers are used.

Some polynomials represent infinitely many primes. The linear polynomial $2x + 1$, where x is any positive integer, gives all the odd numbers, hence infinitely many primes. Each of the linear polynomials $4x + 1$ and $4x + 3$ also represents an infinite number of primes. In 1837 the German mathematician Peter Gustav Lejeune Dirichlet proved that, if h and k are relatively prime positive integers, the linear polynomial $kx + h$ gives infinitely many primes as x runs through the positive integers. This result is known as Dirichlet's theorem on the existence of primes in a given arithmetic progression. To prove this theorem, Dirichlet went outside the discrete realm of integers and appealed to continuity, using tools of calculus such as limits and infinite series. He thereby laid the foundations for the branch of mathematics called analytic number theory, in which methods of real and complex mathematical analysis are applied to problems about the integers (see below *Analytic number theory*).

It is not known whether there is any quadratic polynomial $ax^2 + bx + c$ with $a \neq 0$ that represents infinitely many primes. But Dirichlet proved that, if a, $2b$, and c are relatively prime positive integers with $b^2 \neq ac$, then the quadratic polynomial in two variables $ax^2 + 2bxy + cy^2$ represents infinitely many primes as x and y run through the positive integers.

Although no simple formula gives all the primes, it is known that there is a polynomial in 10 variables with integer coefficients such that the set of primes is equal to

the set of positive values of this polynomial obtained as the variables run through all nonnegative integers.

At first glance it seems likely that primes could be produced by taking a high power of a fixed integer k and adding or subtracting 1. It is easy to show that, if $k^m - 1$ is prime, then $k = 2$ and m is prime, so the problem is reduced to that of Mersenne primes. But, for $k^m + 1$ to be prime, k must be even and the exponent m must be a power of 2. Numbers of the form $2^{2^n} + 1$ were studied by the 17th-century French mathematician Pierre de Fermat, who conjectured that they would always give a prime for $n = 0, 1, 2, \ldots$. These numbers are called Fermat numbers and are denoted by F_n. The first five are

Fermat numbers

$$F_0 = 3,\ F_1 = 5,\ F_2 = 17,\ F_3 = 257,\ \text{and}\ F_4 = 65{,}537,$$

and all of these are primes. But in 1732 Euler found that F_5 is a composite number. This is easily verified on a hand calculator that displays 10 or more digits, because

$$F_5 = 2^{32} + 1 = 4{,}294{,}967{,}297 = 641 \cdot 6{,}700{,}417.$$

Beyond F_4, no further Fermat primes have been found. In fact, F_n is composite for all n from 5 to 20 and for many higher values of n, although for some composite Fermat numbers not all the factors are known. Euler knew that if F_n is composite, it has a factor of the form $k \cdot 2^{n+2} + 1$ for some k. This is how he discovered the factor $641 = 5 \cdot 2^7 + 1$ in F_5. Similarly, the large Fermat number $F_{3,310}$, which has more than 10^{990} digits, is known to be composite because it has the factor $5 \cdot 2^{3,312} + 1$. Fermat's conjecture that F_n is always prime is clearly false. Some mathematicians think there may be no further Fermat primes beyond F_4, or perhaps a finite number at most.

The Fermat numbers have an interesting connection with the classical problem of constructing a regular polygon by Euclidean methods (*i.e.*, by the use of an unmarked straightedge and compass). In 1796 Gauss, at the age of 19, solved a problem that had perplexed professional and amateur mathematicians for 2,000 years. He proved that a circle could be divided into n equal parts with a straightedge and compass if and only if n is the product of distinct Fermat primes and a nonnegative power of 2. For example, regular polygons with n sides can be constructed with a straightedge and compass for $n = 3, 4, 5, 6, 8, 10, 12, 15, 16, 17, 20, 24, 30, \ldots$ but not for $n = 7, 9, 11, 13, 14, 18, 19, 21, 22, 23, 25, \ldots$. This spectacular accomplishment inspired the young Gauss to become a mathematician rather than a philologist.

Formulas have been contrived to produce all the primes, but they usually involve numbers that cannot be determined without previous knowledge of the primes themselves. One example involves a positive number A less than 1 whose decimal representation is

$$A = 0.2003000050000007000000011\ldots.$$

The nonzero digits in A are the primes listed in increasing order. Enough zeros are inserted between successive primes so that the last digit of the nth prime p_n occurs at the decimal place at position n^2. Thus, 2 occurs at the first decimal place, 3 at the fourth, 5 at the ninth, and 7 at the 16th place. The number of decimal places between the prime p_n and its predecessor p_{n-1} is the difference $n^2 - (n-1)^2 = 2n - 1$. Each prime p_n can be recovered from the number A by using the formula

$$p_n = [10^{n^2}A] - 10^{2n-1}[10^{(n-1)^2}A],$$

where the square bracket symbol $[x]$ denotes the largest integer contained in x. For example, when $n = 2$, the formula states that

$$p_2 = [10^4 A] - 10^3[10A].$$

Now $10^4 A = 2{,}003.00005\ldots$, hence $[10^4 A] = 2{,}003$. Similarly, $[10A] = 2$, so $10^3[10A] = 2{,}000$, and the formula gives $p_2 = 2{,}003 - 2{,}000 = 3$.

Although the formula expresses p_n in terms of A, it is useless for determining primes, because to calculate the quantity $10^{n^2}A$ in the square bracket it is necessary to know the first n^2 decimal places of A, and this requires knowing the primes $p_1, p_2, \ldots, p_n$. Formulas such as these are regarded as amusing but of little value.

Congruences and residue classes. If a positive integer m divides the difference $a - b$, then $a = b + qm$ for some integer q. Gauss modified this equation by writing

$$a \equiv b \pmod{m}$$

(read "a is congruent to b modulo m"). This is called congruence notation and is an alternative way of writing $m|(a-b)$. The divisor m is called the modulus and is assumed to be positive, but the integers a and b can be positive, negative, or zero. Examples of congruences are

Congruence notation

$$17 \equiv 5 \pmod{12},\ -47 \equiv 2 \pmod{7},\ \text{and}\ 120 \equiv 0 \pmod{5}.$$

A familiar application is in clock arithmetic, where the 24 hours in a day are represented by the numbers $1, 2, \ldots, 12$. The hours beyond 12 noon are given by the remainders obtained on division by 12. Thus,

$$13 \equiv 1,\ 17 \equiv 5,\ \text{and}\ 21 \equiv 9 \pmod{12}.$$

Gauss's congruence notation made a significant impact on the development of number theory. One reason is that congruences share many properties with equations. For example, every number is congruent to itself, $a \equiv a \pmod{m}$, and, if $a \equiv b \pmod{m}$, then $b \equiv a \pmod{m}$. Also, if $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$, then $a \equiv c \pmod{m}$. For a given modulus, congruences can be added, subtracted, and multiplied as though they were equations. Thus, if $a \equiv b \pmod{m}$ and $c \equiv d \pmod{m}$, then

$$a + c \equiv b + d,\ a - c \equiv b - d,\ \text{and}\ ac \equiv bd \pmod{m}.$$

For example, to verify the result for the sum, write

$$(a + c) - (b + d) = (a - b) + (c - d)$$

and note that m divides both $a - b$ and $c - d$. The conclusion for the product follows from the identity $ac - bd = a(c - d) + d(a - b)$, because both $a(c - d)$ and $d(a - b)$ are divisible by m.

Common nonzero factors cannot always be canceled from both sides of a congruence as they can in equations. Both sides of the congruence $42 \equiv 12 \pmod{15}$ are divisible by 6, but when the common factor 6 is canceled, the result, $7 \equiv 2 \pmod{15}$, is incorrect. The cancellation law for congruences states that $ac \equiv bc \pmod{m}$ implies $a \equiv b \pmod{m/d}$, where $d = (c, m)$, the GCD of c and m. In particular, common factors relatively prime to the modulus can always be canceled: $ac \equiv bc \pmod{m}$ implies $a \equiv b \pmod{m}$ if $(c, m) = 1$. In the congruence $42 \equiv 12 \pmod{15}$ the common factor 2 is relatively prime to the modulus and can be canceled, giving a correct congruence $21 \equiv 6 \pmod{15}$. When the common factor 3 is canceled, the modulus must also be divided by 3 because $(3, 15) = 3$. Thus, canceling the common factor 3 in the congruence $42 \equiv 12 \pmod{15}$ gives $14 \equiv 4 \pmod{5}$.

Congruences place special emphasis on remainders. If a and b leave the same remainder when divided by m, then $a \equiv b \pmod{m}$. Numbers that are congruent mod m are said to belong to the same residue class mod m. There are exactly m distinct residue classes, each of which contains exactly one of the remainders $0, 1, 2, \ldots, m - 1$. For example, the residue class that contains 0 also contains all multiples of m. This is called the residue class 0 mod m and can be displayed as follows:

Residue classes

$$\text{residue class } 0 \bmod m = \{0, \pm m, \pm 2m, \pm 3m, \ldots\}.$$

Similarly, the residue classes 1 and 2 consist of all numbers congruent, respectively, to 1 and 2 mod m:

$$\text{residue class } 1 \bmod m = \{1, 1 \pm m, 1 \pm 2m, 1 \pm 3m, \ldots\},$$
$$\text{residue class } 2 \bmod m = \{2, 2 \pm m, 2 \pm 2m, 2 \pm 3m, \ldots\}.$$

The final entry in this list is

$$\text{residue class } m - 1 \bmod m = \{m - 1, m - 1 \pm m, m - 1 \pm 2m, m - 1 \pm 3m, \ldots\}.$$

A complete residue system mod m is a set of m integers, one selected from each of the m residue classes. For example, the six numbers $\{0, 1, 2, 3, 4, 5\}$ form a complete residue system mod 6, as do the numbers

$$\{0, 1, -1, 2, -2, 3\} \text{ or } \{60, -53, -32, 23, 63, 14\}.$$

In fact, any six integers will form a complete residue sys-

tem mod 6 provided no two of them are congruent mod 6.

Some problems in number theory require solving congruences containing an unknown integer x. For example, a solution of the congruence $2x \equiv 7 \pmod 4$ is any integer x that satisfies the congruence. This congruence has no solutions because $2x - 7$ is odd for every x and cannot be divisible by 4. But the congruence $2x \equiv 7 \pmod{13}$ does have solutions. Trying the numbers

$$0, 1, 2, \ldots, 12$$

in a complete residue system mod 13 shows that only one of these numbers, $x = 10$, is a solution. Therefore, all solutions are given by $x \equiv 10 \pmod{13}$, which means that any number in the residue class 10 mod 13 is a solution.

The same congruence can be solved algebraically without trying all possible candidates. In the linear algebraic equation $2x = 7$, simple division by 2 yields the solution. But the linear congruence $2x \equiv 7 \pmod{13}$ cannot be solved this way because 2 does not divide 7. However, 7 is congruent to some even number mod 13, for example, $7 \equiv 20 \pmod{13}$, so the original congruence is equivalent to $2x \equiv 20 \pmod{13}$. Division by 2 now gives $x \equiv 10 \pmod{13}$, the same set of solutions obtained by trial.

Linear congruences

In general, for given values of a, b, and m, a congruence of the form

$$ax \equiv b \pmod m$$

has a solution x if and only if b is divisible by (a, m), the GCD of the coefficient of x and the modulus. If (a, m) is denoted by d and if $d \mid b$, then the congruence has exactly d solutions in any complete residue system mod m. For example, the congruence

$$6x \equiv 9 \pmod{15}$$

has exactly three solutions in any complete residue system mod 15 because the GCD of 6 and 15 is 3, which divides 9. In fact, the solutions in this case are

$$x \equiv 4,\ x \equiv 9, \text{ and } x \equiv 14 \pmod{15}.$$

If the coefficient of x in a linear congruence

$$ax \equiv b \pmod m$$

is relatively prime to the modulus, then the congruence has exactly one solution in any complete residue system. In particular, when $b = 1$, the solution of the congruence $ax \equiv 1 \pmod m$ is called the reciprocal of a mod m and is often denoted by $\bar{a}$. If the reciprocal $\bar{a}$ is known, the more general linear congruence $ax \equiv b \pmod m$ can be solved by multiplying both members by $\bar{a}$ to obtain

$$x \equiv b\bar{a} \pmod m.$$

For example, the reciprocal of 2 mod 9 is 5 because $2 \cdot 5 \equiv 1 \pmod 9$. To solve the congruence

$$2x \equiv b \pmod 9,$$

multiply both members by 5 to obtain $x \equiv 5b \pmod 9$.

If $(a, m) = 1$, the reciprocal $\bar{a}$ can always be determined by testing all the numbers in a complete residue system to find which one satisfies $ax \equiv 1 \pmod m$. A more expeditious method was introduced by Euler and is based on the following idea. Starting with a complete residue system mod m, such as $0, 1, 2, \ldots, m-1$, all integers not relatively prime to m are deleted. What remains is called a reduced residue system and contains fewer than m integers. Euler introduced the symbol $\varphi(m)$ to denote the number of integers in a reduced residue system. For example, if $m = 14$ and if the multiples of 2 and 7 are removed, the reduced residue system consists of the six integers 1, 3, 5, 9, 11, and 13, so $\varphi(14) = 6$. Euler proved that, if $(a, m) = 1$, then $a^{\varphi(m)} \equiv 1 \pmod m$. Consequently, the reciprocal $\bar{a}$ is $a^{\varphi(m)-1}$, and the solution of the linear congruence $ax \equiv b \pmod m$ is $x \equiv ba^{\varphi(m)-1} \pmod m$.

If the modulus m is a prime p, then $\varphi(p) = p - 1$, and Euler's theorem states that $a^{p-1} \equiv 1 \pmod p$ if $(a, p) = 1$. Therefore $a^p \equiv a \pmod p$, and this holds whether or not a is relatively prime to p. The result $a^{p-1} \equiv 1 \pmod p$ for prime moduli was discovered by Fermat in 1640 and is called the little Fermat theorem. The first known proof seems to have been given by the German mathematician Gottfried Wilhelm Leibniz. Euler extended it in 1760 to any modulus by showing that

$$a^{\varphi(m)} \equiv 1 \pmod m \text{ if } (a, m) = 1.$$

An important congruence known as Wilson's theorem, named after the 18th-century English mathematician John Wilson, states that $(p-1)! \equiv -1 \pmod p$ for any prime p. (The symbol $n!$, read "n factorial," denotes the product of all the integers from 1 to n.) Wilson's theorem can be proved for odd primes by using reciprocals mod p. The numbers 1 and $p - 1$ are the only positive integers less than p that are their own reciprocals mod p. Therefore, in the product $(p-1)! = 1 \cdot 2 \cdot 3 \cdots (p-3)(p-2)(p-1)$, the integers from 2 to $p - 2$ can be grouped into pairs by pairing each integer with its reciprocal. The product of each pair is congruent to 1 (mod p), hence $(p-1)! \equiv p - 1 \equiv -1 \pmod p$.

Wilson's theorem

The converse of Wilson's theorem is also true; that is, if m is a positive integer greater than 1 such that $(m-1)! \equiv -1 \pmod m$, then m is prime. Although the converse of Wilson's theorem provides a test for determining whether an integer is prime, it is not a useful test because of the large number of multiplications required to calculate the residue of $(m-1)! \bmod m$.

The Chinese remainder theorem. The Chinese remainder theorem, sometimes called the Formosa theorem, was described in early Chinese manuscripts that posed problems such as the following: Find an integer that when divided by 3 has remainder 2, when divided by 5 has remainder 3, and when divided by 7 has remainder 4. In the language of congruences, the problem is to find an integer x that simultaneously satisfies the following three congruences:

$$x \equiv 2 \pmod 3,\ x \equiv 3 \pmod 5,\ x \equiv 4 \pmod 7.$$

This problem can be solved by considering the congruences individually. For the third congruence every solution is given by $x = 4 + 7k$, where k is an integer. The numbers of this form less than 105 are 4, 11, 18, 25, 32, 39, 46, 53, 60, 67, 74, 81, 88, 95, and 102. Of these, only 18, 53, and 88 satisfy the second congruence, and only 53 satisfies the first congruence. Thus, 53 is the only solution less than 105. This result can now be used to obtain all solutions. If x is any solution, then $x - 53$ is divisible by 3, 5, and 7 and also by their product, 105, so $x - 53 \equiv 0 \pmod{105}$. Therefore, all solutions are given by $x = 53 + 105t$, where t is any integer.

The existence of a solution in the foregoing example is guaranteed by the Chinese remainder theorem, which states that a system of congruences in different moduli

$$\begin{aligned} x &\equiv a_1 \pmod{m_1} \\ x &\equiv a_2 \pmod{m_2} \\ &\vdots \\ x &\equiv a_k \pmod{m_k} \end{aligned}$$

always has a solution if the moduli are relatively prime in pairs. Moreover, the solution is unique mod m, where $m = m_1 m_2 \cdots m_k$, the product of the moduli.

Several methods are available to determine the actual solution. One of these begins by solving k pairs of congruences:

$$y_i \equiv 1 \pmod{m_i}, \quad y_i \equiv 0 \pmod{q_i},$$

where $q_i = m/m_i$ and $i = 1, 2, \ldots, k$. The moduli m_i and q_i are relatively prime, so this pair of congruences has a unique solution mod m: $y_i \equiv t_i q_i \pmod m$, where t_i is the reciprocal of $q_i \pmod{m_i}$. The solution of the original system is then easily verified to be

$$x \equiv a_1 y_1 + \cdots + a_k y_k \pmod m.$$

To solve the problem given at the beginning of this section by this method, first solve the three pairs of congruences:

$$\begin{aligned} &y_1 \equiv 1 \pmod 3,\ y_1 \equiv 0 \pmod{35}, \\ &y_2 \equiv 1 \pmod 5,\ y_2 \equiv 0 \pmod{21}, \\ &y_3 \equiv 1 \pmod 7,\ y_3 \equiv 0 \pmod{15}. \end{aligned}$$

Solutions easily found by inspection are $y_1 = 70$, $y_2 = 21$, and $y_3 = 15$, so all solutions x are given by

$$x \equiv 2y_1 + 3y_2 + 4y_3 \equiv 263 \equiv 53 \pmod{105},$$

in agreement with the result found earlier.

The Chinese remainder theorem is applicable only when the moduli are relatively prime in pairs. If two or more moduli have a common factor greater than 1, a solution may or may not exist. An extension of the Chinese remainder theorem states that the general system will have a solution if and only if for every pair of moduli m_i and m_j the difference $a_i - a_j$ is divisible by the GCD (m_i, m_j). For example, the system

$$x \equiv 12 \pmod{20},\ x \equiv 3 \pmod{45},\ x \equiv 57 \pmod{72}$$

has no solution because $12 - 3$ is not divisible by the GCD (20, 45). But the system

$$x \equiv 13 \pmod{20},\ x \equiv 3 \pmod{45},\ x \equiv 57 \pmod{72}$$

does have a solution. To determine all solutions, use the first congruence to write $x = 13 + 20k$ and substitute this in the second congruence to obtain $13 + 20k \equiv 3 \pmod{45}$, or $20k \equiv -10 \pmod{45}$, which implies $4k \equiv -2 \pmod{9}$. Multiplication by 2 gives $k \equiv 4 \pmod{9}$, so that $k = 4 + 9s$ for some integer s, and $x = 13 + 20k = 13 + 20(4 + 9s) = 93 + 180s$. Now use this in the third congruence to obtain $93 + 180s \equiv 57 \pmod{72}$, or $180s \equiv -36 \pmod{72}$. Divide by 36 to obtain $s \equiv 1 \pmod{2}$, $s = 2t + 1$. This gives $x = 93 + 180(2t + 1) = 273 + 360t$, where t is any integer.

The solutions of the congruence $x \equiv a \pmod{m}$ can be written as $x = a + tm$, where t runs through all the integers. The solutions form an arithmetic progression in which consecutive terms have a common difference m, and they can be displayed as an infinite sequence:

$$\ldots, a - 3m, a - 2m, a - m, a, a + m, a + 2m, a + 3m, \ldots.$$

The question as to whether $x \equiv a \pmod{m}$ and $x \equiv b \pmod{n}$ have a simultaneous solution is the same as asking whether the progression

$$\ldots, b - 3n, b - 2n, b - n, b, b + n, b + 2n, b + 3n, \ldots$$

has a term in common with the preceding progression. The theory shows that these two sequences have a term in common if and only if $a - b$ is divisible by the GCD of the moduli m amd n. Furthermore, if there is one term in common, there are infinitely many terms in common.

The Chinese remainder theorem has become a useful tool in applications of number theory to other fields. A later section describes how it is used to increase efficiency in machine computation.

Quadratic congruences and the quadratic reciprocity law. Quadratic congruences of the form

$$Ax^2 + Bx + C \equiv 0 \pmod{m}$$

can be reduced to those of a simpler type, $x^2 \equiv a \pmod{p}$, where p is an odd prime and a is not divisible by p. These congruences may or may not have solutions, depending on the value of a. For example, the congruence $x^2 \equiv 3 \pmod{11}$ has two solutions, $x \equiv \pm 5 \pmod{11}$, whereas the congruence $x^2 \equiv 2 \pmod{11}$ has no solutions.

Throughout this section it is assumed that p is an odd prime and a is not divisible by p. If the congruence

$$x^2 \equiv a \pmod{p}$$

has a solution, then a is called a quadratic residue mod p. Otherwise, a is called a quadratic nonresidue mod p. For example, when $p = 11$ the quadratic residues can be found by squaring the numbers from 1 to 10 and reducing mod 11. Squaring the numbers from 1 to 5 gives

$$1^2 \equiv 1,\ 2^2 \equiv 4,\ 3^2 \equiv 9,\ 4^2 \equiv 5,\ 5^2 \equiv 3 \pmod{11}.$$

Squaring the numbers from 6 to 10 yields the same residues in reverse order:

$$6^2 \equiv (-5)^2 \equiv 3,\ 7^2 \equiv (-4)^2 \equiv 5,\ 8^2 \equiv (-3)^2 \equiv 9,$$
$$9^2 \equiv (-2)^2 \equiv 4,\ 10^2 \equiv (-1)^2 \equiv 1 \pmod{11}.$$

Thus, the quadratic residues mod 11 are 1, 3, 4, 5, 9, and the nonresidues are 2, 6, 7, 8, 10. In general, every reduced residue system mod p contains exactly $(p-1)/2$ quadratic residues and exactly $(p-1)/2$ nonresidues. The quadratic residues are obtained by squaring the numbers $1, 2, \ldots, (p-1)/2$ and reducing mod p. The quadratic congruence $x^2 \equiv a \pmod{p}$ has exactly two solutions if a is a quadratic residue mod p, and no solutions if a is a quadratic nonresidue mod p.

Quadratic reciprocity law

A remarkable theorem known as the quadratic reciprocity law connects pairs of quadratic congruences on the question of their solvability. If p and q are primes of the form $4n + 3$, such as 3 and 1,999, then one and only one of the congruences $x^2 \equiv p \pmod{q}$ and $x^2 \equiv q \pmod{p}$ is solvable. On the other hand, if one or both of the odd primes p and q is of the form $4n + 1$, and if one of the two congruences is solvable, so is the other. Thus, either both congruences are solvable or neither is.

For example, both primes $p = 1{,}999$ and $q = 3$ are of the form $4n + 3$. The first congruence,

$$x^2 \equiv 1{,}999 \pmod{3},$$

reduces to the solvable congruence $x^2 \equiv 1 \pmod{3}$, and therefore, by the reciprocity law, the second congruence, $x^2 \equiv 3 \pmod{1{,}999}$, has no solution. One could, of course, verify directly that this congruence has no solution by testing each of the numbers from 1 to 999, but the reciprocity law solves the problem without any need for testing.

If $q = 5$, the congruences $x^2 \equiv 1{,}999 \pmod{5}$ and $x^2 \equiv 5 \pmod{1{,}999}$ provide an example where at least one of the primes p and q is of the form $4n + 1$. The first congruence reduces to $x^2 \equiv 4 \pmod{5}$ and has the solutions $x \equiv \pm 2 \pmod{5}$. By the quadratic reciprocity law the other congruence, $x^2 \equiv 5 \pmod{1{,}999}$, must also be solvable. This guarantees that a search for a solution of the second congruence will not be in vain.

If a is a quadratic residue mod p, there are methods for finding a solution of $x^2 \equiv a \pmod{p}$ without squaring all the numbers from 1 to $(p-1)/2$. If p is of the form $4n + 3$, one solution is $x = a^{(p+1)/4}$. For example, the congruence $x^2 \equiv 5 \pmod{1{,}999}$ has the solution $x \equiv 5^{500} \pmod{1{,}999}$. With modern calculators the residues of the powers of 5 can be easily computed mod 1,999 to give $x \equiv 100 \pmod{1{,}999}$; the other solution is, of course, $x \equiv -100 \pmod{1{,}999}$.

The quadratic reciprocity law was first stated by Euler in the period 1744–46 and rediscovered in 1785 by Adrien-Marie Legendre, who provided a partial proof. In 1795 Gauss discovered the result independently (at the age of 18), and a year later he gave the first complete proof. Gauss was fascinated by the quadratic reciprocity law, which he called "the golden theorem of number theory," and produced at least six different proofs of it. The 19th-century mathematicians Augustin-Louis Cauchy, Karl Gustav Jacob Jacobi, and Leopold Kronecker later devised proofs of their own, as they wanted to determine for themselves why such a deep reciprocal relationship should exist between simple congruences linking any pair of odd primes.

In addition to the quadratic reciprocity law, there are supplementary results that are helpful in solving quadratic congruences. For example, if a and b are not divisible by the odd prime p, then at least one of the three congruences $x^2 \equiv a \pmod{p}$, $x^2 \equiv b \pmod{p}$, and $x^2 \equiv ab \pmod{p}$ is solvable, and, if more than one is solvable, all three must be solvable. This implies that the product of two quadratic residues mod p is a quadratic residue, as is the product of two nonresidues. However, the product of a residue and a nonresidue is a nonresidue. Another result states that $x^2 \equiv -1 \pmod{p}$ is solvable if p is of the form $4k + 1$ but not if p is of the form $4k + 3$. Also, the congruence $x^2 \equiv 2 \pmod{p}$ is solvable if p is of the form $8k + 1$ or $8k + 7$ but not if p is of the form $8k + 3$ or $8k + 5$.

Composite moduli and higher-degree congruences

If the modulus is not a prime, a quadratic congruence can have more than two solutions. For example, the congruence $x^2 \equiv 29 \pmod{35}$ has four solutions: $x \equiv 8, 13, 22,$ and $27 \pmod{35}$. The solutions are found in the following manner. Every solution x must satisfy both congruences $x^2 \equiv 29 \pmod{5}$ and $x^2 \equiv 29 \pmod{7}$. The first has two solutions, $x \equiv \pm 2 \pmod{5}$, and the second has two solutions, $x \equiv \pm 1 \pmod{7}$. The Chinese remainder theorem is now applied to each of the four combinations.

For example, the combination $x \equiv 2 \pmod 5$ and $x \equiv -1 \pmod 7$ gives

$$x \equiv 27 \pmod{35}.$$

In general, any solution x to the quadratic congruence $x^2 \equiv a \pmod m$, where m is a composite modulus, must also satisfy each of the congruences

$$x^2 \equiv a \pmod{p^r}$$

for every prime power p^r that divides m. And for any odd prime p the congruence $x^2 \equiv a \pmod{p^r}$ is solvable if and only if $x^2 \equiv a \pmod p$ is solvable.

The theory of higher-degree congruences is more complicated. One important result, discovered by the 18th-century French mathematician Joseph-Louis Lagrange, states that, if $f(x)$ is a polynomial of degree n with integer coefficients, then the congruence $f(x) \equiv 0 \pmod p$ has at most n solutions if the modulus p is prime. If the modulus is not prime, there may be more than n solutions. For example, the quadratic congruence $x^2 \equiv 29 \pmod{35}$ has four solutions, and the cubic congruence $x^3 - x \equiv 0 \pmod 6$ has six solutions. The problem of solving a polynomial congruence $f(x) \equiv 0 \pmod m$ with a composite modulus m can be reduced to that of solving a system of congruences $f(x) \equiv 0 \pmod{p^r}$ for every prime power p^r that divides m. Each of these congruences, in turn, can be reduced to a polynomial congruence mod p plus a set of linear congruences.

Primitive roots. The positive powers of an integer g form a geometric progression $g, g^2, g^3, \ldots$. If g is relatively prime to a positive integer m, then every member of this progression is also relatively prime to m. There is a smallest positive exponent f such that $g^f \equiv 1 \pmod m$. In fact, the Euler–Fermat theorem states that

$$g^{\varphi(m)} \equiv 1 \pmod m,$$

and hence the smallest exponent f cannot exceed $\varphi(m)$. If the smallest exponent f is equal to $\varphi(m)$, the integer g is called a primitive root mod m. In this case any two powers g^a and g^b are congruent (mod m) if, and only if, the exponents a and b differ by a multiple of $\varphi(m)$. Therefore, if g is a primitive root mod m, the $\varphi(m)$ numbers

$$g, g^2, \ldots, g^{\varphi(m)}$$

form a reduced residue system mod m.

Primitive roots are important because they permit a reduced residue system to be expressed as members of a geometric progression. Gauss proved that a modulus m has primitive roots if, and only if, m has one of the special values

$$m = 1, 2, 4, p^a, \text{ or } 2p^a,$$

where p is an odd prime and $a \geq 1$. For each such m the number of primitive roots mod m is exactly $\varphi(\varphi(m))$. For example, if $m = 17$, then $\varphi(m) = 16$ and $\varphi(\varphi(m)) = \varphi(16) = 8$, so 17 has eight primitive roots. Determining the primitive roots, when they exist, often requires a great deal of computation, especially for large moduli. The following table lists the smallest primitive root $g(p)$ for all primes p less than 50.

p	2	3	5	7	11	13	17	19	23	29	31	37	41	43	47
$g(p)$	1	2	2	3	2	2	3	2	5	2	3	2	6	3	5

The table shows, for example, that for $p = 17$ the smallest primitive root is $g = 3$. The powers

$$3, 3^2, \ldots, 3^{16},$$

when reduced mod 17, form a reduced residue system mod 17. In the following table, the first row contains the powers of 3, and the second row contains their residues mod 17.

3	3^2	3^3	3^4	3^5	3^6	3^7	3^8	3^9	3^{10}	3^{11}	3^{12}	3^{13}	3^{14}	3^{15}	3^{16}
3	9	10	13	5	15	11	16	14	8	7	4	12	2	6	1

Primitive roots provide a useful tool for attacking problems and proving theorems involving reduced residue systems. The following example illustrates how primitive roots can be used to give an alternate proof of Wilson's theorem, $(p-1)! \equiv -1 \pmod p$ for an odd prime p. If g is a primitive root mod p, then $\varphi(p) = p - 1$ and the numbers $g, g^2, \ldots, g^{p-1}$ are congruent, in some order, to the numbers $1, 2, \ldots, p-1$. Therefore their products are congruent mod p, and hence

$$\begin{aligned}(p-1)! &\equiv g \cdot g^2 \cdots g^{p-1} \pmod p \\ &\equiv g^{1+2+\cdots+(p-1)} \pmod p \\ &\equiv g^{p(p-1)/2} \pmod p.\end{aligned}$$

But $g^{(p-1)/2}$ satisfies the congruence $x^2 \equiv 1 \pmod p$ because of the Euler–Fermat theorem. This congruence is the same as $(x-1)(x+1) \equiv 0 \pmod p$, whose only solutions are $x \equiv \pm 1 \pmod p$. The number $g^{(p-1)/2}$ cannot be congruent to 1 (mod p) because the exponent $(p-1)/2$ is smaller than $\varphi(p)$. Therefore

$$g^{(p-1)/2} \equiv -1 \pmod p$$

and hence

$$(p-1)! \equiv g^{p(p-1)/2} \equiv (-1)^p \equiv -1 \pmod p$$

because p is odd. This proves Wilson's theorem.

Primitive roots are especially useful in the study of quadratic residues mod p. If g is a primitive root of an odd prime p, the even powers $g^2, g^4, \ldots, g^{p-1}$ are the quadratic residues mod p, and the odd powers $g, g^3, \ldots, g^{p-2}$ are the quadratic nonresidues mod p. Some properties of quadratic residues and nonresidues are easily deduced from this representation. For example, the product of two quadratic residues is also a quadratic residue, because the product of two even powers of g is another even power. On the other hand, the product of a quadratic residue with a nonresidue is a nonresidue because the product of an even power of g with an odd power is another odd power. Also, the product of two quadratic nonresidues is a quadratic residue, because the product of two odd powers of g is an even power.

Primitive roots can also be used to study congruences of higher degree of the form

$$x^n \equiv a \pmod m,$$

where a and m are relatively prime. This is called a binomial congruence of degree n. If m has a primitive root g then $a \equiv g^c \pmod m$ for some positive exponent c not exceeding $\varphi(m)$. Any solution x of the binomial congruence is congruent mod m to some power of g, say $x \equiv g^y \pmod m$ for some positive exponent y not exceeding $\varphi(m)$. Therefore the binomial congruence becomes

$$g^{ny} \equiv g^c \pmod m.$$

Because the two powers of g are congruent mod m, the exponents ny and c differ by a multiple of $\varphi(m)$. Therefore

$$ny \equiv c \pmod{\varphi(m)}.$$

This auxiliary linear congruence for y has a solution if, and only if, the GCD of n and $\varphi(m)$ divides c. For example, to study the binomial congruence

$$x^{10} \equiv 7 \pmod{17},$$

take $g = 3$, a primitive root mod 17, and write $x \equiv 3^y$ and $7 \equiv 3^{11} \pmod{17}$. The auxiliary linear congruence for y becomes

$$10y \equiv 11 \pmod{16}.$$

This has no solution because the GCD of 10 and 16 does not divide 11. Therefore the original binomial congruence $x^{10} \equiv 7 \pmod{17}$ has no solution. On the other hand, if the same method is applied to the binomial congruence $x^{10} \equiv 2 \pmod{17}$, then $2 \equiv 3^{14} \pmod{17}$ and the auxiliary linear congruence is $10y \equiv 14 \pmod{16}$ or $5y \equiv 7 \pmod 8$. This has the solution $y = 3$, giving the value

$$x \equiv 3^3 \equiv 10 \pmod{17}$$

as a solution of the binomial congruence of degree 10. In other words, $10^{10} \equiv 2 \pmod{17}$.

The same method can be used to study congruences of the form

$$a^x \equiv b \pmod{m},$$

where a and b are given integers relatively prime to m. These are called exponential congruences because the unknown x appears as an exponent. If m has a primitive root g, then a and b can be expressed as powers of $g \pmod m$, say $a \equiv g^c$, $b \equiv g^d \pmod m$, for some choice of positive exponents c and d not exceeding $\varphi(m)$. The exponential congruence $a^x \equiv b \pmod m$ now becomes

$$g^{cx} \equiv g^d \pmod m.$$

This implies that the exponents differ by a multiple of $\varphi(m)$, and hence

$$cx \equiv d \pmod{\varphi(m)}.$$

This auxiliary linear congruence has a solution if, and only if, the GCD of c and $\varphi(m)$ divides d. For example, to solve the exponential congruence

$$10^x \equiv 9 \pmod{17},$$

write $10 \equiv 3^3$, and $9 \equiv 3^2 \pmod{17}$, and the exponential congruence becomes

$$3^{3x} \equiv 3^2 \pmod{17}.$$

The auxiliary linear congruence is $3x \equiv 2 \pmod{16}$, which has the solution $x = 6$. Therefore, $x = 6$ is a solution of the exponential congruence, and hence

$$10^6 \equiv 9 \pmod{17}.$$

If 9 is replaced by any integer b, the exponential congruence $10^x \equiv b \pmod{17}$ always has a solution because in the auxiliary linear congruence,

$$3x \equiv d \pmod{16},$$

the numbers 3 and 16 are relatively prime and the congruence is always solvable.

When this method is used to study binomial or exponential congruences of the form $x^n \equiv a \pmod p$, or $a^x \equiv b \pmod p$, where p is a prime, the major difficulties are to find a primitive root $g \bmod p$ and to express the given integers a and b as powers of $g \pmod p$. The smallest nonnegative power k such that $a \equiv g^k \pmod p$ is called the index of a to the base $g \bmod p$. Extensive tables exist that give the smallest primitive root of many primes, together with a corresponding table of indices of all integers in a reduced residue system modulo these primes. A table of indices reduces the problem of solving a binomial or exponential congruence to the simpler problem of solving an auxiliary linear congruence.

The first large collection of tables of indices for all primes less than 1,000 was the *Canon arithmeticus,* computed under the direction of Jacobi and published in 1839. A revised and extended version was issued in 1956. Since then, high-speed computing machines have been used to extend the range of these tables. For example, one table lists all the primitive roots of the primes less than 5,000. Another lists the smallest primitive root for all primes up to 50,021.

Arithmetical functions. Functions $f(n)$ of the positive integer n that express some arithmetical property of n are called arithmetical functions. An important example is Euler's function $\varphi(n)$, which counts the number of integers in a reduced residue system mod n. Alternatively, $\varphi(n)$ is the number of positive integers less than n that are relatively prime to n, with the value $\varphi(1) = 1$ defined separately. The following examples are concerned with divisors of n:

$d(n)$ = the number of positive divisors of n,
$\sigma(n)$ = the sum of the positive divisors of n,
$\sigma_k(n)$ = the sum of the kth powers of the positive divisors of n.

The divisor function $\sigma_k(n)$ can be expressed in summation notation as follows:

$$\sigma_k(n) = \sum_{d|n} d^k,$$

where the sum is extended over all positive divisors of n. The functions $d(n)$ and $\sigma(n)$ are, of course, special cases of $\sigma_k(n)$ with $k = 0$ and $k = 1$, respectively.

All these examples have the two special properties:

$$f(1) = 1, \text{ and } f(mn) = f(m)f(n) \text{ whenever } (m, n) = 1.$$

Multiplicative arithmetical functions

Arithmetical functions with these properties are called multiplicative; they are completely determined by their values at the prime powers. Thus, if n is a product of powers of distinct primes, say $n = p^a \cdot q^b \cdot \cdots$, then $f(n) = f(p^a)f(q^b) \cdot \cdots$. For example,

$$\begin{aligned}\sigma_2(24) &= \sigma_2(3)\sigma_2(2^3)\\ &= (1^2 + 3^2)(1^2 + 2^2 + 4^2 + 8^2)\\ &= 10 \cdot 85\\ &= 850.\end{aligned}$$

Because of the multiplicative property, this calculation is simpler than adding the squares of all the positive divisors of 24.

The values of the foregoing functions at the prime powers are given by

$$\varphi(p^a) = p^a - p^{a-1},$$

$$d(p^a) = a + 1,$$

$$\sigma(p^a) = 1 + p + p^2 + \cdots + p^a = \frac{p^{a+1} - 1}{p - 1},$$

and

$$\sigma_k(p^a) = \frac{p^{ak+k} - 1}{p^k - 1} \text{ if } k \neq 0.$$

Any arithmetical function $f(n)$ can be used to generate another,

$$F(n) = \sum_{d|n} f(d),$$

by summing $f(d)$ over all positive divisors d of n. This can also be written as $\Sigma_{d|n} f(n/d)$, because as d runs through all positive divisors of n, from smallest to largest, n/d runs through the same divisors in opposite order, from largest to smallest. If the function $f(n)$ is multiplicative, so is the divisor sum $F(n)$, and conversely, if the divisor sum $F(n)$ is multiplicative, so is $f(n)$.

Möbius inversion formula

Any arithmetical function $f(n)$ can be determined from a knowledge of its divisor sum $F(n)$ by the Möbius inversion formula,

$$f(n) = \sum_{d|n} \mu(d) F\left(\frac{n}{d}\right),$$

where $\mu(n)$ is a multiplicative arithmetical function called the Möbius function, after the German mathematician August Ferdinand Möbius. Its values are 0, 1, and -1 and are obtained as follows:

$$\mu(1) = 1,$$
$$\mu(p) = -1 \text{ for each prime } p,$$
$$\mu(p^k) = 0 \text{ if } k > 1.$$

Because $\mu(n)$ is multiplicative, $\mu(n) = 0$ if n is divisible by the square of any prime, and $\mu(n) = (-1)^r$ if n is the product of r distinct primes. The divisor sum

$$\sum_{d|n} \mu(d)$$

is equal to 0 if $n > 1$ and is equal to 1 if $n = 1$.

The Möbius inversion formula often reveals hidden properties of arithmetical functions. For example, it can be used to prove that the Euler function $\varphi(n)$ is multiplicative, a result that is not immediately evident. Each integer k from 1 to n has a GCD $d = (k, n)$ with n; then k/d and n/d are relatively prime. Just as $\varphi(n)$ counts the number of integers from 1 to n that are relatively prime to n, so $\varphi(n/d)$ counts the number of integers $q = k/d$ from 1 to n/d that are relatively prime to n/d. Therefore $\varphi(n/d)$ is also the number of k for which $(k, n) = d$. But there are n values of k altogether, so

$$n = \sum_{d|n} \varphi\left(\frac{n}{d}\right) = \sum_{d|n} \varphi(d).$$

The Möbius inversion formula gives

$$\varphi(n)=\sum_{d|n}\mu(d)\frac{n}{d} \quad \text{or} \quad \frac{\varphi(n)}{n}=\sum_{d|n}\frac{\mu(d)}{d}.$$

Thus $\varphi(n)/n$ is the divisor sum of the multiplicative function $\mu(n)/n$, so $\varphi(n)/n$ is also multiplicative, hence $\varphi(n)$ is multiplicative.

Integers as sums of squares

Another arithmetical function of great interest is $r_2(n)$, the number of ways that n can be written as a sum of squares of two integers, positive, negative, or zero. In counting the number of ways, reversal of order is allowed. For example, there are four ways of writing 1 as a sum of two squares:

$$1 = 1^2 + 0^2 = (-1)^2 + 0^2 = 0^2 + 1^2 = 0^2 + (-1)^2,$$

so that $r_2(1) = 4$. Similarly, $r_2(2) = 4$, $r_2(3) = 0$, $r_2(4) = 4$, $r_2(5) = 8$, and $r_2(6) = 0$. This function is not multiplicative because $r_2(1) \neq 1$. However, $r_2(n)/4$ is multiplicative.

Let p be a prime divisor of n with $p \equiv 3 \pmod 4$, and let p^k be the highest power of p that divides n. Euler proved that n is the sum of two squares if and only if k is even for every such prime divisor p. In other words, $r_2(n) = 0$ if and only if n has at least one maximal prime power divisor p^k with $p \equiv 3 \pmod 4$ and k odd. There are infinitely many such n, the first few being $n = 3, 6, 7, 11, 12, 14, 15, 19, 21, 22, 23, 24, 27, 28, 30$.

Some arithmetical functions are described by giving the first two function values $f(1)$ and $f(2)$ and then expressing $f(n)$ for $n \geq 3$ in terms of $f(n-1)$ and $f(n-2)$. For example, the Fibonacci numbers are defined by

Fibonacci numbers

$$f(1) = 1,\ f(2) = 1, \text{ and } f(n) = f(n-1) + f(n-2)$$

for $n \geq 3$. In this particular case, the function values $f(n)$ are often written as F_n. The first few terms are

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377.$$

These numbers have many interesting properties. For example,

$$F_{2n+1} = 1 + F_2 + F_4 + F_6 + \cdots + F_{2n},$$

and (for those familiar with matrix multiplication) the nth power of the matrix $A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$ is

$$A^n = \begin{bmatrix} F_{n+1} & F_n \\ F_n & F_{n-1} \end{bmatrix}$$

for every $n \geq 2$. The relation $(-1)^n = F_{n+1}F_{n-1} - F_n^2$ is obtained by equating determinants in this matrix equation.

Fibonacci numbers also have an interesting divisibility property:

$$F_n | F_m \text{ whenever } n | m.$$

For example, $F_7 | F_{14}$ because $7 | 14$. This is easily verified because $F_7 = 13$ and $F_{14} = 377 = 13 \cdot 29$.

Linear recurrence sequence

Fibonacci numbers are just one example of a linear recurrence sequence, which is given by any two starting numbers x_1 and x_2 coupled with a recurrence relation of the form $x_n = Ax_{n-1} + Bx_{n-2}$ for $n \geq 3$, where A and B are given constants. This tells how to determine each term x_n from the two earlier terms x_{n-1} and x_{n-2}. Explicit formulas are available for expressing x_n in terms of x_1, x_2, and the roots of the quadratic equation $x^2 = Ax + B$. The formula for the Fibonacci numbers is

$$F_n = \frac{\alpha^n - \beta^n}{\alpha - \beta},$$

where α and β are the roots of the quadratic equation $x^2 = x + 1$, given by

$$\alpha = \frac{1+\sqrt{5}}{2} \text{ and } \beta = \frac{1-\sqrt{5}}{2}.$$

The number α, which also appears as a ratio in various geometric figures, is called the golden ratio. The ancient Greeks considered rectangles whose sides have ratio α to be the most pleasantly proportioned of all rectangles, aside from the square. They called them golden rectangles and believed that, for the ideal beauty of any figure (including the human form), the various parts should have the proportions of the golden ratio.

Diophantine equations. In elementary algebra the linear equation $6x - 9y = 29$ with two unknowns x and y has infinitely many solutions. It can be satisfied by taking any value for x and letting $y = (6x - 29)/9$. But, if it is required that x and y be integers, there are no solutions because $6x - 9y$ is divisible by 3 for all integers x and y, whereas 29 is not. Equations involving one or more unknowns to be solved for integer or rational values of the unknowns are called Diophantine equations in honour of Diophantus of Alexandria, who dealt with such problems in his treatise *Arithmetica* (*c.* AD 250).

Diophantine equations fall into three classes: those with no solutions, those with only finitely many solutions, and those with infinitely many solutions. The equation $6x - 9y = 29$ has no solutions, but the equation $6x - 9y = 30$, which upon division by 3 reduces to $2x - 3y = 10$, has infinitely many. For example, $x = 20$, $y = 10$ is a solution, and so is $x = 20 + 3t$, $y = 10 + 2t$ for every integer t, positive, negative, or zero. This is called a one-parameter family of solutions, with t being the arbitrary parameter.

Congruence methods provide a useful tool in determining the number of solutions to a Diophantine equation. Applied to the simplest Diophantine equation, $ax + by = c$, where a, b, and c are nonzero integers, these methods show that the equation has either no solutions or infinitely many. If the GCD d of a and b does not divide c, there are no solutions. But if d divides c, there are infinitely many solutions, and they form a one-parameter family,

$$x = x_0 + \frac{bt}{d}, \quad y = y_0 - \frac{at}{d},$$

where x_0, y_0 is any particular solution and t is an arbitrary integer. The integer x_0 can be obtained by solving the linear congruence $ax \equiv c \pmod{|b|}$, and then y_0 can be taken equal to $(c - ax_0)/b$, which is one of the solutions of $by \equiv c \pmod{|a|}$. (The symbol $|a|$ denotes the absolute value of a, that is, $|a| = a$ if a is positive, and $|a| = -a$ if a is negative.)

Similarly, the linear Diophantine equation in n unknowns,

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = c,$$

where the coefficients $a_1, a_2, \ldots, a_n$ and c are integers, has a solution in integers $x_1, x_2, \ldots, x_n$ if and only if c is divisible by the GCD $(a_1, a_2, \ldots, a_n)$. The equation

$$6x + 10y + 15z = 1$$

is solvable because the GCD $(6, 10, 15) = 1$. One solution is $x = 1$, $y = 1$, $z = -1$. In this case there is a two-parameter family of solutions given by

$$x = 1 + 5t,\ y = 1 + 6t - 3s,\ z = -1 - 6t + 2s,$$

where t and s are arbitrary integers.

Above it was shown that the Diophantine equation $6x - 9y = 29$ has no integer solutions because the left member is divisible by 3 while the right member is not. This illustrates a general method that is often used to show that a Diophantine equation has no solutions. If a solution exists, the corresponding congruence (mod m) will be satisfied for every modulus m. Therefore, if there is some modulus for which the congruence cannot be satisfied, the equation itself cannot be satisfied. The choice of an appropriate modulus is not always obvious. For example, the Diophantine equation $15x^2 - 7y^2 = 9$, if written mod 7, becomes $x^2 \equiv 9 \pmod 7$, which is solvable. But if it is written mod 5 it becomes $-2y^2 \equiv 4 \pmod 5$, or $y^2 \equiv 3 \pmod 5$, which is not solvable because 3 is a quadratic nonresidue mod 5, the only quadratic residues being 1 and 4. Therefore, the original equation $15x^2 - 7y^2 = 9$ has no integer solutions.

Even if a Diophantine equation has no solutions, it is possible that the corresponding congruence mod m can be satisfied for every $m \geq 2$. For example, the equation

$$(x^2 - 13)(x^2 - 17)(x^2 - 221) = 0$$

has no integer solution x because none of 13, 17, or 221 is a square. Nevertheless, with the aid of the quadratic reciprocity law it can be shown that the congruence

$$(x^2 - 13)(x^2 - 17)(x^2 - 221) \equiv 0 \pmod{m}$$

is solvable for every modulus $m \geq 2$.

The method of congruences can be used to give a partial proof of Euler's theorem on representing numbers as a sum of two squares. According to Euler's theorem, for any prime p the Diophantine equation $x^2 + y^2 = p$ has integer solutions if $p = 2$ and if $p \equiv 1 \pmod 4$ but not if $p \equiv 3 \pmod 4$. The nonexistence of solutions can be seen by writing the equation as a congruence mod 4. If $p \equiv 3 \pmod 4$ the congruence becomes $x^2 + y^2 \equiv 3 \pmod 4$. This is not solvable because the square of any integer is congruent to 0 (mod 4) if the integer is even, and congruent to 1 (mod 4) if the integer is odd, so the sum $x^2 + y^2$ cannot be congruent to 3 (mod 4). It is more difficult to prove that the equation $x^2 + y^2 = p$ is solvable if $p \equiv 1 \pmod 4$. The number of solutions in this case is finite because both x^2 and y^2 are less than p.

A well-known theorem of Lagrange states that every positive integer n is expressible as a sum of four squares of integers. For example, $39 = 1 + 4 + 9 + 25$. However, 39 is not the sum of three squares. In fact, if n is any positive integer of the form $8k + 7$, the Diophantine equation $x^2 + y^2 + z^2 = n$ has no integer solutions. This can be shown by writing the equation as a congruence mod 8,

$$x^2 + y^2 + z^2 \equiv 7 \pmod 8.$$

The square of an integer is congruent to 0, 1, or 4 (mod 8), and the sum of three such numbers cannot be congruent to 7 (mod 8).

One of the oldest Diophantine equations comes from the theorem of Pythagoras, which states that $x^2 + y^2 = z^2$, where x, y, and z are the lengths of the three sides of a right triangle. If the equation is satisfied with integer values of x, y, and z, the triple (x, y, z) is called a Pythagorean triple. The examples (3, 4, 5) and (5, 12, 13) were known to the Babylonians 1,200 years before the time of Pythagoras. A method for obtaining all Pythagorean triples is outlined in Euclid's *Elements* and can be summarized by the formulas

Pythagorean triples

$$x = t(a^2 - b^2),\ y = 2tab,\ z = t(a^2 + b^2),$$

where t, a, and b are arbitrary positive integers such that $a > b$, a and b are relatively prime, and one of a or b is odd and the other even. When $t = 1$, the numbers x, y, z have no prime factor in common and the triple is called primitive. The primitive triples (3, 4, 5) and (5, 12, 13) are obtained by taking $a = 2$, $b = 1$, and $a = 3$, $b = 2$.

There are various ways to derive the formulas for the Pythagorean triples. One method is to factor the left member of the equation $x^2 + y^2 = z^2$ using complex numbers:

$$(x + iy)(x - iy) = z^2,$$

where $i^2 = -1$. If x and y are integers, the complex numbers $x + iy$ and $x - iy$ are called Gaussian integers. If the product of two relatively prime positive integers is a square, according to the fundamental theorem of arithmetic each factor separately must be a square. The same property holds for Gaussian integers, so each of $x + iy$ and $x - iy$ (or with x and y interchanged in both of these) is the square of a Gaussian integer, say

$$x + iy = (a + ib)^2 \text{ and } x - iy = (a - ib)^2.$$

The equations $x = a^2 - b^2$ and $y = 2ab$ follow by equating real and imaginary parts.

A quadratic form is a polynomial in one or more variables all of whose terms are of degree 2, as in the example

$$f(x, y, z) = 3x^2 + y^2 + 20z^2 + 2xy - 14xz - 8yz.$$

The corresponding Diophantine equation $f(x, y, z) = 0$ can be studied by introducing a change of variables that reduces the quadratic form to a simpler quadratic form containing square terms but no cross product terms such as xy, xz, or yz. For example, if x, y, and z are expressed in terms of new variables u, v, w by means of the equations

$$x = u + 2v + 3w,\ y = 2u + v + 4w,\ z = u + v + 2w,$$

then the quadratic form $f(x, y, z)$ given above is transformed into a new quadratic form

$$g(u, v, w) = u^2 + v^2 - w^2.$$

This is easily verified, after some algebra, by direct substitution. Because each of x, y, z is a linear combination of u, v, w with integer coefficients, integer values of u, v, w produce integer values of x, y, z. Therefore, any integer solution of the equation

$$g(u, v, w) = 0$$

leads to an integer solution of $f(x, y, z) = 0$.

For example, the well-known Pythagorean triple

$$(u, v, w) = (3, 4, 5),$$

which satisfies $g(u, v, w) = 0$, gives the solution $(x, y, z) = (26, 30, 17)$ for the equation $f(x, y, z) = 0$. Another Pythagorean triple $(u, v, w) = (4, 3, 5)$ gives a different solution $(x, y, z) = (25, 31, 17)$. The obvious solution $(u, v, w) = (1, 0, 1)$ produces the not-so-obvious solution $(x, y, z) = (4, 6, 3)$. The formulas

$$u = t(a^2 - b^2),\ v = 2tab,\ w = t(a^2 + b^2),$$

which give solutions of

$$u^2 + v^2 = w^2,$$

can be substituted into the equations for x, y, z to give formulas for a family of solutions of $f(x, y, z) = 0$. Because of the lack of symmetry, this method does not give all solutions of $f(x, y, z) = 0$. Further solutions (x, y, z) are obtained by starting with

$$u = 2tab,\ v = t(a^2 - b^2),\ w = t(a^2 + b^2).$$

In this particular example, the three linear equations relating the variables can be solved for u, v, w in terms of x, y, z, the result being

$$u = -2x - y + 5z,\ v = -y + 2z,\ w = x + y - 3z.$$

Because each of u, v, w is a linear combination of x, y, z with integer coefficients, integer values of x, y, z produce integer values of u, v, w. Therefore, any integer solution of the equation $f(x, y, z) = 0$ leads to an integer solution of $g(u, v, w) = 0$. Consequently, there is a one-to-one correspondence between integer solutions of

$$f(x, y, z) = 0$$

and integer solutions of $g(u, v, w) = 0$. As this example suggests, a basic problem in the theory of quadratic forms is to determine linear transformations of the variables that reduce complicated quadratic forms such as $f(x, y, z)$ to simpler forms such as $g(u, v, w)$ that are more easily studied.

The best-known of all Diophantine problems is Fermat's last theorem, which states that, for every integer $n \geq 3$, the equation

Fermat's last theorem

$$x^n + y^n = z^n$$

has no solutions in positive integers x, y, z. A note by Fermat, inscribed in the margin of his copy of Diophantus' *Arithmetica*, asserted that he had found a wonderful proof of this statement, but "the margin is too narrow to contain it." Despite Fermat's assertion, no complete proof was ever discovered in his notebooks. In 1983 the German mathematician Gerd Faltings proved that, for any fixed $n \geq 3$, the Fermat equation $x^n + y^n = z^n$ has at most a finite number of integer solutions with $(x, y) = 1$. Closing the gap from finite to zero integer solutions was achieved in 1995 by the English mathematician Andrew Wiles, thereby solving one of the most famous of all mathematical problems.

Because $x^4 + y^4 = z^4$ has no solutions in positive integers, there is no right triangle whose legs and hypotenuse are squares of integers. In 1772 Euler considered a three-dimensional version of this result and conjectured that $x^4 + y^4 + z^4 = t^4$ has no solutions in positive integers. In 1948 Morgan Ward proved that there is no solution with $t \leq 10^4$, but in 1987 the American mathematician Noam D. Elkies disproved Euler's conjecture when he found the solution

$$2{,}682{,}440^4 + 15{,}365{,}639^4 + 18{,}796{,}760^4 = 20{,}615{,}673^4$$

and showed that there are infinitely many solutions. In 1988 Roger Frye found the solution

$$95{,}800^4 + 217{,}519^4 + 414{,}560^4 = 422{,}481^4$$

and proved that there are no solutions with smaller positive integers. Euler made the more general conjecture that at least n nth powers are required to sum to an nth power if $n > 2$. This was disproved for $n = 5$ in 1966 by the Americans Leon J. Lander and Thomas R. Parkin, who observed that

$$27^5 + 84^5 + 110^5 + 133^5 = 144^5.$$

This is the smallest instance in which four fifth powers sum to a fifth power.

As discussed earlier, for each prime $p \equiv 1 \pmod 4$ the Diophantine equation $x^2 + y^2 = p$ has at most a finite number of solutions because both x^2 and y^2 are less than p. By contrast, there is no apparent reason why the Diophantine equation $x^3 - y^2 = 2$ should have at most a finite number of integer solutions. In fact, Fermat showed that this equation has only two solutions, $x = 3$ and $y = \pm 5$. This result was extended by the German mathematician Carl Ludwig Siegel (1896–1981), who proved that if $P(x)$ is a polynomial with integer coefficients, the Diophantine equation $y^2 = P(x)$ has at most a finite number of integer solutions x, y if $P(x)$ has at least three distinct linear factors. By the fundamental theorem of algebra, $P(x)$ can always be factored as follows:

$$P(x) = a_0(x - a_1)(x - a_2) \cdots (x - a_n),$$

where a_0 is an integer and $a_1, a_2, \ldots, a_n$ are complex numbers. Siegel's theorem requires that at least three of these complex numbers must be different.

For $n \geq 3$, let

$$g(x, y) = a_0x^n + a_1x^{n-1}y + \cdots + a_{n-1}xy^{n-1} + a_ny^n$$

be a homogeneous polynomial with integer coefficients that cannot be factored as a product of polynomials of the same type of lower degree. A theorem of the Norwegian mathematician Axel Thue (1863–1922) states that for every integer m the Diophantine equation $g(x, y) = m$ has at most a finite number of integer solutions x, y. In 1967 the British mathematician Alan Baker proved that for any solution, both $|x|$ and $|y|$ cannot exceed a certain constant C whose value can be calculated explicitly in terms of n, m and the coefficients $a_0, a_1, \ldots, a_n$. On the other hand, the Soviet mathematician Yuri Matijasevic, completing earlier work by the Americans Martin Davis, Hilary Putnam, and Julia Robinson, showed in 1970 that there cannot exist any algorithm for determining whether an arbitrary Diophantine equation does or does not have a solution.

Applications of number theory. Number theory has been applied to such diverse areas as physics, biology, chemistry, computing, cryptography, acoustics, communications, and graphic design. In particular, congruence arithmetic has been used in constructing perpetual calendars, scheduling round-robin tournaments, splicing telephone cables, devising systematic methods for storing computer files, constructing magic squares, and generating random numbers. This section describes two simple applications of congruences in machine computation and cryptography.

Efficiency in machine computation involving large numbers is often limited by the machine's memory (the amount of space available for storing data) and by the length of time required to perform computations. Even simple operations such as addition, subtraction, and multiplication of integers can overwhelm the memory or take an unreasonable amount of time if the integers are large and if millions of these operations must be performed in a given calculation. The Chinese remainder theorem can be used to increase efficiency by making use of relatively small numbers in most of the calculation. Suppose, for example, that a lengthy calculation involves many 10-digit integers. Most of the calculation can be done with five-digit integers by adopting the following procedure. Choose four relatively prime moduli m_1, m_2, m_3, and m_4 of five digits each, for example, $2^{16} - 3$, $2^{16} - 1$, $2^{16} + 1$, and $2^{16} + 3$. Their product $m = m_1m_2m_3m_4$ has 20 digits, which is twice the number of digits in the data. By the Chinese remainder theorem each integer in the data can be uniquely represented by its residues mod m_1, m_2, m_3, and m_4, respectively. Because each of these residues has no more than five digits, the required arithmetic can be performed efficiently on these residues with respect to each of the moduli, yielding integer answers $a_1 \pmod{m_1}$, $a_2 \pmod{m_2}$, $a_3 \pmod{m_3}$, and $a_4 \pmod{m_4}$, where each a_i has no more than five digits. These numbers, in turn, are combined via the Chinese remainder theorem to obtain the final result mod m, which will be accurate to 20 digits. Even though the lengthy calculation is repeated four times, once for each modulus, most of it only uses five-digit integers and is accomplished quickly without overwhelming the memory.

Cryptography

An application to cryptography relies on the disparity in computer time required to find large primes and to factor large integers. For example, a code maker chooses two large primes, say p and q, of about 100 decimal digits each. There are procedures for finding such primes that require only a few minutes of computer time. The primes p and q are kept secret, but their product $n = pq$, a 200-digit number, is made public. A message is coded by a procedure that requires only the knowledge of n. (Because n is made public, these are often called public key codes.) But to decode the message, both factors p and q must be known. With the most efficient computer techniques yet devised, factoring a 200-digit number such as n could require billions of years of computer time on a single computer. For this reason, the codes are considered unbreakable, at least with the present state of knowledge of how to factor large numbers.

The actual procedure for coding the message can be described as follows. Each letter of the message is replaced by two digits, say $A = 01$, $B = 02, \ldots, Z = 26$, and the entire message is broken into pieces of convenient length, where each piece is an integer smaller than the public value n. Each piece is coded separately by the following method. A prime r is chosen that does not divide either $p - 1$ or $q - 1$. Like n, the number r is made public. Let x denote the piece to be coded, raise x to the power r, and reduce this mod n to obtain an integer y between 1 and n. Thus $y \equiv x^r \pmod n$ and $1 \leq y < n$. The number y is the coded form of x.

To decode, one must obtain x from y. If p and q are known, this can be done very simply as follows. Let s denote the reciprocal of $r \bmod \varphi(n)$, so that

$$sr \equiv 1 \pmod{\varphi(n)}.$$

Then $sr = 1 + t\varphi(n)$ for some integer t. By the Euler–Fermat theorem, $x^{\varphi(n)} \equiv 1 \pmod n$, hence

$$x^{t\varphi(n)} \equiv 1 \pmod n.$$

But

$$y^s \equiv x^{rs} \equiv x^{1 + t\varphi(n)} \equiv x \pmod n,$$

so y^s is the same as x modulo n. In other words, to obtain x from y, simply raise y to the power s and reduce mod n. If p and q are known, s and y^s can be determined (mod n) by straightforward computations that require only a few minutes of machine time. But if p and q are not known, the problem of recovering x from y seems insurmountable.

Recreational number theory. Newspapers and magazines challenge their readers with mathematical puzzles, games, and problems that can be understood by the layperson. This section describes some examples drawn from number theory.

Start with any three-digit number, such as 529. Extend it to a six-digit number 529,529 by repeating the digits. Divide the six-digit number by 7. The quotient is 75,647, with no remainder. Now divide this quotient by 11. Again the division is exact, and the new quotient is 6,877. When this quotient is divided by 13 the result is 529, the original number. The problem is to explain why this happens, not only for 529 but for any three-digit number. The explanation is that repetition of the digits is the same as multiplying the original number by 1001, which happens to be the product of 7, 11, and 13, so division by each of these factors restores the original number.

The $3n + 1$ problem

Some problems in recreational number theory, although easily comprehended, are difficult to solve. One well-known example is the $3n + 1$ problem. Start with any

integer n greater than 1. If n is even, replace n by $n/2$. If n is odd, replace it by $3n+1$. Then repeat the process until 1 is obtained. For example, starting with $n=3$ the sequence is 3, 10, 5, 16, 8, 4, 2, 1. Is it true that, starting with any initial integer n, the number 1 will always be obtained after a finite number of steps? To date this question is unresolved, although many professional and amateur mathematicians have worked on it.

Magic squares. Magic squares are among the oldest mathematical amusements. A magic square is a square array of integers in which the rows, columns, and diagonals have the same sums. A square containing the consecutive integers from 1 to n^2 is said to be of order n. The sum of the integers from 1 to n^2 is $n^2(n^2+1)/2$, so in a magic square the numbers in each row, column, and diagonal have the magic sum $n(n^2+1)/2$. The following magic squares have orders 3 and 4, respectively:

8	1	6
3	5	7
4	9	2

16	3	2	13
5	10	11	8
9	6	7	12
4	15	14	1

The square of order 3, with magic sum 15, occurs in Chinese writings as early as 2200 BC. The square of order 4, with magic sum 34, appears in Albrecht Dürer's engraving "Melencolia I." The date of the work, 1514, appears as part of the fourth row.

Interest in magic squares spread from China to Japan, India, and the Middle East, and they were introduced to Europe in Byzantine times.

Many methods are known for constructing magic squares of orders greater than 2, but despite intense efforts by professional and amateur mathematicians (including the celebrated American amateur Benjamin Franklin), no methods are known for constructing all such magic squares. There are exactly eight magic squares of order 3, containing the numbers 1 to 9:

8	1	6
3	5	7
4	9	2

4	9	2
3	5	7
8	1	6

2	9	4
7	5	3
6	1	8

6	1	8
7	5	3
2	9	4

8	3	4
1	5	9
6	7	2

4	3	8
9	5	1
2	7	6

2	7	6
9	5	1
4	3	8

6	7	2
1	5	9
8	3	4

These squares can be obtained by interchanging the first and third rows of the Chinese square to construct a second square, interchanging the first and third columns of the second square to construct a third square, and interchanging the first and third rows of the third square to construct a fourth square. The rows and columns of each of these four squares are then interchanged to obtain four new ones, for a total of eight altogether.

Methods for constructing magic squares usually prescribe rules for entering the numbers $1, 2, \ldots, n^2$ into the cells of the square. One method that works for any odd order n was given in 1693 by Simon de la Loubère, who learned of it while ambassador to Siam from the court of Louis XIV. In 1929 D.N. Lehmer applied congruences to investigate the uniform step method, a generalization of Loubère's method. The uniform step method can be described as follows. Two positive integers a and b, each relatively prime to n and less than n, are chosen in advance. The cells of the square are assigned coordinates (x, y), x being the number of the column counting from the left, and y the number of the row counting from the bottom. The number 1 is put in an arbitrary cell with coordinates (p, q). The number 2 is then placed in the cell $(p+a,\ q+b)$, the number 3 in the cell $(p+2a,\ q+2b)$, and so on, proceeding from one cell to the next by making the uniform step (a, b), the coordinates being reduced to their least positive residues mod n to keep the numbers inside the square. This procedure places the number $n+1$ in the cell $(p+na,\ q+nb)$, which, mod n, is the same as (p, q), the cell already occupied by 1. The uniform step (a, b) is said to fail at this point because two numbers occupy the same cell. A new step (c, d), called a break step, is now used, and the number $n+1$ is placed in the cell $(p+c,\ q+d)$, where c and d are chosen to arrive at an unoccupied cell. The uniform step (a, b) is again resorted to; $n+2$ is placed in the cell $(p+c+a,\ q+d+b)$, $n+3$ in the cell $(p+c+2a,\ q+d+2b)$, and so on. The uniform step fails again when $2n+1$ is placed in $(p+c+na,\ q+d+nb)$, which, mod n, is the cell already occupied by $n+1$, namely $(p+c,\ q+d)$. The break step (c, d) is then used once more to get out of this occupied cell, and $2n+1$ is placed in the cell $(p+2c,\ q+2d)$. Then the uniform step (a, b) is returned to, with $2n+2, 2n+3, \ldots$ being placed in the cells $(p+2c+a,\ q+2d+b)$, $(p+2c+2a,\ q+2d+2b), \ldots$. If all five numbers a, b, c, d, and $ad-bc$ are relatively prime to n, this method will fill the square, and the rows and columns will have the magic sum. The diagonals will also have the magic sum if the coordinates of the initial cell (p, q) satisfy the congruences $2p \equiv a+c+1 \pmod{n}$, and $2q \equiv b+d+1 \pmod{n}$.

The uniform step method

In the Chinese square of order 3, the initial cell (p, q) is $(2, 3)$, the uniform step (a, b) is $(1, 1)$, the break step is $(2, 1)$, and $ad-bc=-1$. The other magic squares of order 3 containing integers 1 to 9 can all be obtained by the uniform step method, with different choices of initial cell (p, q), uniform step (a, b), or break step (c, d). However, the method does not produce all magic squares of order n except for $n=3$.

Magic squares with entries $1, 2, \ldots, n^2$ can be used to construct new magic squares whose entries are not necessarily consecutive integers. For example, if each entry in the Chinese square is multiplied by 2 and these numbers are added, entry by entry, to the second square of order 3, a new magic square will be produced whose rows, columns, and diagonals have the magic sum 45:

20	11	14
9	15	21
16	19	10

In general, if the entries of a magic square with magic sum S are multiplied by an integer A, a new magic square is obtained with magic sum AS. If two magic squares of the same size with magic sums S_1 and S_2 are added, entry by entry, a new magic square is obtained with magic sum S_1+S_2.

Wythoff's game. A game that was popular in the United States in the 19th century involves two players and a number of objects (counters) arranged in several heaps. Each player, in turn, takes counters from a heap; a player may take only one or any number up to the entire heap but is allowed to touch one heap only. The player who takes the last counter wins the game. The game was analyzed mathematically in 1902 by the American mathematician Charles Leonard Bouton, who gave it the name "Nim" and showed that one of the players can always force a win. A simple variation of this game, invented by W.A. Wythoff in 1907, is described here. Two players A and B take turns removing counters from two heaps. Each player in turn must take either any number of counters from the first heap, or any number from the second heap, or the same number from both heaps. Whoever takes the last counter from the table wins the game.

A remarkable feature of this game is that the first player A will always win if A knows a special sequence of number pairs that begins as follows:

$$(1, 2), (3, 5), (4, 7), (6, 10), (8, 13), (9, 15), (11, 18), (12, 20), (14, 23), (16, 26), (17, 28), \ldots.$$

If the two heaps have the same number of counters, A takes everything and wins. Otherwise, A leaves the num-

bers of counters in the two piles to be one of the number pairs in the sequence. For example, A could leave 14 counters in one pile and 23 in the other, represented by the pair (14, 23). Whatever player B does with this pair, the outcome will be a pair not in the list, and then A can move back to another pair in the list. For example, if B is faced with the pair (14, 23) and removes nine counters from each pile to produce the pair (5, 14), then A can remove 11 counters from the 14 to get (5, 3), which appears in the list as (3, 5).

Two general properties of the sequence explain why it is a list of winning positions for the first player. If (a, b) is the nth pair in the sequence, then a is the smallest positive integer not present in any of the earlier pairs of the sequence, and $b = a + n$. These properties suffice to determine all pairs in the list, starting with (1, 2) and extending beyond those given. The winning strategy depends on the following theorem of the Canadian mathematician Samuel Beatty, proved in 1926, that gives an alternative way of constructing the sequence: for any two positive irrational numbers α and β satisfying the equation

$$\frac{1}{\alpha} + \frac{1}{\beta} = 1,$$

the two sequences

$$[\alpha], [2\alpha], [3\alpha], \ldots \text{ and } [\beta], [2\beta], [3\beta], \ldots$$

together contain every positive integer exactly once. Here $[x]$ denotes the greatest integer in x. For example, if α is the golden ratio, $\alpha = (1 + \sqrt{5})/2$, and, if $\beta = \alpha^2$, it is easily verified that $\alpha^2 = 1 + \alpha$, hence

$$\frac{1}{\alpha} + \frac{1}{\beta} = 1.$$

For this choice of α and β, the sequence of pairs of integers $([n\alpha], [n\beta])$ is the list of winning positions.

Approximating irrational numbers by rational numbers. One of the most interesting numbers in mathematics is π, which can be defined as the ratio of the circumference of a circle to its diameter or as the ratio of the area of a circle to the square of its radius. Its numerical value has been calculated to more than a billion decimal places, the first 10 being given by 3.1415926535 . . . .

The number π is irrational; that is, it is not the quotient of two integers. (Quotients of integers are called rational numbers.) Rational approximations of π to any desired degree of accuracy can be obtained by truncating its decimal expansion. For example, each of the first few decimal approximations 3.1, 3.14, 3.141, 3.1415, and 3.14159 can be written in the form h/k, where h and k are relatively prime integers, giving the following rational approximations:

$$\frac{31}{10}, \frac{157}{50}, \frac{3{,}141}{1{,}000}, \frac{6{,}283}{2{,}000}, \frac{314{,}159}{100{,}000}.$$

On the other hand, the fraction $^{22}/_{7} = 3.142857\ldots$ approximates π nearly as well as $^{3.141}/_{1{,}000}$, even though its denominator, 7, is very small compared to 1,000. An approximation better than any of these is the fraction

$$\frac{355}{113} = 3.141592920\ldots,$$

whose first six decimals agree with those of π, yet its denominator, 113, is very small compared to that of its nearest competitor in the above list, $^{314.159}/_{100{,}000}$. In fact, no fraction with denominator less than 113 is closer to π than $^{355}/_{113}$.

An important branch of number theory has to do with approximating a given irrational number by rational numbers h/k whose denominators are not too large. One approach is to use continued fractions, which are obtained in the following manner.

Continued fractions

Every real number x can be written as a sum, $x = [x] + (x)$, where $[x]$ denotes the integer part of x (the largest integer not exceeding x), and (x) denotes the difference $x - [x]$, called the fractional part of x. If x is an integer the fractional part (x) is zero. If x is not an integer the fractional part (x) is positive and less than 1, so it can be written as $1/y$ for some real $y > 1$. The number y in turn can be written as the sum $[y] + (y)$, giving

$$x = [x] + \frac{1}{y} = [x] + \frac{1}{[y] + (y)}.$$

If $(y) = 0$, then $x = [x] + 1/[y]$, a rational number, and the continued fraction terminates. But if $(y) > 0$, then $(y) = 1/z$ for some $z > 1$, and

$$x = [x] + \cfrac{1}{[y] + \cfrac{1}{[z] + (z)}}.$$

Repetition of this process gives a continued fraction representation of x,

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{a_3 + \ldots}}}$$

where $a_0 = [x]$, $a_1 = [y]$, $a_2 = [z]$, and so on. This is written more simply as $x = [a_0; a_1, a_2, a_3, \ldots]$, where all the integers after a_0 are positive. If x is rational the process terminates after a finite number of steps and the continued fraction is finite. For example,

$$\frac{22}{7} = [3; 7] \text{ and } \frac{355}{113} = [3; 7, 16].$$

But if x is irrational the continued fraction is infinite. The continued fraction for π begins as follows:

$\pi = [3; 7, 15, 1, 292, 1, 1, 1, 2, 1, 3, 1, 14, 2, 1, 1, 2, 2, 2, \ldots$

Truncating the continued fraction of an irrational number θ at the integer a_n produces a rational number

$$\frac{p_n}{q_n} = [a_0; a_1, \ldots, a_n],$$

called the nth convergent to θ. Continued fractions are useful because p_n/q_n is closer to θ than any other rational number h/k with denominator k not exceeding q_n. For example, the third convergent to π is

$$[3; 7, 15, 1] = \frac{355}{113},$$

a fraction closer to π than any other fraction with denominator less than 113. Moreover, it can be shown that the nth convergent p_n/q_n to any irrational θ is smaller than θ if n is even, larger than θ if n is odd, and no farther away from θ than $1/q_n^2$, as summarized by the inequality

$$\left| \theta - \frac{p_n}{q_n} \right| < \frac{1}{q_n^2}.$$

In 1891 the German mathematician Adolf Hurwitz proved that infinitely many convergents satisfy the stronger inequality

$$\left| \theta - \frac{p}{q} \right| < \frac{1}{\sqrt{5}q^2},$$

but this is not true for all θ if $\sqrt{5}$ is replaced by any larger number. The golden ratio

$$\alpha = \frac{1 + \sqrt{5}}{2}$$

illustrates this last point. If $A > \sqrt{5}$, it can be shown that the inequality

$$\left| \alpha - \frac{p}{q} \right| < \frac{1}{Aq^2}$$

is satisfied only by a finite number of fractions p/q.

The continued fraction for the golden ratio is

$$\alpha = [1; 1, 1, 1, \ldots],$$

with an endless succession of 1s, as is easily verified by noting that

$$\alpha = 1 + \frac{1}{\alpha} = 1 + \cfrac{1}{1 + \cfrac{1}{\alpha}} = \cdots.$$

The golden ratio α satisfies $\alpha^2 = \alpha + 1$, a quadratic equation easily put into the form $ax^2 + bx + c = 0$, where a, b, and c are integers. Irrational numbers with this property are called real quadratic irrationals and have the form

$(A+\sqrt{B})/C$, where A, B, and C are integers, with B positive and not a perfect square. They can also be described as those irrational numbers whose continued fractions are periodic; that is, the continued fraction consists of endless repetition of the same block of terms, except possibly for a few terms at the beginning. Examples are

$$\sqrt{7} = [2; 1, 1, 1, 4, 1, 1, 1, 4, 1, 1, 1, 4, \ldots]$$
$$= [2; \overline{1, 1, 1, 4}]$$

and

$$\frac{63+\sqrt{30}}{13} = [5; 3, 1, 2, 1, 4, 1, 2, 1, 4, 1, 2, 1, 4, \ldots]$$
$$= [5; 3, \overline{1, 2, 1, 4}].$$

A bar over a block of numbers denotes endless repetition of that block. Another example is

$$\sqrt{43} = [6; \overline{1, 1, 3, 1, 5, 1, 3, 1, 1, 12}].$$

Although the nth convergent in the continued fraction of any irrational number θ can be made arbitrarily close to θ by making n sufficiently large, the rate of convergence is very slow if θ has a periodic continued fraction. For example, the first four convergents to $\sqrt{43} = 6.557438\ldots$ are

$$[6; 1] = 7,$$

$$[6; 1, 1] = 6.5,$$

$$[6; 1, 1, 3] = \frac{46}{7} = 6.571\ldots,$$

$$[6; 1, 1, 3, 1] = \frac{59}{9} = 6.555\ldots,$$

the fourth of which agrees with $\sqrt{43}$ to two decimals. By contrast, the third convergent to π,

$$[3; 7, 15, 1] = \frac{355}{113} = 3.141592920\ldots,$$

agrees with π to six decimals.

Algebraic and transcendental numbers. Besides classification into rational and irrational numbers, real numbers can also be classified as algebraic or transcendental numbers. A number is said to be algebraic if it is the root of a polynomial equation of the form

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2x^2 + a_1x + a_0 = 0,$$

where the coefficients are integers and not all zero. The real roots are called real algebraic numbers. If the coefficient a_n of the highest power of x is 1, the roots are called algebraic integers. For example, the roots of the equation $x^3 - 1 = 0$ are algebraic integers, only one of which, $x = 1$, is real. The other two roots are complex numbers, $x = (-1 \pm i\sqrt{3})/2$, where $i^2 = -1$.

If an algebraic number θ is a root of the polynomial equation given above and is not a root of any equation of this type of lower degree, then θ is said to be algebraic of degree n. For example, none of the roots of $x^3 - 1 = 0$ is algebraic of degree 3 because they all satisfy polynomial equations of lower degree with integer coefficients. The root $x = 1$ satisfies $x - 1 = 0$ and therefore is algebraic of degree 1, whereas the other two roots satisfy $x^2 + x + 1 = 0$, which makes them algebraic of degree 2. Algebraic numbers of degree 1 are rational numbers, while algebraic integers of degree 1 are the usual integers, often referred to as "rational integers" in algebraic number theory to distinguish them from irrational algebraic integers of higher degree, such as $\alpha = (1+\sqrt{5})/2$, which is of degree 2 because it satisfies $x^2 - x - 1 = 0$.

Transcendental numbers

A real number that is not algebraic is called transcendental. Such numbers were first shown to exist in 1844, when the French mathematician Joseph Liouville proved that there is a strong restriction on how well algebraic numbers can be approximated by rational numbers. Specifically, if θ is a real algebraic number of degree n and if C is any positive constant, there is at most a finite number of rational fractions p/q such that

$$\left|\theta - \frac{p}{q}\right| < \frac{C}{q^{n+1}}.$$

This inequality can be used to show that any number θ with infinitely many very good rational approximations must necessarily be transcendental. An example is the Liouville number $\lambda = 0.110001000\ldots$, where a 1 is placed in the rth decimal place if $r = m!$ for some positive integer m, and the remaining digits are 0s. Thus, 1 appears in the rth decimal place for $r = 1, 2, 6, 24, 120, 720, \ldots$. The number λ can be written as an infinite series,

$$\lambda = \sum_{m=1}^{\infty} \frac{1}{10^{m!}},$$

and good rational approximations to λ can be obtained by taking the sum of the first k terms,

$$\lambda_k = \sum_{m=1}^{k} \frac{1}{10^{m!}} = \frac{c_k}{10^{k!}}$$

where c_k is an integer. The difference between λ and λ_k is

$$\sum_{m=k+1}^{\infty} 10^{-m!} < \frac{2}{10^{(k+1)!}},$$

which can be written as

$$|\lambda - \lambda_k| = \left|\lambda - \frac{c_k}{10^{k!}}\right| < \frac{2}{(10^{k!})^{k+1}}.$$

It follows that λ cannot be an algebraic number of degree n for any n because there are infinitely many integers $k > n$, contradicting Liouville's theorem that there is at most a finite number of rational fractions p/q satisfying the inequality

$$\left|\lambda - \frac{p}{q}\right| < \frac{2}{q^{n+1}}.$$

An irrational number θ is said to be approximated to order n if there is a constant C, depending only on θ, such that the inequality

$$\left|\theta - \frac{p}{q}\right| < \frac{C}{q^n}$$

has infinitely many solutions in integers p and q. The number λ defined above can be approximated to any order whatever. It has already been observed that every irrational θ can be approximated to order 2 at least, and that a real algebraic number of degree n cannot be approximated to order $n+1$. This result was improved by many authors until 1955, when the British mathematician Klaus Friedrich Roth established the best possible conclusion, that no irrational algebraic number can be approximated to any order greater than 2.

Impossibility of squaring the circle

In 1882 the German mathematician Ferdinand Lindemann proved that π is a transcendental number. This settled the ancient question of squaring the circle: given any circle, is it possible to construct a square of equal area, using Euclidean constructions, with an unmarked straightedge and compass? The answer is no, because from any given unit length, the only lengths that can be constructed with an unmarked straightedge and compass are algebraic numbers.

If θ is irrational and if p_n/q_n is the nth convergent to θ, then the inequality relating p_n/q_n to θ,

$$\left|\theta - \frac{p_n}{q_n}\right| < \frac{1}{q_n^2},$$

implies

$$|q_n\theta - p_n| < \frac{1}{q_n},$$

which is satisfied with arbitrarily large values of q_n. Therefore, the product $q_n\theta$ can be made arbitrarily close to the integer p_n by taking n sufficiently large, which makes the fractional part $(q_n\theta)$ arbitrarily close to 0 or 1. In 1884 Kronecker proved a more general result: for any irrational θ and any prescribed real number x in the interval from 0 to 1, there are positive integers n such that $(n\theta)$ lies arbitrarily close to x. This result is described by saying that the fractional parts $(n\theta)$ are dense in the unit interval. Even more is known about the distribution of these numbers. About half of them lie in any interval of length $1/2$, about one-third lie in any interval of length $1/3$, and so on. More precisely, if I is any subinterval of length L

lying within the interval from 0 to 1, and if N denotes the number of fractional parts $(\theta), (2\theta), \ldots, (n\theta)$ that lie in I, then the quotient N/n approaches L as a limit as n tends to infinity. This property is described by saying that the fractional parts $(n\theta)$ are uniformly distributed over the unit interval.

ALGEBRAIC NUMBER THEORY

Algebraic integers and unique factorization. As discussed in the previous section, an algebraic number θ of degree n is any root of a polynomial equation of degree n with integer coefficients that is not a root of any such equation of lower degree. If the equation is

$$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 = 0,$$

with $a_n = 1$, then θ is called an algebraic integer. The algebraic integers of degree 1 are the ordinary rational integers.

Examples of algebraic integers of degree 2 are the quadratic irrationals $\sqrt{3}$, $1+\sqrt{3}$, and $(1+\sqrt{5})/2$. They satisfy, respectively, the quadratic equations $x^2 - 3 = 0$, $x^2 - 2x - 2 = 0$, and $x^2 - x - 1 = 0$. Because a quadratic equation with integer coefficients can have complex roots, algebraic integers of degree 2 can be complex numbers. Examples are

$$i,\ 1+i, \text{ and } \frac{1+i\sqrt{3}}{2},$$

which satisfy, respectively,

$$x^2 + 1 = 0,\ x^2 - 2x + 2 = 0, \text{ and } x^2 - x + 1 = 0.$$

More generally, the algebraic integers of degree 2 can be described in terms of $\sqrt{d}$, where d is any square-free integer (that is, d is not divisible by the square of any prime, and $d \neq 1$). The permissible values of d are 2, 3, 5, 6, 7, 10, 11, . . . , together with their negatives and also -1. All numbers of the form $a + b\sqrt{d}$, where a and b are rational integers, are algebraic integers of degree 2; they satisfy the quadratic equation $x^2 - 2ax + a^2 - db^2 = 0$. But if $d \equiv 1$ (mod 4), numbers of the form $(a + b\sqrt{d})/2$ with a and b odd are also algebraic integers of degree 2; they satisfy the quadratic equation

$$x^2 - ax + \frac{a^2 - db^2}{4} = 0.$$

The number $(a^2 - db^2)/4$ is a rational integer in this case because $a^2 \equiv b^2 \equiv d \equiv 1$ (mod 4).

Gauss systematically extended the ideas of number theory to algebraic integers, beginning with those of degree 2. He considered algebraic integers of the form $a + bi$, where a and b are rational integers. These are called Gaussian integers, and they share many properties with ordinary rational integers. For example, the sum, difference, or product of two Gaussian integers is a Gaussian integer. But the quotient of two Gaussian integers, although an algebraic number, need not be a Gaussian integer, as illustrated by the quotient $1/(1+i) = (1-i)/2$.

Properties of Gaussian integers

Divisibility of Gaussian integers (denoted here by Greek letters) is defined in the same way as for rational integers. An integer β is called a divisor of an integer α if there is an integer γ such that $\alpha = \beta\gamma$. If so, then γ is also a divisor (or factor) of α. A nonzero integer α is said to be prime if there are no integers β and γ, neither equal to ± 1 or $\pm i$, such that $\alpha = \beta\gamma$. This restriction on β and γ rules out such factorings as $1 + 2i = i(2 - i)$; if this were allowed there would be no primes. The four integers ± 1 and $\pm i$ are examples of units. More generally, an algebraic integer α (Gaussian or not) is called a unit if it is a divisor of 1, that is, if there is another algebraic integer β such that $\alpha\beta = 1$. The other factor, β, is also a unit. If α is a unit and if γ is any algebraic integer, then α and $\alpha\gamma$ are called associates. For example, every Gaussian integer γ has four associates, $\pm\gamma$ and $\pm i\gamma$.

Rational primes of the form $4k+3$ are also Gaussian primes, but the rational prime 2 and those of the form $4k+1$ can be factored as the product of two Gaussian primes. For example,

$$2 = (1+i)(1-i),$$
$$5 = (2+i)(2-i),$$
$$13 = (2+3i)(2-3i),$$
$$73 = (8+3i)(8-3i),$$

each factor being a Gaussian prime. In the factorization of 2, the two factors are associates because $1 + i = i(1 - i)$, but in the other examples the factors are not associates. The Gaussian primes consist of the rational primes $p \equiv 3$ (mod 4) and their associates, the Gaussian prime factors of 2 and their associates, and the Gaussian prime factors of the rational primes $p \equiv 1$ (mod 4) and their associates.

The unique factorization property holds for Gaussian integers. That is, every Gaussian integer α other than zero and the four units ± 1 and $\pm i$ can be factored into Gaussian primes in exactly one way, apart from the order of the factors and ambiguities between associated primes. Thus, if

$$\alpha = \pi_1 \pi_2 \cdots \pi_r$$

is one factorization into primes, then

$$\alpha = (\varepsilon_1 \pi_1)(\varepsilon_2 \pi_2) \cdots (\varepsilon_r \pi_r)$$

is essentially the same factorization if each of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_r$ is one of the four units and the product $\varepsilon_1 \varepsilon_2 \cdots \varepsilon_r = 1$.

If θ is a given nonzero algebraic number of degree n, the set of numbers of the form

$$q_0 + q_1\theta + q_2\theta^2 + \cdots + q_{n-1}\theta^{n-1},$$

where the coefficients $q_0, q_1, \ldots, q_{n-1}$ are rational numbers, is denoted by $Q(\theta)$ and is called an algebraic number field of degree n. It is the largest set of numbers obtainable from θ and the rational numbers by addition, subtraction, multiplication, and division (with division by zero excluded). The power θ^n and all higher powers of θ are already included in this set because of the basic equation that makes θ an algebraic number of degree n. The quotient of two numbers in this set (division by zero excluded) is also a number in the set. When $\theta = 1$, $Q(1)$ is a field of degree 1 and consists of all the rational numbers. If $\theta = \sqrt{d}$, where d is a square-free rational integer, then $Q(\sqrt{d})$ is a field of degree 2, called a real quadratic field if d is positive and an imaginary quadratic field if d is negative.

Unique factorization in quadratic fields

Algebraic number theory studies properties of the algebraic integers in a given field $Q(\theta)$. One of the basic questions is that of unique factorization. For the Gaussian integers, which lie in the imaginary quadratic field $Q(i)$, unique factorization holds. There is also unique factorization in the real quadratic fields $Q(\sqrt{d})$ for $d = 2, 3, 5, 6$, and 7 but not for $d = 10$. Gauss conjectured that there are infinitely many real quadratic fields with the unique factorization property, but this conjecture has not been proved.

Unique factorization also holds in the imaginary quadratic fields $Q(\sqrt{-d})$ for $d = 1, 2$, and 3 but not for $d = 5$. For example, 6 and 21 have two factorizations in $Q(\sqrt{-5})$,

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$$

and

$$21 = 3 \cdot 7 = (4 + \sqrt{-5})(4 - \sqrt{-5}),$$

the factors 2, 3, 7, $1 \pm \sqrt{-5}$, and $4 \pm \sqrt{-5}$ being primes in the field $Q(\sqrt{-5})$. Gauss noted that $Q(\sqrt{-d})$ has the unique factorization property in the nine cases

$$d = 1, 2, 3, 7, 11, 19, 43, 67, \text{ and } 163,$$

and he conjectured that there are no other imaginary quadratic fields with this property. Little progress was made on this problem until 1934, when Hans Heilbronn and Edward Linfoot proved that there is at most one more d not in this list. Computations by D.H. Lehmer established that d would have to be enormously large. In 1967 the American mathematician Harold M. Stark proved that there is no 10th case, and therefore Gauss's list of nine imaginary quadratic fields with unique factorization is complete. Such detailed analysis has not yet been achieved for general algebraic fields.

Algebraic integers and Diophantine equations. A significant difference between real and imaginary quadratic fields is revealed by the number of units in the field. Ev-

ery imaginary quadratic field has only a finite number of units, all of which are roots of unity, that is, solutions of the equation $x^n = 1$ for some n. For example, the units in the Gaussian field $Q(i)$ are ± 1 and $\pm i$, the roots of $x^4 = 1$. The only units in $Q(\sqrt{-2})$ are ± 1, whereas in $Q(\sqrt{-3})$ there are six units, the roots of $x^6 = 1$. By contrast, every real quadratic field has infinitely many units, each of which can be represented in the form $\pm\varepsilon^n$, where ε is a particular unit (called a fundamental unit) and $n = 0, \pm 1, \pm 2, \ldots$. For example, in the field $Q(\sqrt{2})$ the number $\varepsilon = 1 + \sqrt{2}$ is a fundamental unit.

Fundamental units in real quadratic fields can be surprisingly large. For example, the number $1{,}520 + 273\sqrt{31}$ is a fundamental unit in $Q(\sqrt{31})$, and the number

$$2{,}143{,}295 + 221{,}064\sqrt{94}$$

is one in $Q(\sqrt{94})$. One method for determining fundamental units uses continued fractions. Another, which will be described here, makes use of Diophantine equations. As noted in the previous section, the algebraic integers θ in $Q(\sqrt{d})$ consist of those of the form $\theta = a + b\sqrt{d}$, where a and b are rational integers, together with those of the form $\theta = (a + b\sqrt{d})/2$ with both a and b odd if $d \equiv 1 \pmod 4$. The other root of the quadratic equation satisfied by θ is $\theta' = a - b\sqrt{d}$ if $\theta = a + b\sqrt{d}$, or $\theta' = (a - b\sqrt{d})/2$ if $\theta = (a + b\sqrt{d})/2$. The algebraic integer θ is a unit in $Q(\sqrt{d})$ if $\theta\theta' = \pm 1$, which means $a^2 - db^2 = \pm 1$ or $a^2 - db^2 = \pm 4$ if both a and b are odd and $d \equiv 1 \pmod 4$. Therefore, finding the units in $Q(\sqrt{d})$ amounts to finding all solutions of the Diophantine equation $x^2 - dy^2 = \pm 1$, together with all odd solutions of $x^2 - dy^2 = \pm 4$ if $d \equiv 1 \pmod 4$. These Diophantine equations are called Pell equations, after the English mathematician John Pell (1610–1685). The name is the result of a misunderstanding on Euler's part, because Pell never studied these equations.

Pell equations

If $d > 0$ and $d \equiv 2$ or $3 \pmod 4$, the quadratic field $Q(\sqrt{d})$ has a fundamental unit $\varepsilon = x + y\sqrt{d}$ where x and y are the least positive solutions of the Pell equation $x^2 - dy^2 = -1$, if positive solutions exist; otherwise x and y are the least positive solutions of $x^2 - dy^2 = 1$, which always exist. A fundamental unit is similarly obtained by solving one of the Pell equations $x^2 - dy^2 = \pm 4$ if $d > 1$ and $d \equiv 1 \pmod 4$.

Algebraic number theory is often used to study Diophantine equations. For example, in the Diophantine equation $y^3 - x^2 = 1$, both x and y cannot be even and both cannot be odd because then the difference $y^3 - x^2$ would be even. Also, x cannot be odd when y is even because in that case $y^3 - x^2 \equiv -1 \pmod 4$. Therefore, in any solution x must be even and y odd. One such solution is $x = 0$, $y = 1$, and this is the only solution with $x = 0$. A method to show that there are no further solutions is to enlarge the setting to the field $Q(i)$ and write the equation in the form

$$y^3 = x^2 + 1 = (x - i)(x + i),$$

where x is assumed to be nonzero. Neither factor is a unit because $x \neq 0$. If the Gaussian integers $x + i$ and $x - i$ had a common divisor π among the primes of $Q(i)$, then π would divide their difference $2i$, so π would be either $1 + i$ or $1 - i$, the prime divisors of 2. In that case $\pi^2 = \pm 2i$, which is impossible because π^2 divides y^2 and y is odd. Unique factorization implies that both $x + i$ and $x - i$ are cubes of integers in $Q(i)$. Hence $x + i = (a + bi)^3$ for some choice of rational integers a and b, from which it follows, by equating real and imaginary parts, that $x = a^3 - 3ab^2$ and $1 = 3a^2b - b^3 = b(3a^2 - b^2)$. Thus, $b = \pm 1$ and $3a^2 - b^2 = \pm 1$, from which it follows that $b = -1$, $a = 0$. Thus, $x = 0$, $y = 1$ is the only solution in integers of $y^3 - x^2 = 1$.

A similar argument can be used to show that the Diophantine equation $y^3 = x^2 + 2$ has only two solutions in integers, $x = \pm 5$, $y = 3$. The argument makes use of unique factorization of integers in the field $Q(\sqrt{-2})$. There are only two units in this field, ± 1, which simplifies matters.

Unique factorization of integers in the field $Q(\sqrt{-3})$ can be used to show that the Diophantine equation $z^3 = x^3 + y^3$ has no solutions in positive integers. This field has six units, the three cube roots of 1 and their negatives. The roots are determined by solving the cubic equation

$$u^3 = 1, \text{ or } (u - 1)(u^2 + u + 1) = 0.$$

The solutions are 1, ω, and ω^2, where

$$\omega = \frac{-1 + \sqrt{-3}}{2} \text{ and } \omega^2 = \frac{-1 - \sqrt{-3}}{2}.$$

The number $\sqrt{-3}$ is a prime in this field. The Diophantine equation can now be written in the form

$$z^3 = (x + y)(x + \omega y)(x + \omega^2 y),$$

from which it can be shown that each factor on the right is the cube of an integer in the field $Q(\sqrt{-3})$ multiplied by a unit of the field and some power of the prime $\sqrt{-3}$. This information suffices to show that the equation $z^3 = x^3 + y^3$ has no solutions in positive integers. The argument makes essential use of unique factorization in the field $Q(\sqrt{-3})$.

In the mid-19th century the German mathematician Ernst Eduard Kummer and others attempted to extend this argument to prove the Fermat conjecture that for any positive integer $n > 2$ there are no solutions of $x^n + y^n = z^n$ in positive integers. It suffices to consider only the case in which n is an odd prime $p > 3$. The idea was to factor $x^p + y^p$ into linear factors and rewrite the equation in the form

$$z^p = (x + y)(x + \zeta y)(x + \zeta^2 y) \cdots (x + \zeta^{p-1} y),$$

where

$$\zeta = \cos\frac{2\pi}{p} + i \sin\frac{2\pi}{p}.$$

The number ζ satisfies $\zeta^p = 1$, so ζ is a pth root of 1. The field $Q(\zeta)$ is called a cyclotomic field. Kummer realized that if the factor $x + \zeta y$ has the form $\varepsilon\alpha^p$ for some algebraic integer α and some unit ε in the cyclotomic field $Q(\zeta)$, he could settle the Fermat conjecture by an argument similar to that described above for the case $p = 3$. In fact, he showed that the equation $x + \zeta y = \varepsilon\alpha^p$, with x and y not divisible by p, implies $x \equiv y \pmod p$. A similar argument applied to the equation $x^p + (-z)^p = (-y)^p$ shows that $x \equiv -z \pmod p$. Therefore

Cyclotomic fields

$$2x^p \equiv x^p + y^p \equiv z^p \equiv -x^p \pmod p,$$

so that $3x^p \equiv 0 \pmod p$. But this is a contradiction because $p > 3$ and p does not divide x.

The equation $x + \zeta y = \varepsilon\alpha^p$, which is crucial to the foregoing argument, is valid if unique factorization holds for the integers in the cyclotomic field $Q(\zeta)$. The smallest prime p for which unique factorization fails in $Q(\zeta)$ is $p = 23$, but for values of p for which unique factorization holds, Kummer's argument shows that the Fermat equation $x^p + y^p = z^p$ has no positive solutions.

Kummer was aware that the unique factorization property does not hold in every cyclotomic field, and he searched for another way to deduce the crucial equation $x + \zeta y = \varepsilon\alpha^p$. He succeeded, in part, by introducing a new kind of complex number, which he called an ideal complex number and which permitted him to generalize the concept of unique factorization. He was then able to prove that the Fermat equation $x^p + y^p = z^p$ has no solutions for all primes p satisfying two special properties, but he was not able to determine whether all primes have these properties. The German mathematician Richard Dedekind rephrased Kummer's ideas and introduced the concept of an ideal.

Ideals and class numbers of algebraic number fields. An ideal I is a set of algebraic integers in an algebraic number field K with the following three properties: I contains 0; if α and β are in I, their sum and difference are also in I; and, if α is in I and if β is any algebraic integer in K, then $\alpha\beta$ is in I.

For example, suppose $K = Q(1)$, the field of rational numbers. The set of all integer multiples of a fixed integer a is an ideal, denoted by (a) and called a principal ideal. Thus, the principal ideal (7) consists of all multiples of 7. Another ideal is the set of all linear combinations $7m + 3n$, where m and n run through all the rational integers. This ideal is denoted by (7, 3) and is said to be generated by 7 and 3. The ideal (14, 49), generated by 14 and 49, is easily seen to be the same as the principal ideal (7) because $7 = 49 - 3 \cdot 14$, a linear combination of 14 and 49.

In any algebraic number field K, the set of all algebraic integers in K is itself an ideal, in fact a principal ideal, denoted by (1). The ideal (α, β) generated by two algebraic integers α and β is the set of all linear combinations $\lambda\alpha + \mu\beta$, where λ and μ run through all algebraic integers in K.

If the algebraic integers in K have the unique factorization property, then every ideal is principal and the introduction of ideals is superfluous. But in a field where the unique factorization property fails, ideals can be used to restore unique factorization. This is done by introducing multiplication of ideals in a very natural way. The product AB of two ideals A and B in an algebraic number field K is the ideal obtained by taking all products $\alpha\beta$ of an element α in A with an element β in B, together with all sums and differences of such products. For example, the product of any ideal A with the principal ideal (1) is A, so the principal ideal (1) plays the role of the number 1 in the arithmetic of ideals.

A prime ideal is one that is not the principal ideal (1) and that is not the product of two ideals unless one factor is the principal ideal (1). Unique factorization holds for the algebraic integers in K if and only if all ideals are principal. The failure of unique factorization in the field $Q(\sqrt{-5})$ is illustrated by the example $6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$. Unique factorization can be restored by factoring the ideal (6) as a product of four prime ideals,

$$(6) = (2, 1 + \sqrt{-5})(2, 1 - \sqrt{-5})(3, 1 + \sqrt{-5})(3, 1 - \sqrt{-5}).$$

Unique factorization by prime ideals

More generally, in any algebraic number field K, every nonzero ideal other than the principal ideal (1) can be factored as a product of prime ideals, and this factorization is unique apart from the order of the factors.

For fields in which not all ideals are principal, Kummer introduced a way of measuring the extent to which unique factorization fails. He did this by considering two ideals A and B to be equivalent if there are two nonzero principal ideals (α) and (β) such that $(\alpha)A = (\beta)B$. With this definition of equivalence, the ideals of an algebraic number field are separated into equivalence classes, of which there are only finitely many, one of them being the class of principal ideals. This finite number of equivalence classes is called the class number of the field. The class number is 1 if and only if the algebraic integers in the field have the unique factorization property, which, of course, means that all ideals are principal. The size of the class number is a rough measure of how badly unique factorization fails to hold. Let $h(d)$ denote the class number of the quadratic field $Q(\sqrt{d})$. Some examples of class numbers are $h(-1) = 1$, $h(-5) = 2$, $h(-127) = 5$, $h(-329) = 24$, and $h(235) = 6$. Except for the first one, these numbers are not easy to calculate. It is known, however, that there are algebraic number fields with arbitrarily large class numbers.

Given any algebraic number field K with class number greater than 1, there is a larger field L containing K, called the class field of K, which has the significant property that every ideal in K becomes a principal ideal in L. This does not mean that L has class number 1, even though every ideal in K is a principal ideal in the larger field L, because some new ideals in L may have arisen that are not principal. Iteration of the process leading from K to the class field L gives the class field of L, and so on, ultimately producing a class field tower. It was hoped that the class field tower would be finite, terminating in a class field at the top of the tower with class number 1. However, in 1964 the Russian mathematicians Yevgeny Solomonovich Golod and Igor Rostislavovich Shafarevich gave an example of a starting field K that had an infinite class field tower.

The class field tower

ANALYTIC NUMBER THEORY

Although most of the following discussion requires only a good knowledge of algebra, some arguments draw on concepts from elementary calculus, such as limits, summations, the natural logarithm, the exponential function, infinite series, and infinite products. In these sections the symbol $\log x$ is used for the natural logarithm of x (the logarithm of x to the base e). The notation $\zeta(s)$ represents the sum of the infinite series

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s},$$

which converges if $s > 1$.

The order of magnitude of arithmetical functions. Many investigations in number theory are concerned with the behaviour of arithmetical functions $f(n)$ for large values of n. The difficulty in determining this behaviour is exemplified by $d(n)$, the number of divisors of n, whose values fluctuate considerably. This function has the value 2 whenever n is prime; it therefore has the value 2 infinitely often. It also takes on arbitrarily large values when n has a large number of divisors. For example, if n is the product of the first k primes, say $n = 2 \cdot 3 \cdot 5 \cdot \cdots p_k$, then $d(n) = 2^k$ because $d(n)$ is multiplicative and $d(p) = 2$ for each prime p. If k is large, the prime number theorem (see below) implies that k is approximately $\log n/\log \log n$, so that $d(n)$ is approximately $2^{\log n/\log\log n}$. This idea leads to the following inequalities for $d(n)$: For every positive number ε there is an integer $N(\varepsilon)$(depending on ε) such that

$$d(n) < 2^{(1+\varepsilon)\log n/\log\log n}$$

for all $n \geq N(\varepsilon)$, whereas

$$d(n) > 2^{(1-\varepsilon)\log n/\log\log n}$$

for infinitely many n. There are analogous inequalities for the Euler function $\varphi(n)$ involving a certain positive constant c less than 1. For every positive number ε there exists an integer $N(\varepsilon)$ such that

$$\varphi(n) > (1 - \varepsilon)\frac{cn}{\log\log n}$$

for all $n \geq N(\varepsilon)$, whereas

$$\varphi(n) < (1 + \varepsilon)\frac{cn}{\log\log n}$$

for infinitely many n.

For these and other examples, it is more fruitful to study the arithmetic mean,

$$\hat{f}(n) = \frac{1}{n}\sum_{k=1}^{n} f(k),$$

because fluctuations tend to cancel one another on the average. For example, Dirichlet showed that the average of the divisor function increases at the same rate as $\log n$ because the ratio $\hat{d}(n)/\log n$ approaches 1 as a limit as n approaches infinity (denoted $n \to \infty$). This is a consequence of a formula of Dirichlet which states that

$$\sum_{k=1}^{n} d(k) = n \log n + (2C - 1)n + O(\sqrt{n}),$$

where $C = 0.57721\ldots$ is Euler's constant, and the symbol $O(\sqrt{n})$ represents an unspecified function of n whose absolute value is less than some constant times $\sqrt{n}$. The principal term, $n \log n$, is called the asymptotic value of the sum; the other two terms are of smaller order and represent the error made by approximating the sum by its asymptotic value. The error term $O(\sqrt{n})$ has been improved to $O(n^{\theta})$ for various values of θ less than $1/2$, but θ cannot be replaced by $1/4$. Determining the smallest exponent θ in the error term is an unsolved problem called the Dirichlet divisor problem. By 1988 it was known that θ can be any number greater than $7/22$.

The Dirichlet divisor problem

Asymptotic formulas of this type have been found for many arithmetical functions. The following examples are typical, although in many cases the error terms have been improved beyond what appears here.

$$\sum_{k=1}^{n} \varphi(k) = \frac{3}{\pi^2}n^2 + O(n \log n)$$

$$\sum_{k=1}^{n} \sigma_1(k) = \frac{\pi^2}{12}n^2 + O(n \log n)$$

$$\sum_{k=1}^{n} \sigma_a(k) = \frac{\zeta(a+1)}{a+1}x^{a+1} + O(x^a) \qquad (a > 1)$$

$$\sum_{k=1}^{n} r_2(k) = \pi n + O(\sqrt{n})$$

$$\sum_{k=1}^{n} |\mu(k)| = \frac{6}{\pi^2} n + O(\sqrt{n})$$

The last formula contains the absolute value of the Möbius function, $|\mu(k)|$, not $\mu(k)$ itself. Determining the asymptotic behaviour of the summatory function

$$M(n) = \sum_{k=1}^{n} \mu(k)$$

is one of the most difficult unsolved problems in number theory. For a long time it was thought that perhaps $|M(n)| < \sqrt{n}$, but in 1984 Andrew Odlyzko and Herman te Riele showed that this inequality is violated infinitely often. It has been conjectured that the inequality $|M(n)| < n^{1/2+\varepsilon}$ holds for every $\varepsilon > 0$. If this were true it would have important implications concerning the distribution of primes. It is known that $M(n)/n \to 0$ as $n \to \infty$, a result equivalent to the prime number theorem.

The prime number theorem. The search for the order of magnitude of two arithmetical functions related to primes led to one of the most profound theorems in number theory, the prime number theorem. If p_n denotes the nth prime number and $\pi(n)$ denotes the number of primes in the interval from 1 to n, Euclid's theorem on the infinitude of primes guarantees that $\pi(n) \to \infty$ as $n \to \infty$. It is natural to ask how p_n and $\pi(n)$ grow as functions of n for large n. Irregularities in the distribution of primes and the lack of a simple formula for determining primes suggest that precise answers to these questions might be difficult or impossible to obtain. It seems astonishing to learn that answers can be obtained, and they are remarkably simple: For large n, the nth prime p_n grows as $n \log n$, whereas $\pi(n)$ grows as $n/\log n$. Each of these statements implies the other, and each is known as the prime number theorem. More precisely, the prime number theorem can be stated in two equivalent forms:

The distribution of primes

$$\lim_{n\to\infty} \frac{\pi(n)}{n/\log n} = 1, \tag{1}$$

and

$$\lim_{n\to\infty} \frac{p_n}{n \log n} = 1. \tag{2}$$

These relations are written more simply as $\pi(n) \sim n/\log n$ and $p_n \sim n \log n$ and are expressed verbally by saying that $\pi(n)$ is asymptotic to $n/\log n$ and that p_n is asymptotic to $n \log n$, respectively.

The first form was conjectured independently by Gauss and by Legendre around 1800, but neither provided proof. Gauss was led to his conjecture by examining a table of primes $\leq 10^6$. He calculated $\pi(n)$ simply by counting the number of primes in the interval from 1 to n. This is similar to the exercise done in an earlier section, in which primes were counted in blocks of 1,000 consecutive integers. However, if the blocks are not of equal length but are allowed to grow by a factor of 100, the following values of $\pi(n)$ and of the ratio $n/\pi(n)$ are obtained:

n	$\pi(n)$	$n/\pi(n)$
10^2	25	4.000
10^4	1,229	8.137
10^6	78,498	12.739
10^8	5,761,455	17.537
10^{10}	455,052,511	21.975
10^{12}	37,607,912,018	26.590
10^{14}	3,204,941,750,802	31.202

The table shows that the ratio $n/\pi(n)$ grows very slowly compared to n. As the exponents on the powers of 10 increase by 2, the ratio $n/\pi(n)$ seems to increase by about 4.6, or 2.3 times the exponent of 10. The exponent of 10 is the logarithm of n to the base 10, so the table indicates that $n/\pi(n)$ grows at about 2.3 times this logarithm. But 2.3 times the logarithm of a number to base 10 is approximately equal to the natural logarithm of that number. Comparison of the natural logarithm $\log n$ with the ratio $n/\pi(n)$ reveals the following values:

n	$\log n$	$n/\pi(n)$	$\dfrac{\log(n)}{n/\pi(n)}$
10^2	4.605	4.000	1.151
10^4	9.210	8.137	1.132
10^6	13.816	12.379	1.085
10^8	18.421	17.357	1.061
10^{10}	23.026	21.975	1.048
10^{12}	27.631	26.590	1.039
10^{14}	32.236	31.202	1.033

This remarkable agreement between $\log n$ and $n/\pi(n)$ suggests that their ratio approaches 1 as $n \to \infty$. Gauss, Legendre, and many other eminent mathematicians of the early 19th century tried unsuccessfully to prove this conjecture. The first step toward a proof was made in 1851 by Chebyshev, who showed that if $\pi(n)\log n/n$ has a limit as $n \to \infty$, then this limit must equal 1.

In 1859 Bernhard Riemann attacked the problem with a new method, using a formula of Euler relating the sum of the reciprocals of the powers of the positive integers with an infinite product extended over the primes. Euler's product formula, devised in 1737, states that, for every real $s > 1$,

$$\sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{n=1}^{\infty} \frac{1}{1 - \frac{1}{p_n^s}},$$

where the product runs through all the primes. Euler derived this formula by writing each factor on the right as an infinite geometric series

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots,$$

with $x = p_n^{-s}$. When all these series are multiplied together and the denominators are arranged in increasing order, the result is the series on the left (because of the fundamental theorem of arithmetic).

Riemann replaced the real variable s in Euler's product formula with a complex number and showed that the distribution of prime numbers is intimately related to properties of the function $\zeta(s)$ defined by the series (now called the Riemann zeta function). Riemann came close to proving the prime number theorem, but not enough was known during his lifetime about the theory of functions of a complex variable to complete the proof successfully.

Riemann's zeta function

Thirty years later the necessary analytic tools were at hand, and in 1896 Jacques-Salomon Hadamard and Charles Jean de la Vallée-Poussin independently proved the first form of the prime number theorem. The proof was one of the great achievements of analytic number theory. Subsequently, new proofs were discovered, including an elementary proof found in 1949 by Paul Erdős and Atle Selberg that makes no use of complex function theory.

The second form of the prime number theorem is easily deduced from the first form by taking logarithms of both members of equation (1) and then removing a factor $\log n$ to obtain

$$\lim_{n\to\infty} \left\{ \log n \left(\frac{\log \pi(n)}{\log n} + \frac{\log \log n}{\log n} - 1 \right) \right\} = 0.$$

Because $\log n \to \infty$, the factor multiplying $\log n$ must tend to zero. The quotient $\log \log n/\log n$ also tends to zero, hence

$$\lim_{n\to\infty} \frac{\log \pi(n)}{\log n} = 1.$$

This relation, multiplied by equation (1), yields

$$\lim_{n\to\infty} \frac{\pi(n) \log \pi(n)}{n} = 1.$$

Now replace n by the nth prime, p_n, in this equation. Then $\pi(p_n) = n$, and the equation becomes

$$\lim_{n\to\infty} \frac{n \log n}{p_n} = 1,$$

which implies equation (2). Equation (1) can also be

deduced from (2), so the two statements are equivalent.

Each of the following statements involving the Möbius function is also equivalent to the prime number theorem:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mu(k) = 0, \text{ and } \lim_{n\to\infty} \sum_{k=1}^{n} \frac{\mu(k)}{k} = 0.$$

The prime number theorem is important not only because it makes an elegant and simple statement about primes and has many applications but also because much new mathematics was created in the attempts to find a proof. This is typical in number theory. Some problems, very simple to state, are often extremely difficult to solve, and mathematicians working on these problems often create new areas of mathematics of independent interest. Another example is the creation of algebraic number theory as a result of work on the Fermat conjecture.

Primes in arithmetic progressions. An arithmetic progression of integers with first term h and common difference k consists of all numbers of the form $kn + h$ as n runs through the nonnegative integers 0, 1, 2, ... If h and k have a common prime factor p, each term of the progression is divisible by p, and there can be no more than one prime in the progression. In 1803 Legendre considered whether there must be infinitely many primes in the progression if h and k have no common prime factor. For example, all odd numbers fall into two progressions, one containing numbers of the form $4n - 1$, and the other containing numbers of the form $4n + 1$, so at least one of these progressions must contain infinitely many primes. In fact, both of these progressions contain infinitely many primes. This is easily proved for the progression $4n - 1$ by a minor modification of the Euclidean proof of the infinitude of primes. The proof is as follows: For any finite set S of primes of the form $4n - 1$, let p be the largest prime in S, and consider the integer $N = 4 \cdot 3 \cdot 5 \cdot \cdots p - 1$, where the product $3 \cdot 5 \cdot \cdots p$ contains all the odd primes $\le p$ as factors. Because N is greater than p it cannot be in S. No prime $\le p$ divides N, and so all the prime factors of N must exceed p. But all the prime factors of N cannot be of the form $4n + 1$ because the product of two such numbers, and therefore the product of any finite collection of such numbers, is again of the same form. Hence there must be some prime factor of N of the form $4n - 1$ that is not in S. This proves that there must be an infinite number of primes of the form $4n - 1$.

This argument breaks down for the progression $4n + 1$, because the product of two numbers in this progression is again of the form $4n + 1$. But an elementary argument of a different type can be used to show that the progression $4n + 1$ also contains infinitely many primes. There are also elementary arguments showing that there are infinitely many primes in other special arithmetic progressions such as $5n - 1$, $8n - 1$, $8n - 3$, and $8n + 3$. For any positive integer k, the sequence $(kn + 1 : n = 1, 2, 3, 4, \ldots)$ contains infinitely many primes. This special case of Dirichlet's theorem has been proved by elementary methods, but the proof is not as simple as in the other cases cited.

In a celebrated memoir published in 1837, Dirichlet showed that every arithmetic progression $kn + h$, where h and k are relatively prime, contains infinitely many primes. This was the first major triumph of analytic number theory. Dirichlet was guided by Euler's proof of Euclid's theorem on the infinitude of primes, which shows that the sum of the reciprocals of the primes diverges. Dirichlet used an ingenious argument to show that the sum of the reciprocals of all the primes in the progression $kn + h$ also diverges. The method can be understood by referring to the following asymptotic formula for the partial sums of the series of reciprocals of the primes:

$$\sum_{p\le x} \frac{1}{p} = \log(\log x) + O(1),$$

where the sum is extended over all primes $p \le x$, and the symbol $O(1)$ represents a quantity that remains bounded for all x. This formula shows that the partial sums are asymptotic to $\log(\log x)$ and therefore tend to ∞ as $x \to \infty$. Dirichlet's method gives a corresponding formula for sums extended over those primes $p \le x$ that also lie in the arithmetic progression $kn + h$:

$$\sum_{\substack{p\le x \\ p\equiv h \pmod k}} \frac{1}{p} = \frac{1}{\varphi(k)} \log(\log x) + O(1),$$

where the notation $p \equiv h \pmod k$ indicates that each prime in the sum is congruent to h mod k. Again, the principal term on the right tends to ∞ as $x \to \infty$, so there must be infinitely many primes $p \equiv h \pmod k$, hence infinitely many in the progression $kn + h$. Because the principal term is independent of h, the formula also shows that each of the $\varphi(k)$ residue classes mod k makes the same contribution to the principal term.

The key to Dirichlet's method was the introduction of new arithmetical functions $\chi(n)$, now called Dirichlet characters mod k. These are multiplicative functions that have period k and vanish only on the numbers not relatively prime to k. There are exactly $\varphi(k)$ Dirichlet characters mod k, and they make it possible to extract the integers congruent to h mod k from a given set of integers.

Dirichlet's memoir also introduced functions analogous to the Riemann zeta function $\zeta(s)$ of prime number theory. These functions, now called Dirichlet L-functions, are defined by infinite series of the form

Dirichlet L-functions

$$L(s, \chi) = \sum_{n=1}^{\infty} \frac{\chi(n)}{n^s},$$

where $\chi(n)$ is a Dirichlet character mod k and s is a real number greater than 1 (or a complex number with real part greater than 1). The study of L-functions is an active area of contemporary mathematical research that has profound applications to both analytic and algebraic number theory.

Generating functions. Dirichlet's work on L-functions led to the description of a more general class of functions defined by infinite series of the form

$$F(s) = \sum_{n=1}^{\infty} \frac{f(n)}{n^s},$$

where $f(n)$ is a given arithmetical function. This is called a Dirichlet series with coefficients $f(n)$, and the function $F(s)$ is called a generating function of the coefficients. The prototype is the Riemann zeta function $\zeta(s)$, which generates the constant function $f(n) = 1$ for all n,

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

The square of the zeta function generates the divisor function $d(n)$,

$$\zeta(s)^2 = \sum_{n=1}^{\infty} \frac{d(n)}{n^s},$$

and the reciprocal of the zeta function generates the Möbius function $\mu(n)$,

$$\frac{1}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\mu(n)}{n^s}.$$

More generally, if a Dirichlet series $F(s)$ generates an arithmetical function $f(n)$, and another Dirichlet series $G(s)$ generates $g(n)$, then the product $F(s)G(s)$ generates the arithmetical function

$$h(n) = \sum_{d|n} f(d) g\left(\frac{n}{d}\right),$$

called the Dirichlet product of the two arithmetical functions f and g. The appearance of the Dirichlet product is seen by multiplying the two Dirichlet series and rearranging terms, as follows:

$$F(s)G(s) = \sum_{n=1}^{\infty} \frac{f(n)}{n^s} \sum_{m=1}^{\infty} \frac{g(m)}{m^s}$$

$$= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \frac{f(n)g(m)}{(nm)^s} = \sum_{k=1}^{\infty} \frac{h(k)}{k^s},$$

where

$$h(k) = \sum_{nm=k} f(n)g(m) = \sum_{d|k} f(d) g\left(\frac{k}{d}\right).$$

The results are valid whenever the series for $F(s)$ and $G(s)$ converge absolutely.

Products of Dirichlet series can be used to generate many arithmetical functions occurring in multiplicative number theory. For example,

$$\frac{\zeta(s-1)}{\zeta(s)} = \sum_{n=1}^{\infty} \frac{\varphi(n)}{n^s}$$

and

$$\zeta(s)\zeta(s-a) = \sum_{n=1}^{\infty} \frac{\sigma_a(n)}{n^s}.$$

In additive number theory it is more convenient to use generating functions represented by power series. A function $A(x)$ with the power series expansion

$$A(x) = \sum_{n=0}^{\infty} a(n)x^n$$

is said to generate the coefficients $a(n)$. If $B(x)$ generates the coefficients $b(m)$, so that

$$B(x) = \sum_{m=0}^{\infty} b(m)x^m,$$

then, for those x for which both series converge absolutely, the product of the two series is given by

$$A(x)B(x) = \sum_{n=0}^{\infty} a(n)x^n \sum_{m=0}^{\infty} b(m)x^m$$

$$= \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} a(n)b(m)x^{n+m} = \sum_{k=0}^{\infty} c(k)x^k,$$

where

$$c(k) = \sum_{n+m=k} a(n)b(m).$$

The new sequence $c(k)$ is called the Cauchy product of the two sequences $a(n)$ and $b(m)$.

For example, suppose $a(n) = 1$ if n is a square, and $a(n) = 0$ otherwise, so that

$$A(x) = 1 + x + x^4 + x^9 + x^{16} + \cdots + x^{n^2} + \cdots.$$

Thus, $A(x)$ involves only exponents that are squares. The square of $A(x)$ is given by the series

$$A(x)^2 = 1 + 2x + x^2 + 2x^4 + 2x^5 + x^8 + 2x^9 + \cdots + b(n)x^n + \cdots,$$

where $b(n)$ is the number of ways that n can be expressed as the sum of two squares of nonnegative integers, with the order of the summands taken into account. The fourth power of $A(x)$ is given by

$$A(x)^4 = 1 + 4x + 6x^2 + 4x^3 + 5x^4 + 12x^5 + \cdots + c(n)x^n + \cdots,$$

where $c(n)$ is the number of ways that n can be expressed as the sum of four nonnegative squares. For example, there are exactly six ways of writing 2 as the sum of four nonnegative squares:

Integers expressed as sums of squares

$$\begin{aligned} 2 &= 1+1+0+0 = 1+0+1+0 \\ &= 1+0+0+1 = 0+1+0+1 \\ &= 0+1+1+0 = 0+0+1+1, \end{aligned}$$

hence $c(2) = 6$. Euler observed that Lagrange's four squares theorem is equivalent to the statement that $c(n)$ is positive for every n, but he did not succeed in showing that $c(n) > 0$.

If squares of negative numbers are also allowed among the summands, the number of representations of an integer as a sum of squares is much greater. For example, if squares of -1 are allowed, there are 24 ways to represent 2 as a sum of four squares. The number of representations of an integer n as a sum of k squares of integers (positive, negative, or zero) is denoted by $r_k(n)$. Its generating function is

$$\theta(x)^k = 1 + \sum_{n=1}^{\infty} r_k(n)x^n,$$

where $\theta(x)$, called a theta function, is defined by the infinite series

$$\theta(x) = \sum_{n=-\infty}^{\infty} x^{n^2}.$$

The example $r_2(n)$ was mentioned above.

Jacobi showed that properties of theta functions often lead to information concerning representations of integers by squares. For example, one of Jacobi's identities from the theory of elliptic functions implies that

$$\theta(x)^2 = 1 + 4\sum_{n=1}^{\infty} \{d_1(n) - d_3(n)\}x^n,$$

where $d_1(n)$ and $d_3(n)$ are the number of divisors of n congruent to 1 and 3 mod 4, respectively. When the coefficient of x^n in this formula is compared with the coefficient of x^n in the generating function for $r_2(n)$, it gives the explicit formula

$$r_2(n) = 4\{d_1(n) - d_3(n)\}$$

for the number of representations of n as the sum of two squares. By similar methods, Jacobi also showed that $r_4(n) = 8\sigma(n)$ if n is odd, whereas $r_4(n) = 24$ times the sum of the odd divisors of n if n is even. Explicit formulas for $r_k(n)$ have been obtained by similar methods for $k = 6$, 8, 10, and 12, but they are more complicated. Explicit formulas for $r_k(n)$ are known for all even $k \leq 24$. Formulas for odd values of k are more difficult to derive.

Additive number theory. In contrast to multiplicative number theory, which is concerned with properties arising from prime factorization, additive number theory is concerned with arithmetical functions related to additive properties of the integers. The basic problem of additive number theory is that of expressing a given positive integer n as a sum of integers from some set

$$A = \{a_1, a_2, \ldots\},$$

where the summands a_i are special numbers such as primes, squares, cubes, and so on. Each representation of n as a sum of elements from A is called a partition of n. The number of such partitions is often of great interest.

The problem of representation by squares was discussed in the previous section. Another famous problem of additive number theory is the Goldbach conjecture, formulated in 1742, which states that every even number greater than 4 is the sum of two odd primes. This conjecture has been confirmed numerically up to very large numbers but remains unproved. In 1937 the Russian mathematician Ivan Matveyevich Vinogradov proved that every sufficiently large odd number is the sum of three odd primes, and in 1966 the Chinese mathematician Chen Jing-run (Ch'en Ching-jun) showed that every sufficiently large even integer is the sum of a prime plus a number with no more than two prime factors.

Waring's problem is named after the English mathematician Edward Waring, who in 1770 stated, without proof and with limited numerical evidence, that every positive integer n is the sum of four squares, of nine cubes, of 19 fourth powers, and so on. The problem is to determine, for a given positive integer k, whether there is an integer s (depending only on k) such that the equation

Waring's problem

$$n = x_1^k + x_2^k + \cdots + x_s^k$$

has integer solutions for all $n \geq 1$. The least value of s that exists for a given k is denoted by $g(k)$. Similarly, $G(k)$ denotes the smallest s required to so represent all sufficiently large n.

In 1770 Lagrange proved that $g(2) = 4$, and during the next 139 years the existence of $g(k)$ was shown for $k = 3$, 4, 5, 6, 7, 8, and 10. In 1909 the German mathematician David Hilbert proved the existence of $g(k)$ for every k but did not determine its numerical value. The exact value of $g(k)$ is now known for every k. For example, $g(3) = 9$, $g(4) = 19$, $g(5) = 37$, $g(6) = 73$, $g(7) = 143$, and $g(8) = 279$. A general formula states that

$$g(k) \geq 2^k + \left[\frac{3^k}{2^k}\right] - 2$$

for all $k \geq 2$, with equality if $4 \leq k \leq 200{,}000$.

The existence of $G(k)$ follows from that of $g(k)$ because $G(k) \leq g(k)$, but only the values $G(2) = 4$ and $G(4) = 16$

are known exactly. Some upper bounds smaller than $g(k)$ are known, for example, $G(3) \leq 7$, $G(5) \leq 23$, $G(6) \leq 36$, $G(7) \leq 53$, and $G(8) \leq 73$. The British mathematicians Godfrey Harold Hardy and John E. Littlewood conjectured that $G(k) < 2k + 1$ when k is not a power of 2 and that $G(k) \leq 4k$ when it is, but the most that is known is $G(k) < ck \log k$ for some constant c.

Unrestricted partition function

Perhaps the most fundamental arithmetical function in additive number theory is the unrestricted partition function $p(n)$, which counts the number of ways that a positive integer n can be written as a sum of natural numbers not exceeding n. For example, $p(5) = 7$ because there are 7 partitions of 5, given by

$$\begin{aligned} 5 &= 4 + 1 \\ &= 3 + 2 \\ &= 3 + 1 + 1 \\ &= 2 + 2 + 1 \\ &= 2 + 1 + 1 + 1 \\ &= 1 + 1 + 1 + 1 + 1. \end{aligned}$$

In counting partitions, the number of summands (parts) is unrestricted, repetition is allowed, and the order of the summands is not taken into account.

Euler showed that the reciprocal of the infinite product

$$\varphi(x) = \prod_{m=1}^{\infty} (1 - x^m)$$

is a generating function for $p(n)$:

$$\frac{1}{\varphi(x)} = \sum_{n=0}^{\infty} p(n)x^n,$$

where $p(0) = 1$. He also showed that

$$\varphi(x) = 1 - x - x^2 + x^5 + x^7 - x^{12} - x^{15} + \cdots,$$

where the exponents 1, 2, 5, 7, 12, 15, ... are called pentagonal numbers; they can be derived from the expressions $(3k^2 - k)/2$ and $(3k^2 + k)/2$ by taking $k = 1, 2, 3, 4, \ldots$. By multiplying the power series for $\varphi(x)$ with that for its reciprocal $1/\varphi(x)$, Euler obtained the following recursion formula for computing $p(n)$:

$$p(n) = p(n-1) + p(n-2) - p(n-5) - p(n-7) + \cdots,$$

where $p(k) = 0$ if $k < 0$. Another recursion formula, obtained by logarithmic differentiation of the generating function $1/\varphi(x)$, reveals a remarkable connection between $p(n)$ and the function $\sigma(n)$ of multiplicative number theory:

$$np(n) = \sum_{k=1}^{n} \sigma(k)p(n-k).$$

The function $p(n)$ grows rapidly. For example, $p(10) = 42$, $p(100) = 190{,}569{,}292$, and $p(200) = 3{,}972{,}999{,}029{,}388$. In 1917 Hardy and the Indian mathematician Srinivasa Ramanujan used the transformation theory of elliptic modular functions to show that $p(n)$ has the asymptotic value

$$p(n) \sim \frac{e^{K\sqrt{n}}}{4n\sqrt{3}},$$

where $K = \pi\sqrt{2/3}$. In 1937 the German-American mathematician Hans Rademacher showed that $p(n)$ could be expressed as the sum of a convergent series, the first term of which gives the foregoing asymptotic value.

Ramanujan discovered a number of remarkable identities that imply divisibility properties of the partition function. For example, the Ramanujan identity

$$5\frac{\varphi(x^5)^5}{\varphi(x)^6} = \sum_{m=0}^{\infty} p(5m+4)x^m$$

implies that $p(5m + 4) \equiv 0 \pmod 5$. He also found that $p(7m + 5) \equiv 0 \pmod 7$ and $p(11m + 6) \equiv 0 \pmod{11}$.

Investigations on the partition function in the early 20th century initiated a new era of research that has revealed profound connections among elliptic modular functions, abstract algebra, and number theory.

Lattice point problems. In a rectangular coordinate system, each point in the plane is represented by a pair of real numbers (x, y), its rectangular coordinates. Those points with integer coordinates are called lattice points, and they form a configuration of equally spaced rows and columns called the unit square lattice. Since the time of Gauss, many interesting and deep properties of the integers have been discovered by studying the unit square lattice. These investigations, together with their extension to higher-dimensional space, led to the creation of a new branch of mathematics called geometric number theory, or the geometry of numbers. This section describes a small sample of theorems concerning lattice points and illustrates some of the methods used to prove them. Most of the discussion is concerned with lattice points in the plane, with some indication of how the results are extended to spaces of higher dimension.

If a plane is covered with unit squares, like those on graph paper, the vertices of the squares form the unit square lattice. The same lattice can be generated by covering the plane with parallelograms having lattice points as vertices, provided there are no further lattice points inside or on the boundary of each parallelogram. Any parallelogram whose vertices generate the unit square lattice has area equal to 1 (that is, equal to the area of one unit square of the lattice). This fact is easily deduced from a theorem, discovered by the Austrian mathematician Georg Pick, concerning polygons whose vertices are lattice points (lattice polygons). The theorem states that the area of the region enclosed by any lattice polygon is one less than the number of lattice points inside the polygon plus half the number on the boundary. An example is shown in Figure 1. Therefore, any lattice parallelogram that generates the unit square lattice has area 1, because there are no lattice points inside the parallelogram and exactly four on the boundary.

Pick's theorem

Pick's theorem also implies that every regular lattice polygon must necessarily be a square. Thus, for example, there is no equilateral lattice triangle. This can be proved by calculating the area of such a polygon in two ways. If the regular polygon has n sides and is subdivided into n isosceles triangles with a common vertex at the center, its area is a rational multiple of $\cot(\pi/n)$. However, Pick's theorem implies that the area is rational, because the area is either an integer or an integer plus $\frac{1}{2}$; this implies that $\cot(\pi/n)$ must be a rational number. But it is known that, for $n \geq 3$, $\cot(\pi/n)$ is rational only for $n = 4$. For example, $\cot(\pi/3) = 1/\sqrt{3}$, an irrational number. Thus, there is no equilateral lattice triangle.

Because lattice polygons are special figures, it is not surprising that there is a simple formula for determining their areas by counting lattice points. But no three-dimensional analogue of Pick's formula can exist for finding the volume of a lattice polyhedron by counting lattice points inside, on the faces, and on the vertices. This can be demonstrated with a tetrahedron whose vertices are at the lattice points (0, 0, 0), (0, 1, 0), (0, 0, 1), and (1, 1, k), where k is a positive integer. There are lattice points at the four vertices but none inside, on the faces of, or on the edges of this tetrahedron. As k increases the volume of the tetrahedron increases, but the number of lattice points inside the figure or on its boundary does not change. Therefore, no three-dimensional analogue of Pick's formula can exist.

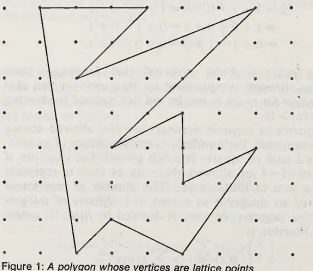


Figure 1: *A polygon whose vertices are lattice points.* It contains 15 lattice points inside and 18 on the boundary. By Pick's theorem the area of the region enclosed is $15 + 9 - 1 = 23$ (see text).

The number of lattice points inside a plane figure can change quite drastically if the figure is translated, even by a small amount. For example, a rectangle with vertices at the four lattice points (0, 0), (3, 0), (3, 1), and (0, 1) has exactly eight lattice points on its boundary and none inside. If this figure is shifted slightly to the right, the number of lattice points on the boundary decreases to six, again with none inside. If the figure is now shifted upward a small amount, the number of lattice points on the boundary drops to zero and the number inside increases to three. This simple example shows that the number of lattice points captured in a given region will depend on where the region is placed on the unit square lattice. It also shows that relations between areas of plane regions and numbers of lattice points captured may be harder to derive if the regions are not bounded by lattice polygons.

Blichfeldt's theorem

In 1914 the American mathematician Hans Frederick Blichfeldt discovered a remarkable theorem relating areas and lattice points: Any bounded plane region with positive area A placed in any position on the unit square lattice can be translated so that the number of lattice points inside the region will be at least A. The theorem allows the shape of the region to be arbitrary. The only requirements are that the region be bounded (which means that it can be enclosed inside some large square) and have positive area. If the area A is an integer, say $A = 14$, the theorem asserts that the region can be translated so that it encloses at least 14 lattice points inside. If A is not an integer, say $A = 14.2$, the region can be translated so that the number of lattice points inside is at least 15, the next integer larger than A. An example is shown in Figure 2.

Blichfeldt's theorem can be extended to higher-dimensional space. The version for three-dimensional figures states that any bounded solid region with positive volume V placed on the unit cubic lattice in any position can be translated so that the number of lattice points inside the region will be at least V. There is also an n-dimensional version in which V is the n-dimensional volume.

Minkowski's convex body theorem

The use of geometric methods in number theory was introduced in 1889 by Hermann Minkowski when he created a subject he called the "geometry of numbers." The cornerstone of this theory is his convex body theorem: A bounded plane convex region, symmetric about a lattice point and with area greater than 4, must contain at least three lattice points in its interior. (A region is called convex if for every pair of points in the region the line segment joining them also lies in the region. Familiar examples of convex regions are those bounded by parallelograms, circles, or ellipses. A region is symmetric about a point P if for every point Q in the region, Q' is also present, where P is the midpoint of the straight line segment QQ'.)

Minkowski's theorem can be deduced from Blichfeldt's theorem. Denote the convex plane region by K and assume that its centre of symmetry is at the origin. Denote by $K/2$ the region obtained by multiplying all the coordinates of points of K by $\frac{1}{2}$. The new region $K/2$ has the same shape as K but is only half as large. In the language of geometry, $K/2$ is similar to K with similarity ratio $\frac{1}{2}$. Therefore, the region $K/2$ is also convex and symmetric about the origin, but its area is $\frac{1}{4}$ that of K, because areas of similar figures are related by the square of the similarity ratio. The area of K is greater than 4, so the area of $K/2$ is greater than 1. By Blichfeldt's theorem, some translation of $K/2$ contains at least two distinct lattice points in its interior. Denote these lattice points by $\boldsymbol{x}$ and $\boldsymbol{y}$ in vector notation, and let $\boldsymbol{a}$ be the vector through which $K/2$ was translated to capture $\boldsymbol{x}$ and $\boldsymbol{y}$. (See Figure 3.) The lattice points $\boldsymbol{x}$ and $\boldsymbol{y}$ came from two points $\boldsymbol{x} - \boldsymbol{a}$ and $\boldsymbol{y} - \boldsymbol{a}$ interior to $K/2$. These are not necessarily lattice points because the translation vector $\boldsymbol{a}$ need not have integer coordinates. Because $K/2$ is symmetric about the origin, the point $\boldsymbol{a} - \boldsymbol{x}$ is also interior to $K/2$. Now $K/2$ is convex, so the point midway between $\boldsymbol{y} - \boldsymbol{a}$ and $\boldsymbol{a} - \boldsymbol{x}$, that is, the point $(\boldsymbol{y} - \boldsymbol{x})/2$, is also interior to $K/2$. Therefore, the point with coordinates twice as large, $\boldsymbol{y} - \boldsymbol{x}$, is interior to K. But $\boldsymbol{y} - \boldsymbol{x}$ is a lattice point, different from the origin. This lattice point, together with the symmetric point $\boldsymbol{x} - \boldsymbol{y}$ and the origin, are three lattice points interior to K, and this proves Minkowski's theorem.

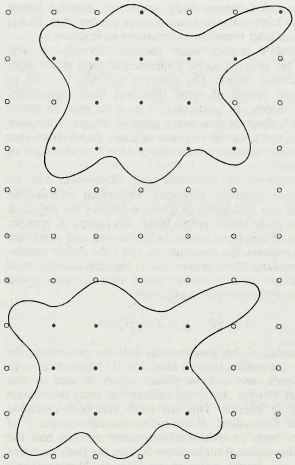


Figure 2: *An example illustrating Blichfeldt's theorem.* The lower region of area 14.2 (with 12 lattice points inside), has been translated to the upper region to capture 15 lattice points inside.

The three-dimensional version of Minkowski's theorem asserts that a bounded convex body, symmetric about a lattice point and with volume greater than 8, must contain at least three lattice points in its interior. The most important applications require an n-dimensional version of the theorem, in which the number 8 is replaced by 2^n. A proof, virtually identical to the one given above for the two-dimensional case, can be based on the n-dimensional version of Blichfeldt's theorem.

Minkowski extended his theorem to nonrectangular lattices in which the unit squares of the rectangular lattice are replaced by parallelograms of area $|\Delta|$, where Δ is a number called the determinant of the lattice. In the three-dimensional case the parallelograms become parallelepipeds of volume $|\Delta|$. For these more general lattices, the n-dimensional version of Minkowski's theorem states

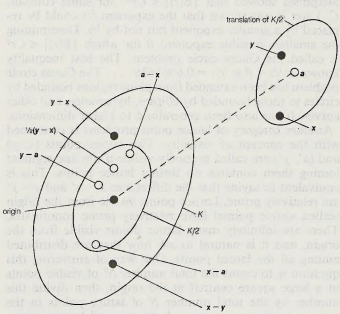


Figure 3: *Minkowski's convex body theorem deduced from Blichfeldt's theorem.* The lattice points x and y are in a translation of $K/2$. The lattice points $x - y$ and $y - x$ are in K.

that a bounded convex body, symmetric about a lattice point and with n-dimensional volume greater than $2^n|\Delta|$, contains at least three lattice points in its interior.

Minkowski's convex body theorem provides a new method for proving some fundamental facts about algebraic number theory. The theorem can also be used to derive new proofs of some theorems from elementary number theory. In particular, it can be used to prove Lagrange's theorem that every positive integer is the sum of four squares, or the theorem of Euler and Fermat that every prime number of the form $4k+1$ is the sum of two squares.

The theorems of Blichfeldt and Minkowski can be regarded as existence theorems. Prescribed information about the size and shape of a region implies the existence of one or more lattice points inside the region. A number of results of this type involve both the area and perimeter of plane regions. For example, in 1947 the Polish mathematician Maria Nosarzewska found that the number N of lattice points inside a convex plane region of area A and perimeter P satisfies the inequalities

$$A - \frac{1}{2}P < N \leq A + \frac{1}{2}P + 1.$$

In particular, if the area exceeds half the perimeter, the leftmost inequality implies that $N > 0$, hence $N \geq 1$. In other words, any convex plane region of area A and perimeter P with $A/P > 1/2$ contains at least one lattice point in its interior. This last result has been extended to higher dimensions. If the n-dimensional volume V of a convex body in n-dimensional space exceeds half the $(n-1)$-dimensional surface area S, then the body contains at least one lattice point in its interior. Nosarzewska's inequality improves a theorem of the Czech mathematician Vojtech Jarnik, which states that $|N - A| < P$ for any plane region, convex or not, with perimeter $P \geq 1$.

Many problems in number theory involve counting lattice points in special regions. For example, while investigating the number of representations of a number as a sum of two squares, Gauss proposed the problem of counting the number $N(r)$ of lattice points inside or on the boundary of a large circle of radius r with centre at the origin. Because every lattice point in the plane can be regarded as belonging to exactly one square of the unit square lattice (for example, that square having the lattice point as its lower left corner), counting lattice points in a region is essentially the same as counting squares in the region. Therefore, the number $N(r)$ should be approximately πr^2, the area of the region, with an error of the order of magnitude of the circumference, $2\pi r$. Gauss showed that $N(r) = \pi r^2 + E(r)$, where the error $E(r)$ satisfies $|E(r)| \leq 2\sqrt{2}\pi r$. Jarnik's inequality implies $|E(r)| \leq 2\pi r$ if $r \geq 1/(2\pi)$, a minor improvement. In 1904 the Polish mathematician Wacław Sierpiński showed that $|E(r)| \leq Cr^{2/3}$ for some constant C. Later it was shown that the exponent $2/3$ could be replaced by a smaller exponent but not by $1/2$. Determining the smallest possible exponent θ for which $|E(r)| \leq Cr^{\theta}$ is called the Gauss circle problem. The best inequality known is $1/2 < \theta \leq 7/11 = 0.6363636\ldots$. The Gauss circle problem has been extended from plane regions bounded by circles to those bounded by ellipses, hyperbolas, and other curves. It has also been generalized to higher dimensions.

Counting lattice points in circles

Another category of lattice point problems is concerned with the concept of visibility. Two lattice points (x, y) and (x', y') are called mutually visible if the line segment joining them contains no further lattice points. This is equivalent to saying that the differences $x - x'$ and $y - y'$ are relatively prime. Lattice points visible from the origin (called visible points) have relatively prime coordinates. There are infinitely many lattice points visible from the origin, and it is natural to ask how they are distributed among all the lattice points. One way of answering this question is to count the total number N' of visible points in a large square centred at the origin, then divide this number by the total number N of lattice points in the square. As the square expands to cover all lattice points, the limiting value of the ratio N'/N is $6/\pi^2$, called the density of the visible points. This result can be stated in the language of probability. If a lattice point is selected at random from the unit square lattice, the probability that it is visible from the origin is $6/\pi^2$. Or, if two integers are selected at random, the probability that they are relatively prime is $6/\pi^2$. The density of visible points in an n-dimensional lattice is $1/\zeta(n)$, where $\zeta(n)$ is the sum of the reciprocals of the nth powers of all the positive integers.

Visibility

UNSOLVED PROBLEMS

There are hundreds of unsolved problems in number theory, some of which have already been mentioned in this article. This concluding section lists examples of problems that are easy to state but which remain unsolved.

Prime numbers and divisibility. Are there infinitely many even perfect numbers?

Are there any odd perfect numbers?

Is there an even number greater than 2 that is not the sum of two primes?

Is there an even number greater than 2 that is not the difference of two primes?

Are there infinitely many twin primes—that is, primes p such that $p+2$ is also prime?

Are there infinitely many primes of the form $2p+1$, where p is prime?

Are there infinitely many Mersenne primes $2^p - 1$ with prime p?

Are there infinitely many Fermat primes—that is, primes of the form $2^{2^n} + 1$?

Are there infinitely many composite numbers $2^{2^n} + 1$?

Are there infinitely many primes of the form $x^2 + k$, where k is a given integer?

Are there infinitely many primes of the form $n! + 1$?

Are there infinitely many primes that are the sum of three cubes—that is, primes of the form $x^3 + y^3 + z^3$, where x, y, z are integers?

Is there always a prime between n^2 and $(n+1)^2$ for every positive integer n?

Are there infinitely many primes p satisfying the congruence $2^{p-1} \equiv 1 \pmod{p^2}$?

Are there infinitely many composite n satisfying the congruence $a^{n-1} \equiv 1 \pmod{n}$ for all a relatively prime to n?

Diophantine equations. For an arbitrary integer $n \geq 2$, are there positive integers x, y, z such that

$$\frac{4}{n} = \frac{1}{x} + \frac{1}{y} + \frac{1}{z}?$$

Are there integer solutions to the equation

$$x^3 + y^3 + z^3 = 3$$

other than the four solutions

$(x, y, z) = (1, 1, 1), (4, 4, -5), (4, -5, 4)$, and $(-5, 4, 4)$?

Are there integer solutions to the equation

$$x^3 + y^3 + z^3 = 30?$$

Are there positive integers x, y, z and $n > 2$ such that

$$x^n + y^n = z^n?$$

Does the equation $x^y = z^z$ have solutions in odd integers x, y, z greater than 1?

Arithmetical functions. Is there a composite n such that Euler's function $\varphi(n)$ divides $n - 1$?

Is there an integer n for which the equation $\varphi(x) = n$ is satisfied for exactly one integer x?

If $\pi(x)$ is the number of primes $\leq x$, is the inequality

$$\pi(x + y) \leq \pi(x) + \pi(y)$$

satisfied for all integers x and y greater than 1?

Irrational and transcendental numbers. Does the digit 1 occur infinitely often in the decimal expansion of $\sqrt{2}$?

Does the decimal expansion of π contain the sequence of nine digits 123456789 in succession?

Is the number $e + \pi$ transcendental?

The intellectual challenge of unsolved problems in number theory has kept the subject alive and active since the time of the ancient Greeks. Because new problems arise more rapidly than the old ones are solved, number theory will continue to thrive and to stimulate other branches of mathematics for centuries to come.

BIBLIOGRAPHY. Among the many books on number theory, the following are recommended for readers with a good background in high-school algebra: H. DAVENPORT, *The Higher Arithmetic: An Introduction to the Theory of Numbers*, 5th ed. (1982); IVAN NIVEN and HERBERT S. ZUCKERMAN, *An Introduction to the Theory of Numbers*, 4th ed. (1980); KENNETH H. ROSEN, *Elementary Number Theory and Its Applications*, 2nd ed. (1988); WACŁAW SIERPIŃSKI, *Elementary Theory of Numbers*, trans. from Polish (1964); and HAROLD M. STARK, *An Introduction to Number Theory* (1970). Introductory books at a slightly more advanced level include GEORGE E. ANDREWS, *Number Theory* (1971); and G.H. HARDY and E.M. WRIGHT, *An Introduction to the Theory of Numbers*, 5th ed. (1979, reprinted 1983). See also J.S. CHAHAL, *Topics in Number Theory* (1988). Collections of abstracts of articles published in *Mathematical Reviews* are found in WILLIAM J. LEVEQUE (ed.), *Reviews in Number Theory*, 6 vol. (1974), covering 1940–72; and RICHARD K. GUY (ed.), *Reviews in Number Theory 1973–83*, 6 vol. (1984).

Topics in analytic and algebraic number theory are presented in TOM M. APOSTOL, *Introduction to Analytic Number Theory* (1976); EMIL GROSSWALD, *Topics from the Theory of Numbers*, 2nd ed. (1984); and LOO KENG HUA (LO-KENG HUA), *Introduction to Number Theory* (1982; originally published in Chinese, 1957). More advanced treatments are Z.I. BOREVICH and I.R. SHAFAREVICH, *Number Theory* (1966; originally published in Russian, 1964); and KENNETH IRELAND and MICHAEL ROSEN, *A Classical Introduction to Modern Number Theory* (1982).

For the history of number theory, see LEONARD EUGENE DICKSON, *History of the Theory of Numbers*, 3 vol. (1919–23, reprinted 1966); OYSTEIN ORE, *Number Theory and Its History* (1948, reprinted 1988); and ANDRÉ WEIL, *Number Theory: An Approach Through History from Hammurapi to Legendre* (1984).

Specialized topics are covered by GEORGE E. ANDREWS, *The Theory of Partitions* (1976, reissued 1984); J.W.S. CASSELS, *An Introduction to Diophantine Approximation* (1957, reprinted 1972), and *An Introduction to the Geometry of Numbers* (1959, reprinted 1971); HAROLD M. EDWARDS, *Fermat's Last Theorem: A Genetic Introduction to Algebraic Number Theory* (1977); MARK KAC, *Statistical Independence in Probability, Analysis and Number Theory* (1959, reprinted 1964); L.J. MORDELL, *Diophantine Equations* (1969); IVAN NIVEN, *Irrational Numbers* (1956); C.D. OLDS, *Continued Fractions* (1963, reprinted 1975); PAULO RIBENBOIM, *13 Lectures on Fermat's Last Theorem* (1979), and *The Book of Prime Number Records* (1988); and M.R. SCHROEDER, *Number Theory in Science and Communication: With Applications in Cryptography, Physics, Digital Information, Computing, and Self-Similarity*, 2nd enl. ed. (1986).

Books devoted primarily to problems in number theory are RICHARD K. GUY, *Unsolved Problems in Number Theory* (1981); WŁADYSŁAW NARKIEWICZ, *Classical Problems in Number Theory* (1986); DANIEL SHANKS, *Solved and Unsolved Problems in Number Theory*, 3rd ed. (1985); WACŁAW SIERPIŃSKI, *A Selection of Problems in the Theory of Numbers*, trans. from Polish (1964), and *250 Problems in Elementary Number Theory* (1970); and I.M. VINOGRADOV, *Elements of Number Theory* (1954; originally published in Russian, 5th ed., 1949).

DERRICK HENRY LEHMER, *Guide to Tables in the Theory of Numbers* (1941), is a highly recommended, comprehensive, and descriptive account. Tables of primitive roots and indices are found in C.G.J. JACOBI (ed.), *Canon arithmeticus*, rev. and expanded ed. edited by HEINRICH BRANDT (1956); A.E. WESTERN and J.C.P. MILLER, *Tables of Indices and Primitive Roots* (1968); and HERBERT HAUPTMAN, EMANUEL VEGH, and JANET FISHER, *Tables of All Primitive Roots for Primes Less Than 5000* (1970).

(T.M.A./I.N.)

Numerical Analysis

Numerical analysis comprises three related activities: the study of a problem by the computation of numerical values that solve the mathematical equations (called mathematical models) describing the behaviour of some system, the development of methods (called numerical algorithms) for finding those values, and the analysis of the properties of those methods. The mathematical models are developed by professionals in various subject areas of application, that is, by scientists, engineers, or economists. A numerical analyst is a person who develops or analyzes methods for computing the numerical values.

Mathematical models are the starting point for the numerical analyst. Examples of models include an astronomer's mathematical description of the motion of a planet or a satellite (called its equations of motion), the equations used by a designer to describe the radioactivity in a nuclear reactor and the operation of its control system, the equations hypothesized by a research chemist to describe a chemical reaction, the equations used by government planners to express the relationships between economic and social indicators such as gross national product (GNP), prime rate, and unemployment rate, and the equations used in computer-aided tomography (CAT scans) to construct a view of the interior of a human body from an X-ray scan.

Among the uses of models are the following: prediction of future behaviour—for example, the determination of the future locations of the heavenly bodies (a matter of continuing interest in scientific calculation but of great concern in previous centuries because of its importance to navigation before the development of modern instruments); the selection of the best design—for example, the choice of the lowest cost nuclear reactor that meets the requirements of power output and safety; the comparison of computed values with experimental observations to test the validity of the theory represented by the model or to estimate the values of unknown constants that appear in the model—for example, the research chemist's effort to determine the rate at which materials react by comparing changes in chemical concentrations predicted from model equations with the values measured in the laboratory; and the determination of values that are not directly observable from other values that are—for example, in computer-aided tomography, the use of X-ray observations to provide the average density of a large set of narrow, pencil-shaped sections of the body, which then are used to construct a map of the density of the interior of the body.

Some problems are described by models so simple that they can be manipulated without numerical computation. For example, if a metal rod is subject to a stretching force (the stress, F) and it lengthens a certain amount (the strain, L), the stress-strain relationship might be modeled by the equation $L = \lambda F$, in which λ is a numerical constant. This says that the stretching is linearly proportional to the applied force, so that if the force is doubled, the stretching doubles. A person can understand this relationship and use it to solve problems such as "What force is needed to stretch the rod L millimetres?" in the symbolic or explicit form $F = L/\lambda$, so that it is unnecessary to compute actual values until they are needed in a design task. However, many models are so complex that it is not possible to solve problems explicitly or to understand the relationships expressed by the model. In that case, the model user turns to numerical methods, either to compute specific solutions or to study the behaviour of the model as it is changed.

The goal of numerical analysis is the efficient computation of accurate approximations of the values that satisfy mathematical equations. Two major problems confront the numerical analyst: the round-off errors that unavoidably arise during computation, and the representation of problems involving infinite amounts of information with the finite number of values that a person (or computer) can handle.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 10/23, and the *Index*.

This article is divided into the following sections:

Errors 38
  Finite-precision errors: round-off 38
  Finite-dimension errors: truncation 39
Approximation of functions 41
  Polynomial interpolation 41
  Difference calculus 42
  Computation of elementary functions 43
  Spline approximation 43
  Least squares approximation 43
  Minimax approximation 43
Solution of equations 44
  Linear equations and direct methods 44
  Nonlinear equations and iterative methods 45
  Differential equations 47
Applications and implementation 48
  Numerical optimization 48
  Inverse problems 48
  Large-scale scientific computation 49
  Mathematical software 49
Bibliography 49

ERRORS

Most problems involve infinite sets of values, each of which can potentially require an infinite number of digits for exact representation. Digital computation, either human, mechanical, or electronic, is, by its very nature, finite: it involves a finite set of numerical values, each of which is represented in a finite set of digits. These two approximations of infinite quantities by finite ones lead to the two types of error in numerical computation, round-off and truncation error. A numerical analyst is interested in determining the size of possible errors in a calculation.

Bounds and estimates of errors

Analysts try to find bounds on errors or estimates of errors. A bound is a value that the analyst can guarantee the error will not exceed. An estimate is a value that the analyst believes is an approximation of the error (error estimates typically vary from the true error by a factor of 2). The word *error* is unfortunate because it suggests a mistake or sin. Round-off and truncation errors are unavoidable: the job of the numerical analyst is to find out how to compute accurately and efficiently in their presence.

Finite-precision errors: round-off. The precision of a value is indicated by the number of digits used in its representation. It is typically between seven and 15 decimal digits, depending on the type of computer and calculation. Finite precision means that most values cannot be represented exactly, but that there is some round-off error in the computed approximation. For example, if $^4/_3$ is represented by five significant decimal digits as 1.3333 (the most accurate possible using five digits), there is a round-off error of $0.0000333\ldots$ or $^1/_3 \times 10^{-4}$, as illustrated below:

	decimal representation
correct value	1.3333333...
5-digit representation	1.3333
round-off error	0.0000333...

All calculations are subject to the effects of round-off errors. Round-off errors, which are proportional to the size of the value containing the error when a fixed number of significant digits is used, can be as large as 5 in the first neglected place or one-half in the last retained place. When a value is added to a much smaller value, the round-off can be large relative to the smaller value. For example, if 23,456 is added to 10.518 in five digits, the result is 23,467, with a round-off error of −0.482. If 23,456 is subtracted from the answer, the result is 11. Thus, in five-digit arithmetic, (23,456 + 10.518) − 23,456 = 11.0, which has an error in the third place. After the addition, the error was less than one-half in the last place, but the subtraction removed the three leading digits, 234, moving this error to the third place. This phenomenon is called cancellation. It does not cause errors but makes the size of errors already introduced larger relative to the computed result. Thus, although round-off errors are small, their effect in the final answer can be large, so that one of the tasks of the numerical analyst is to devise or modify computational schemes to minimize the effect of these errors.

The effect of a change in the computational scheme is illustrated by a simple example related to an exterior ballistics problem. (Exterior ballistics deals with the trajectory of a projectile after it leaves the barrel of a gun. Much early numerical table building and computing was devoted to this problem, which is related to the orbit problem.) When a projectile is fired from ground level (that is, from an initial height, $h = 0$) and with an initial vertical velocity of v_0, its height, h, after t seconds is $h = tv_0 - t^2g/2$, where g is the acceleration due to the force of gravity. The projectile will hit a target at height h_0 at a time that satisfies the quadratic equation

$$h_0 = tv_0 - t^2\frac{g}{2}. \tag{1}$$

This equation can be rewritten in the form of the standard quadratic equation $ax^2 + bx + c = 0$ whose roots, r, are given by the formula

$$r = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Upon replacing a, b, and c by $g/2$, $-v_0$, and h_0, the roots of (1) are found to be

$$r = \frac{v_0 \pm \sqrt{v_0^2 - 2gh_0}}{g}.$$

For the case $v_0 = 500$, $h_0 = 75$, and $g = 9.82$, the two roots can be computed from

$$r = \frac{500 \pm \sqrt{250{,}000 - 2 \times 9.82 \times 75}}{9.82}.$$

The value of the quantity under the square root symbol is 248,527. If three-digit arithmetic is used, it becomes 249,000 (trailing zeros do not count as significant digits). The square root of 249,000 is 499 correct to three digits. Hence the two roots are computed as (500 + 499)/9.82 = 102 and (500 − 499)/9.82 = 0.102 to three digits. The correct solutions to three digits are 102 and 0.150. The 50 percent error in the smaller root occurs because the root is computed by the subtraction of the two nearly equal values 500 and 499. This illustrates the phenomenon of cancellation. Because of three-digit accuracy, each of these two values can be in error by as much as 0.5. In fact, to six digits the second value is 498.525 so that the approximation 499 is in error by 0.475. After subtraction, the computed result of 500 − 499 is 1.0, instead of 1.475, and is in error in the second place. This large error can be avoided by computing the smaller root with a different version of the formula. It is easy to show (by dividing (1) by t^2 and solving for $1/t$) that the second root is given by

$$r = \frac{2h_0}{v_0 + \sqrt{v_0^2 - 2gh_0}},$$

from which the result, 0.150, can be computed using three-digit arithmetic. In the article "What Is a Satisfactory Quadratic Equation Solver?" George E. Forsythe discusses further the difficulties that can arise even in this simple problem.

The quadratic equation above is a problem whose solutions can be computed accurately if the calculation is organized correctly. Numerical analysis is used to investigate the errors that can arise in such computations. In some problems it is very difficult to compute accurate answers because a slight change in the problem can cause a large change in the answer. Such problems are said to be ill-conditioned. In some types of problems, an arbitrarily small change in the problem can lead to an arbitrarily large change in the answer. Such problems are said to be ill-posed. A properly posed problem is one in which bounded changes to the problem lead to bounded changes in the answer. A well-conditioned problem is one whose answers change only in a small way if the initial problem is changed in a small way. The equation

Ill-conditioned and ill-posed problems

$$(x - 1.1)^4 = 0$$

has the answer $x = 1.1$ by inspection. When the expression on the left-hand side is multiplied out, the equation

$$x^4 - 4.4x^3 + 7.26x^2 - 5.324x + 1.4641 = 0 \tag{2}$$

is obtained. In this form it is ill-conditioned. If five-digit arithmetic is used, the coefficients in (2) can be represented exactly. On the other hand, if four-digit arithmetic is used, the last coefficient would be represented as 1.464, to get the equation

$$x^4 - 4.4x^3 + 7.26x^2 - 5.324x + 1.464 = 0, \tag{3}$$

which is the same as

$$x^4 - 4.4x^3 + 7.26x^2 - 5.324x + 1.4641 = 0.0001 = 10^{-4}. \tag{4}$$

In this particular case, the roots of the changed equation can be computed exactly because it has the form $(x - 1.1)^4 = 10^{-4}$, of which one root is given by $(x - 1.1) = 10^{-1}$ or $x = 1.2$. Thus, a change in the fifth decimal place, or 0.01 percent, in the original problem (2) to get (3) has caused a change in the second place, or 10 percent, in the answer. Because this problem is ill-conditioned, no numerical method can avoid the fact that a slight change to the problem can cause a large change in the answer. Because round-off errors in the early stages of the calculation are indistinguishable from changes in the original problem, such problems are susceptible to large errors in their numerical answers, and a simple error analysis shows these errors.

A direct analysis of the effect of the errors on the answers is called a forward error analysis because it analyzes the effects of the errors as they propagate forward through the steps of the calculation to the final result. A forward analysis leads to the statement "Round-off errors have caused this much error in the answer." As illustrated above, a small round-off can cause a large error in the answer of an ill-conditioned problem. Another useful measure of the computational error is the amount by which the computed answer fails to satisfy the problem. This is called the residual. For example, if the computed answer of 1.2 is substituted into equation (2), the left-hand side is $(1.2 - 1.1)^4 = 10^{-4}$. Although the error in the answer is 0.1, the residual is 10^{-4}. In this example and in many others, the size of the residual can be estimated by a backward error analysis. This leads to the statement "Round-off errors can change the answers so that they are the exact answers of a problem that is different from the original problem by this much." A backward error analysis applied to a good method for the problem discussed above might show that the method solved a problem whose coefficients were within 0.01 percent of those in (2), so that (3) could have been the problem solved. Although the answers to the changed problem may be very different from those of the original, they are as good as can be expected from an ill-conditioned problem. A good method is one that has small backward errors or residuals: for a well-conditioned problem, such a method always gives accurate answers because, by definition, in a well-conditioned problem the answer changes by only a small amount when the problem changes by a small amount.

Forward and backward error analysis

Finite-dimension errors: truncation. The types of problems discussed in the previous section are finite dimensional because only a finite number of values are involved:

the values of the unknown, x, and the coefficients. In many problems an infinite amount of information is inherently present, and the problem cannot be solved exactly by numerical methods that deal with only a finite number of values. This circumstance introduces a fundamental concept, that of a finite approximation to an infinite-dimensional object, which is illustrated by the approximation of a smooth graph of a function by a series of straight lines (a piecewise-linear approximation).

A function, $f(x)$, is a prescription of a value corresponding to each value of its argument, x. For example, $f(x) = 2 - x + x^2$ is a function whose value is 2 when $x = 0$, 2 when $x = 1$, 4 when $x = 2$, etc. A description of a function by a mathematical formula defines the value of the function for all values of its argument. The mathematical definition of a function is that it is a (possibly infinite) set of pairs of values, (x_i, y_i). A function of a single variable x can be represented graphically by placing a point at $[x, f(x)]$ for each value of x. If this is done for a few values of x, say five values, the set of points shown in Figure 1 is obtained. This graph can be described with just 10 numbers, $x_1, y_1, x_2, y_2, \ldots, y_5$, where $y_i = f(x_i)$. The members of each pair (x_i, y_i) are the coordinates of one of the points. These five points, however, do not prescribe the value of the function for other than the five given values of x. More pairs of values (x_i, y_i) could be added, but they cannot specify the value of the function for every value of x unless there is an infinite number of pairs of values. Hence, this graph needs infinitely much information if it is to describe the function completely. Thus, a general function is an infinite-dimensional object because it requires infinitely much information to specify it. If the graph is "smooth," however, it is reasonable to approximate it by joining a small set of points with a continuous line. If consecutive points are joined by a sequence of straight lines, as shown in Figure 2, a piecewise-linear approximation is obtained. Such an approximation requires the values of only a finite number of values, the values of the end points of each of the straight-line segments shown in Figure 2. The piecewise-linear approximation is a finite-dimensional approximation to an infinite-dimensional object. This is critically important in numerical computation by hand or computer because computers and people can deal only with finite amounts of information. More closely spaced points can be used in a piecewise-linear approximation to get a more accurate approximation, as also shown in Figure 2, but the additional points mean that more information is used to approximate the function.

The error caused by replacing an infinite-dimensional object by a finite-dimensional approximation is called truncation error. A large part of numerical analysis is

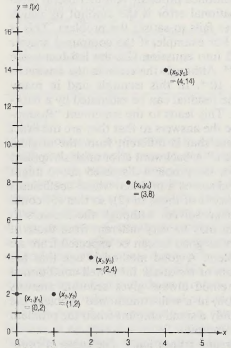


Figure 1: Graph of values of the function $y = f(x) = 2 - x + x^2$ for $x = 0, 1, 2, 3$, and 4.

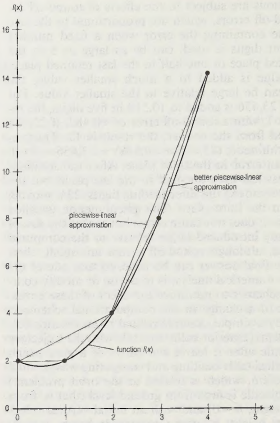


Figure 2: Piecewise-linear approximations to a continuous function (see text).

concerned with the accuracy of finite-dimensional approximations, examining questions such as "What type of approximation can be used?", "How much information is needed for a given accuracy?", and "Will the approximation become more accurate as more information is used?". As more information is used in an approximation, more computation is required to evaluate it, so that it is natural to expect more accuracy. A method that can achieve arbitrarily much accuracy by using sufficient information is said to be convergent. Most methods have one or a few parameters that determine the amount of information used in the approximation. For example, if, in the piecewise-linear approximation discussed above, the points x_i are equally spaced, that is, if $x_{i+1} - x_i = h$ for all i, where h is the mesh spacing, then the number of points used over a fixed interval, say (0, 1), is determined by the parameter h. As h is reduced, the amount of information used to approximate the function increases, and it is visually obvious from Figure 2 that the approximation becomes more accurate. How much more accurate it becomes is typical of the questions that numerical analysts investigate. In this case it is relatively easy to show with calculus that if h is halved, the error decreases by a factor of approximately 4, provided that the function $f(x)$ is smooth enough, that is, if it is continuous and has no sharp turns in it, or, mathematically, if it has a second derivative. The relationship between the error and the mesh size for the piecewise-linear approximation is that the error is no larger than h^2 times a constant times the second derivative of the function. This method is said to have second-order convergence. (If a method depends on a parameter h and its error is proportional to h^p, it is said to have pth-order convergence.) Numerical analysts look for methods that have high orders of convergence because, as more information is used, their accuracy increases more rapidly than lower-order methods. Straightforward methods for simple problems, such as the approximation of functions, are usually convergent, but much care must be exercised in the choice of methods for complex problems to get convergence. A significant part of the theoretical work in numerical analysis is concerned with proofs of the convergence of methods.

Convergent methods

In practical calculation, round-off errors are also present. As the amount of information is increased to decrease truncation error, the size of each round-off error, which is determined by the number of digits used in the calculation, remains constant. Hence, no amount of additional infor-

mation can reduce the error below some minimum unless the precision of the calculation is increased. Furthermore, in methods for many problems the number of arithmetic operations in the calculation increases as the amount of information increases, so that the total error due to round-off increases because it is the combined effect of an increasing number of round-off errors. In these calculations, the error in the answer, which is the sum of the effects of truncation errors and round-off errors, initially decreases as the amount of information used increases and the truncation error decreases, but then increases as the round-off errors become dominant. This is illustrated in Figure 3.

APPROXIMATION OF FUNCTIONS

An approximation to a function may be needed for several reasons. In some cases, values of the function are known at only a finite number of points, where they have been previously obtained by measurement. In that case, a method of estimating the values of the function between those points is needed. The process of obtaining these values is called interpolation. In other cases, it is computationally expensive to calculate the function every time it is needed, but less expensive to compute an approximation, which requires fewer computational operations. In still other cases, it is not possible to compute the function exactly with the standard arithmetic operations of addition, multiplication, etc., even with exact arithmetic, so that an approximate formula must be used.

Many approximations to functions consist of polynomials. A polynomial is a function of the form

$$P(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n. \tag{5}$$

This polynomial has degree n. It can be evaluated using n multiplications and additions when written in the form

$$(\cdots((a_n \times x + a_{n-1}) \times x + a_{n-2}) \times x + \cdots + a_1) \times x + a_0.$$

A polynomial of degree n contains $n+1$ coefficients a_i, $i = 0, 1, \ldots, n$, so that it has $n+1$ degrees of freedom, or has dimension $n+1$. It is formed as a linear sum of multiples of the $n+1$ basis functions x^i, $i = 0, 1, \ldots, n$. These functions are independent, which means that no linear combination of them will be identically zero for all x. When an infinite-dimensional object has to be approximated by a finite-dimensional one, any set of independent functions can be used. If $\varphi_i(x)$, $i = 0, 1, \ldots, n$, are a set of $n+1$ independent functions, they form a basis, and the approximation

$$f(x) \approx a_0\varphi_0(x) + a_1\varphi_1(x) + \cdots + a_n\varphi_n(x)$$

can be used. If the $\{\varphi_i(x)\}$, $i = 0, 1, 2, \ldots$, are an infinite set of independent basis functions that satisfy certain technical restrictions, they can be used to get arbitrarily accurate approximations by including enough of them. If the basis function $\varphi(x)$ is a polynomial of degree i, these technical restrictions are satisfied. The approximations generally have better properties if the basis functions are as different from one another as possible. In this respect, the functions x^i are not very good because, for example, the functions x^3 and x^4 are not greatly different in overall shape, as can be seen by drawing them. An important class of basis functions are those that are mutually orthogonal, which is a mathematical prescription for requiring them to be as different as possible. (This use of orthogonal is related to its use in specifying directions that are at right angles, and thus very different, such as north and east.)

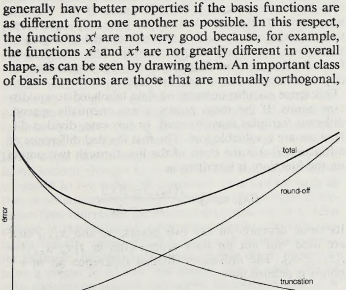


Figure 3: The effect of the amount of information on error in numeric computations.

Polynomial interpolation. Interpolation is used to compute the approximate value of a function, given its values at a set of points. To approximate the function at a point whose value is not given, the value of a polynomial passing through a few neighbouring points is computed, as shown in Figure 4. Usually, the points with known values closest to the point of unknown value are chosen because the accuracy of interpolation is proportional to the distance of the unknown point from each of the known points. An important fact about polynomials is that a unique polynomial of degree n can be constructed to pass through any set of $n+1$ points; that is, if (x_i, y_i), $i = 0, 1, \ldots, n$, are a set of points with all x_i different, there is a unique set of coefficients a_i such that $P(x)$ given by equation (5) satisfies $P(x_i) = y_i$ for $0 \le i \le n$. A different set of points, and hence a different polynomial, is used to interpolate for different regions of x values. For example, if a third-degree polynomial (a cubic) were to be used, four points would be needed. If the x_i are ordered, that is, if $x_0 < x_1 < x_2 < \ldots$, and x lies between, say, x_7 and x_8, the points x_6, x_7, x_8, and x_9 would be used, whereas if x lies between x_{13} and x_{16} the points x_{14}, x_{15}, x_{16}, and x_{17} would be used to construct the interpolating polynomial. The coefficients of that polynomial could be calculated for each interval, and then the value of $P(x)$ could be calculated from the polynomial. However, computational schemes have been devised to compute the interpolated value directly without finding the coefficients of the underlying polynomial. The user of such schemes is interested in rapid calculation with a minimum susceptibility to round-off error. The Lagrange formula is an important form of polynomial interpolation. To interpolate an nth-degree polynomial that matches the $n+1$ points (x_i, y_i), $i = 0, 1, \ldots, n$, this formula incorporates the set of $n+1$ polynomial basis functions, $\varphi_i(x)$, each of degree n, which are such that $\varphi_i(x)$ is zero at all of the $n+1$ points except at x_i, where its value is 1. The interpolating polynomial can be written as

Lagrange formula

$$P(x) = y_0\varphi_0(x) + y_1\varphi_1(x) + \cdots + y_n\varphi_n(x).$$

This clearly satisfies the interpolation conditions: each of the basis functions is a polynomial of degree n, so their sum is also an nth-degree polynomial; when $x = x_i$, all of the terms in the sum vanish except for $y_i\varphi_i(x_i)$, which is just y_i. Definition of the $\varphi_i(x)$ is straightforward. On the basis of the function

$$\Pi(x) = (x - x_0)(x - x_1) \cdots (x - x_n),$$

which is an $(n+1)$st-degree polynomial that is zero at $x = x_i$ for $0 \le i \le n$, a set of functions

$$\pi_i(x) = \Pi(x)/(x - x_i)$$

may be defined. These are nth-degree polynomials that are zero except at x_i. Division of each of them by $\pi_i(x_i)$ gives the functions

$$\varphi_i(x) = \pi_i(x)/\pi_i(x_i),$$

each of which has the desired property that its value at x_i is 1. Although these formulas are simple, they are not the basis for computationally efficient methods. The importance of the Lagrange formula arises mainly from its use in the mathematical development of methods and proofs of their properties. Although there are many different polynomial interpolation formulas, all nth-degree polynomial interpolation formulas that are correct at $n+1$ points are equivalent except in the amount of computation and the effects of round-off.

The approximation of a function by a polynomial forms the basis of many other numerical calculations besides interpolation. If the integral or derivative of a function is needed and it is not possible to compute it directly, it can be approximated by computing the integral or derivative of a polynomial approximant since there are straightforward rules for evaluating the integral or derivative of a polynomial.

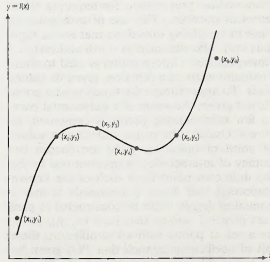


Figure 4: *Polynomial interpolation.*
The six points x_i, y_i, etc., represent values of an unknown function. A third-degree polynomial has been constructed so that four of its values match four of the values of the unknown function. Other third-degree polynomials could be made to match other sets of four values of the unknown function, or a polynomial of higher degree could be found to match all six.

Difference calculus. Differences are the basis for the development of many methods of polynomial approximation. If the values of a function $f(x)$ are given at a set of equally spaced argument values x_0, x_1, $x_2 \ldots$, where $x_n = x_0 + nh$, the forward difference of $f(x)$ at x_n, written as $\Delta f(x_n)$, is defined as $f(x_{n+1}) - f(x_n)$. Piecewise-linear interpolation between x_n and x_{n+1} is given by the formula

$$f(x) \approx f(x_n) + \frac{(x - x_n)}{(x_{n+1} - x_n)} \Delta f(x_n).$$

Since the spacing between the x_i's is h, this can be written as

$$f(x) \approx f(x_n) + s\Delta f(x_n),$$

where $s = (x - x_n)/h$. Higher degree polynomials can be used by using higher order differences. The second difference, written $\Delta^2 f(x_n)$, is defined as $\Delta f(x_{n+1}) - \Delta f(x_n)$, which can be rewritten as $f(x_{n+2}) - 2f(x_{n+1}) + f(x_n)$. In general, the mth-order difference is given by

$$\Delta^m f(x_n) = \Delta^{m-1} f(x_{n+1}) - \Delta^{m-1} f(x_n).$$

The polynomial of degree m passing through the points x_n, $x_{n+1}, \ldots, x_{n+m}$ can be written as

$$P(x) = f(x_n) + \frac{(x - x_n)}{(x_{n+1} - x_n)} \Delta f(x_n) + \frac{(x - x_{n+1})(x - x_n)}{(x_{n+2} - x_{n+1})(x_{n+2} - x_n)} \Delta^2 f(x_n)$$
$$+ \cdots + \frac{(x - x_{n+m-1}) \cdots (x - x_n)}{(x_{n+m} - x_n) \cdots (x_{n+1} - x_n)} \Delta^m f(x_n).$$

This can be rewritten as

$$f(x) \approx f(x_n) + s\Delta f(x_n) + \frac{s(s-1)}{2} \Delta^2 f(x_n)$$
$$+ \cdots + \frac{s(s-1) \cdots (s-m+1)}{m!} \Delta^m f(x_n).$$

If the function is smooth, its differences get smaller as their order increases until round-off errors cause them to increase again. The point at which they become smallest gives a direct indication of how many differences can be used to achieve the maximum accuracy possible. A list of the differences of various orders can be easily computed and used for interpolation. An important property of differences is that if the function is a polynomial of degree m in x, the differences of order higher than m are zero and $\Delta^m f(x_n)$ is independent of n, that is, is a constant.

Difference calculus is a formal manipulation scheme for developing useful approximations. It uses the operators Δ (already introduced as the forward difference operator); E, which is defined as the step operator so that $Ef(x_n) = f(x_{n+1})$; and others. Because of the relationship

$$f(x_{n+1}) = f(x_n) + f(x_{n+1}) - f(x_n) = f(x_n) + \Delta f(x_n),$$

the formal relationship

$$E = 1 + \Delta$$

can be written. This can be squared to get

$$E^2 = (1 + \Delta)^2 = 1 + 2\Delta + \Delta^2.$$

This is the formal statement that

$$f(x_{n+2}) = f(x_n) + 2\Delta f(x_n) + \Delta^2 f(x_n),$$

whose truth can be verified by simple algebra. Since E^m is the equivalent of moving forward m values in the table, it seems reasonable that $E^{0.5}$ should be the equivalent of moving forward half a step in the table, that is, to interpolating for the value of $f(x)$ at $x_{n+1/2}$. Similarly, E^s should represent interpolation s steps forward for any s, whether integer or not. Formally, E^s can be rewritten as

$$E^s = (1 + \Delta)^s$$

and expanded according to the binomial expansion to get

$$E^s = 1 + s\Delta + \frac{s(s-1)}{2} \Delta^2 + \frac{s(s-1)(s-2)}{3!} \Delta^3 + \cdots.$$

When this is applied to $f(x_n)$ and the infinite series is terminated after $m + 1$ terms, the previous interpolation formula is obtained. Thus, difference calculus has provided a formal verification of the formula stated earlier.

To illustrate how differences can be used for many other approximations, it is convenient to define the operator D as differentiation. Formally,

$$Df(x) = \frac{d}{dx} f(x).$$

The Taylor's series approximation to a function states that

$$f(x + a) = f(x) + a \frac{d}{dx} f(x) + \frac{a^2}{2} \frac{d^2}{dx^2} f(x) + \frac{a^3}{3!} \frac{d^3}{dx^3} f(x) + \cdots$$

This can be written in operator form as

$$E^{a/h} = 1 + aD + \frac{(aD)^2}{2} + \frac{(aD)^3}{3!} + \cdots.$$

The series on the right-hand side is the power-series expansion of the exponential function and is equal to e^{aD}. On the replacement $a = sh$ this relation becomes

$$E^s = e^{shD} = (e^{hD})^s.$$

This leads to $E = e^{hD}$, from which the relationship

$$hD = \log(E) = \log(1 + \Delta)$$

is obtained. Expansion of the logarithm then yields

$$hD = \Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} - \cdots,$$

which gives a numerical approximation for the derivative of a function in terms of its differences. If this relation is terminated after $m + 1$ terms, it is exactly equivalent to computing the derivative of a polynomial of degree m through $m + 1$ values of the function. This fact follows from the fact that the differences of order higher than m of an mth-degree polynomial are zero.

Difference calculus operates on data tabulated at equidistant points. If the mesh points x_i are unequally spaced, different formulas must be used. In this case, divided differences are a valuable tool. The first divided difference of a function $f(x)$ is the slope of the line through two points on the function. It is written as

$$f[x_n, x_{n+1}] = \frac{f(x_{n+1}) - f(x_n)}{x_{n+1} - x_n}.$$

Its value depends on the two points, x_n and x_{n+1}, that are used, but not on their order. That is, $f[x_n, x_{n+1}] = f[x_{n+1}, x_n]$. The mth-order divided difference on $m + 1$ points is defined as

$$f[x_n, x_{n+1}, \ldots, x_{n+m}]$$
$$= \frac{f[x_{n+1}, x_{n+2}, \ldots, x_{n+m}] - f[x_n, x_{n+1}, \ldots, x_{n+m-1}]}{[x_{n+m} - x_n]}.$$

If the points x_i are equally spaced with the distance between them equal to 1, then the mth-order divided difference is identical to the mth-order difference divided by $m!$. Polynomial interpolation by an mth-degree polynomial can be written as

$$f(x) \approx f(x_n) + (x - x_n)f[x_n, x_{n+1}] + \cdots$$
$$+ (x - x_n)(x - x_{n+1}) \cdots (x - x_{n+m-1})f[x_n, x_{n+1}, \cdots, x_{n+m}].$$

Computation of elementary functions. Elementary functions are functions that occur very frequently in scientific and engineering calculations. They are usually the solution of simple relations that cannot be expressed directly in terms of other, simpler operations or functions. Examples include sine, cosine, and exponential. The solutions to many problems can be written explicitly in terms of elementary functions. Many elementary functions can be expressed as power series, which are like polynomials but which have infinitely many terms. For example, the exponential function has the power series expansion

Power series approximations

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \cdots . \qquad (6)$$

This summation has the property that as more and more terms are included it converges; that is, its value comes closer and closer to the true value. If (6) is evaluated to seven-digit accuracy with $x = 0.1$ for a number of terms, the following results are obtained:

number of terms (n)	value of nth term	sum of n terms
1	1.000000	1.000000
2	0.100000	1.100000
3	0.005000	1.105000
4	0.000167	1.105167
5	0.000004	1.105171
6	0.000000	1.105171

Further terms would not change the seven-digit sum. The series provides a way of computing an approximation of e^x to any desired accuracy by adding enough terms. Many elementary functions are calculated in this way (although the minimax method described later can be used to reduce the amount of arithmetic and hence the time taken).

Spline approximation. In many applications the shape of an object is defined by the positions of a number of points on its surface, for example, the hull of a ship or the body of an automobile. The builder, to construct a smooth-surfaced object from this information, needs a mechanism for determining the position of any point on its surface. A simpler case of this problem is the construction of a smooth curve through a set of points—the problem solved by a drafter asked to draw a graph. The numerical problem is to approximate a function smoothly from the knowledge of a few of its points. Polynomial interpolation provides an interpolant, which is the approximating function, but it is not usually smooth because different polynomials are used between different pairs of points. Thus, a piecewise-linear interpolant has angles between the straight lines in adjacent regions, as shown in Figure 2.

A spline approximant is a piecewise polynomial interpolant with the property that it matches the function values at the given points, called the knots, and that it is smooth. The mathematical definition of a smooth function is that it has a certain number of continuous derivatives. The first derivative is the slope of the curve. If the slope is continuous, the curve has no angles, that is, no sudden changes in direction. The second derivative is related to the radius of curvature. In a surface such as that of an automobile body, the radius of curvature determines how reflections are magnified or reduced. A sudden change in the radius of curvature results in a surface that is not visually pleasing if it is supposed to be uniform—for example, the hood of an automobile might appear to have a crease in it. For this reason, curves and surfaces with continuous second derivatives are often desirable. A common spline approximation is a piecewise cubic with continuous second derivatives.

Smooth functions

Least squares approximation. Interpolation approximates functions whose form is known approximately with a polynomial (or other easily computable interpolant) that matches the function values assumed to be known exactly. In some situations the form of the function is known exactly (for example, it is known that it is a linear function), but the only other information known is the approximate value at a few points. This happens when, for example, the values are obtained from experimental measurements. In this situation it is desirable to compute a function of the known form that is the best fit to the given values. To make this definition precise, the meaning of "best" must be specified. The least-squares approximation is the one that minimizes the sum of the squares of the differences between the given values and the values of the function. For example, if it is known that the function is linear, that is, that $f(x) = ax + b$, and the three measurements x_1, y_1; x_2, y_2; and x_3, y_3 are given, where $y_i \approx f(x_i)$, the three differences are $y_i - (ax_i + b)$, $i = 1, 2, 3$, and the least squares method chooses a and b such that

$$[y_1 - (ax_1 + b)]^2 + [y_2 - (ax_2 + b)]^2 + [y_3 - (ax_3 + b)]^2$$

is as small as possible. If at least two of the x_i are distinct, there is a unique set a and b that achieves a minimum of this expression. Formulas for the values of a and b can be derived by the use of calculus. The general linear least-squares problem, which is a least-squares problem in which the unknown parameters (a and b in the example) appear linearly in the original errors ($y_i - ax_i - b$ in the example), has the same properties: if there are n unknown parameters and there are at least n distinct points x_i, there is a unique minimum.

The calculation of the solution to a least-squares problem is a good example of the application of numerical analysis. Straightforward application of calculus to the problem of the least-squares nth-degree polynomial approximant leads to a system of linear equations whose solution is discussed below. For large n the system is very poorly conditioned, so that a poor numerical answer is obtained. This occurs when the polynomials x^i are used as a basis; the use of orthogonal polynomials as a basis leads to much better accuracy.

Minimax approximation. Minimax approximation is another example of choosing the best approximation from a set of possible approximations. A measure of the error in an approximation to a function is first defined as the size of the largest difference between the approximation and the function at any of the points of interest. The approximation that has the smallest error measure is the minimax approximant. For example, the minimax linear approximation to the three points (x_i, y_i), $i = 1, 2$, and 3, is the line shown in Figure 5. It is equidistant from each of the points, as shown. Minimax approximations are related to least-squares approximations in that both are best approximations; in both approximations, "best" is defined as the minimum of some measure of error, but the measure of the error is different. In least squares, it is the sum of the squares of the errors; in the minimax case, it is the size of the largest error. (Both methods can be viewed as special cases of general minimum-error approximations.)

Measures of error

Minimax approximations are most frequently used to approximate one known function by another function that is less difficult to compute. For example, the sine function might be approximated by a polynomial in a computer program that computes a numerical value of sin(x) by

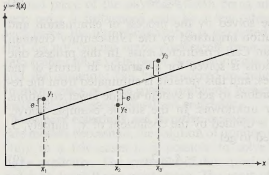


Figure 5: *Linear minimax approximation.* The line has been drawn so that its vertical distance, e, is the same from each of the three points.

evaluation of the polynomial. In this approach, a polynomial that is the best approximation in the minimax sense is used. The minimax polynomial may require fewer terms to get a given accuracy than would be needed if the first few terms of the power series were used. Although it takes a large amount of computation to determine the minimax polynomial for a given accuracy, that computation can be carried out once, and then the resulting polynomial can be evaluated whenever an approximation to the function is needed.

The nth-degree minimax polynomial approximant to a function has the important property that the difference between the polynomial and the function (the error) reaches its maximum value at least $n+2$ times, and the sign of the error has alternate signs at each extremum. Figure 5 illustrates this in the case $n=1$. The Chebyshev orthogonal polynomials are a set of polynomials, $T_0(x)$, $T_1(x)$, $T_2(x), \ldots$, where $T_n(x)$ has degree n and is such that it has $n+1$ equal and opposite extrema. The values of the extrema alternate between $+1$ and -1. These polynomials are effective as basis functions for the representation of minimax approximations.

SOLUTION OF EQUATIONS

An equation is a relation between two expressions, for example, $3z+4=(y-2)^2$. The expressions may contain known values, such as the numbers in the example, and unknowns, such as z and y, for which values must be computed. An equation says that the values of the unknowns must be such that the values of the expressions on either side of the equal sign are the same. An unknown, say x, appears linearly in an expression if it appears only in the form ax, where a has a known value independent of x, after all terms in the expression have been multiplied out. Thus, z appears linearly in the above example, but y does not—because the right-hand side is y^2-4y+4 and the term y^2 is nonlinear. A linear equation is one in which all unknowns appear linearly and each is multiplied only by a constant. Most phenomena that physicists and engineers model are nonlinear, so that the effective solution of nonlinear equations is important. Many of the models can be approximated by linear equations, however, so the efficient solution of linear equations is equally important. Indeed, the solution of nonlinear equations often involves the solution of linear equations within the process.

The first section below discusses the solution of linear equations. In addition to their importance in many problems, they also illustrate the problems of round-off errors, backward error analysis, and the difficulty of ill-conditioned problems. Linear equations have the property that they can be solved exactly using the usual arithmetic operations of addition, subtraction, etc., if no round-off is involved. Methods that solve a problem exactly in the absence of round-off are called direct methods because they lead directly to the solution. Nonlinear equations cannot be solved exactly; rather, their solution must be approximated. So-called iterative methods generate a sequence of closer and closer approximations called iterates.

Linear equations and direct methods. The standard linear-equation problem is a set of N linear equations in N unknown variables. An example of a simple system of two equations is

$$3x_1+6x_2=12 \tag{7}$$

$$5x_1-3x_2=7. \tag{8}$$

Gauss's procedure

These can be solved by the process of elimination and back substitution originated by the 19th-century German mathematician Carl Friedrich Gauss. In this process one of the equations is solved for a variable in terms of the other variables, and this variable is eliminated from the remaining equations to get a system of one fewer equations in one fewer unknowns. In the simple example above, equation (7) is divided by the coefficient of x_1, namely, 3, and rearranged to get

$$x_1=4-2x_2. \tag{9}$$

This is substituted into equation (8) to get

$$5(4-2x_2)-3x_2=7, \tag{10}$$

which can be rearranged to get

$$-13x_2=-13$$

or $x_2=1$. This result can now be substituted back into equation (9) to get $x_1=2$. The key idea is that, with each elimination step, the number of equations and number of unknowns involved is reduced by one. If there are initially N equations, $N-1$ elimination steps reduce the system to one equation in one unknown, which can be solved directly. A system of N equations in N unknowns is usually written as

$$A_{i1}x_1+A_{i2}x_2+\cdots+A_{iN}x_N=b_i,\quad i=1,\cdots,N, \tag{11}$$

where the A_{ij} are the coefficients, the b_i are the right-hand side values, and the x_j are the unknowns to be computed. The first of these equations ($i=1$) can be solved for x_1 to get

$$x_1=(b_1-A_{12}x_2-\cdots-A_{1N}x_N)/A_{11} \tag{12}$$

and this value of x_1 can be substituted into the remaining $N-1$ equations, which will then involve only $N-1$ unknowns. The same process can be repeated until a set of equations, each like equation (12) and each specifying an x_i in terms of x_j for $j>i$, is obtained. The Nth such equation specifies x_N directly. The value of x_N can then be substituted into the $(N-1)$st equation to get x_{N-1}, and so on. This is an entirely mechanical process that can be programmed easily for a computer. It is called an algorithm, a specification of a computational process that will determine the answer to a problem in a finite number of steps. It should be noted that if exact arithmetic could be used in this algorithm and if the division in equation (12) and subsequent divisions were possible (that is, if A_{11} and subsequent A_{ij} were not zero), the solution of the linear equations could be computed exactly.

In a practical computation, arithmetic cannot be performed exactly, so round-off errors will affect the answers. Their effect can be seen by considering a graphic interpretation of an equation: an equation of the form (7) represents a straight line on a graph relating x_1 and x_2, as in Figure 6. Any point on the line λ_1 shown has coordinates x_1 and x_2 that satisfy equation (7). Similarly, equation (8) represents another straight line shown as λ_2 in Figure 6. Any point on that line satisfies equation (8). Hence, the values of x_1 and x_2 that satisfy both (7) and (8) must lie on both lines, so that these values are the coordinates of the intersection of the two lines. If the lines are not parallel, they have a unique intersection, so that the equations have a unique solution.

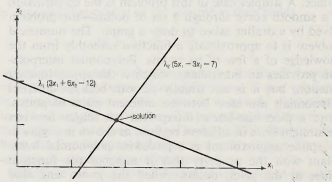


Figure 6: Graphic solution of a pair of linear equations (see text).

Round-off errors in the computation have an effect similar to that of changing the coefficients of the equations. Hence, in the presence of round-off errors, a computer will work with a narrow region that can be thought of as a "thick line," as shown in Figure 7, rather than an ideal line. The coordinates of any point inside the shaded region satisfy the equation of the line within the accuracy of the errors introduced into the coefficients. A pair of equations represents two thick lines within the accuracy of their coefficients, and the solution determined by the pair of equations lies somewhere within the intersection of the two thick lines, as shown in Figure 8.

If the equations represent a pair of lines that are nearly parallel, the numerical problem represented by the pair

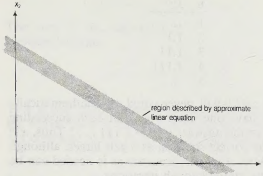


Figure 7: *Effect of round-off errors on the graphic representation of a linear equation.*
If the coefficients could be expressed exactly, the equation would be represented by an ideal line (having zero width), but because the coefficients can be expressed with limited accuracy, the line must be replaced by a band of finite width.

of thick lines in Figure 9 does not have an accurately determined answer because the region of intersection is large. The problem is said to be ill-conditioned. Any slight changes in the coefficients can cause large changes in the answer, so any round-off errors in the numerical procedure may cause large changes in the answer.

In the Gauss elimination–back substitution procedure described above, it was noted that it could always be completed unless one of the A_{ii} divisors was zero. If one of those divisors is very small, intermediate results may be very large, and the final answer may be obtained as the difference of two large numbers. This means that cancellation may occur so that there may be a large effect due to round-off. This effect can be seen in a graphic view of the process of Gaussian elimination. As described above, the equations represent a pair of lines. Dividing an equation by a number, as in dividing (7) by 3 to get (9), does not change the line the equation represents. Substituting one equation into another, as in substituting (9) into (8) to get (10), gives another equation that also represents a straight line, λ_4. Since the solution of the original equations lies on the new line, this new line must pass through the intersection of the two original lines. It should be noted that equation (10) does not involve x_1, so that line λ_4 is horizontal, as shown in Figure 10. The final step, the back substitution, evaluates x_2 from (10) and substitutes into (9)—which represents the same line as (7)—to get x_1. Thus, the solution is determined from lines λ_1 and λ_4. In Figure 6 they intersect nearly at right angles, as do λ_1 and λ_2. Hence, both the original problem of finding the intersection of λ_1 and λ_2, and the intermediate problem of finding the intersection of λ_1 and λ_4, are well-conditioned. If the line λ_1 had been nearly horizontal and the line λ_2 had been nearly vertical, however, the result of the Gaussian elimination would have been a nearly horizontal line, λ_4, to be used with the nearly horizontal line λ_1 to determine the solution. As described earlier, this is an ill-conditioned problem subject to large numerical changes in the answer. In this case, although the initial problem is well-conditioned (lines λ_1 and λ_2 intersect nearly at right angles), the intermediate problem given by lines λ_1 and λ_4 is ill-conditioned because of the poor choice of the numbering of the equations. The algorithmic solution to this

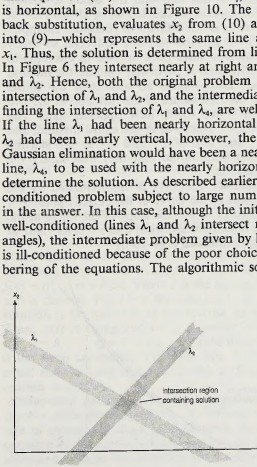


Figure 8: *Graphic solution of two linear equations, showing the effect of round-off error.*
Instead of two ideal lines intersecting in a single point, two narrow bands intersect in a region of finite area.

is to reorder the equations during the elimination process to ensure that the A_{ii} elements are large. This is called partial pivoting. Even with partial pivoting, solution errors for ill-conditioned problems can be large. A formal error analysis provides a pessimistic bound of the errors that can occur even for a well-conditioned problem, especially as the number of equations increases. Before the introduction of electronic computers this situation did not present a difficulty because large problems could not be tackled. Around 1950 it was thought unlikely that computers could produce accurate solutions of large systems of linear equations. The English mathematician J.H. Wilkinson and other researchers, who had been solving problems by hand methods, showed that problems could frequently be solved much more accurately than the conventional analysis predicted. Finally, Wilkinson's backward error analysis showed that the Gaussian elimination process would solve a system of equations with a residual or backward error that did not increase too rapidly as the size of the problem increased. Experience has confirmed Wilkinson's theoretical result: the error does not increase very much with the size of the problem.

Partial pivoting

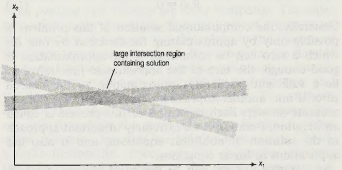


Figure 9: *Graphic solution of two linear equations for which the slopes of the lines are nearly the same.*
Because the region of intersection is large, the error in the solution is large.

Today it is common to solve sets of hundreds of thousands of linear equations on computers. There is major interest in ways in which such large problems can be solved rapidly. If the system is dense, which means that most of the coefficients, A_{ij}, are not zero, the Gaussian elimination process takes an amount of time equal to $aN^3 + bN^2 + cN + d$, where a, b, c, and d are constants that depend on the computer. As N gets large, the dominant term in the time is aN^3, and subsequent doublings in the problem size result in eightfold increases in the solution time. The method is said to have a running time asymptotic to N^3. The asymptotic behaviour of a method provides a measure of its speed for very large problems. For example, if a method has an asymptotic speed of $N^{2.5}$, it will execute more rapidly than one with asymptotic speed N^3 for very large N, although for small N it may not be faster. The asymptotically fastest method for linear equations is not known, and it is a question of considerable theoretical interest. It is known that they can be solved at asymptotic speeds close to $N^{2.3}$. For practical problems that will fit into computers, the Gaussian elimination technique remains the best method known for dense problems.

Comparing speeds of computations

Many large problems are sparse; that is, most of the coefficients A_{ij} are zero (because each equation describes a small piece of the physical system being modeled, and each piece is influenced only by a few other pieces). In such cases, the numerical analyst is interested in methods that do not change many coefficients from zero to non-zero (called fill-in). Gaussian elimination can still be used, and the equations can be re-ordered to minimize fill-in. In many cases of large sparse problems, however, iterative methods (discussed below) are preferable.

Nonlinear equations and iterative methods. Most accurate models necessitate the solution of nonlinear equations. Only in a few cases is it possible to solve such equations explicitly. The example of the quadratic equation $ax^2 + bx + c = 0$ was used above. The roots of this can be written explicitly, but polynomial equations involving x^5 or higher powers cannot be solved explicitly. The general

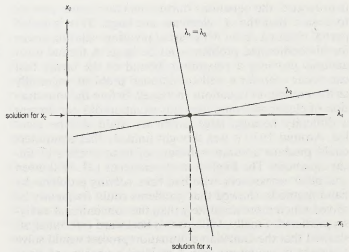


Figure 10: Graphic representation of the Gaussian elimination–back substitution procedure (see text).

problem is to find the value of x that gives a zero of a function $f(x)$, that is, to find a root of

$$f(x) = 0. \tag{13}$$

Generally, the computational solution of this problem is possible only by approximating the function by one for which a zero can be computed. If the approximation is good enough, the zero of the approximate function may be a sufficiently accurate approximation to the desired zero. If not, another approximation is used to get a more accurate answer. Each repetition of this process is called an iteration. Iteration is an extremely important approach to the solution of nonlinear equations, and it also has applications to linear equations.

An iterative method can be described as one in which knowledge about each succeeding approximate solution is used to compute a better solution. This knowledge is used to get a better approximation to the function. To illustrate the principle, the simplest of iterative methods is considered here, even though it is of value only in very limited applications. If x is added to both sides of equation (13), it becomes

$$x + f(x) = x, \tag{14}$$

an equation with the same roots as (13). It is now rewritten as

$$x = x - f(x). \tag{15}$$

Equation (15) is no easier to solve than (13), but the nonlinear right-hand side can be replaced with an approximation. In the first method, it is replaced by the constant value obtained by replacing x with a known value, say, c, to get

$$x = c - f(c). \tag{16}$$

This is certainly easy to solve for x since it is linear in x; in fact, it is necessary only to evaluate the right-hand side. This process can be turned into an iterative algorithm by taking the value computed for x and using it for the value of c in a second application of the process, and so on. An initial guess for the root is obtained somehow. (The problem solver will probably have some knowledge of an approximate value of the solution from experience. If not, any reasonable value, for example, $x = 0$, is chosen.) This initial guess is called $x^{(0)}$, and each new approximation is named $x^{(i)}$, $i = 1, 2, \ldots$. Then the relationship

$$x^{(n+1)} = x^{(n)} - f(x^{(n)}) \tag{17}$$

is obtained. The values $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$, form an infinite sequence. If they are such that $x^{(n+1)} = x^{(n)}$, it is clear that $x = x^{(n)} = x^{(n+1)}$ satisfies (15) and hence the original problem (13). This method is illustrated with the linear equation $f(x) = 0.9x - 1$ to keep the exposition simple (it is obvious that the zero of $0.9x - 1$ could be found directly with simple arithmetic). The iteration is

$$x^{(n+1)} = x^{(n)} - (0.9x^{(n)} - 1) = 1 + 0.1x^{(n)}. \tag{18}$$

If $x^{(0)} = 0$, the following sequence of values for $x^{(n)}$ is obtained:

n	$x^{(n)}$
1	1.
2	1.1
3	1.11
4	1.111
5	...

It is clear that $x^{(n+1)}$ will never equal $x^{(n)}$ mathematically because it will have one more digit, but each succeeding value is closer to the answer, $10/9 = 1.111\ldots$. Thus, $x^{(n)}$ converges to the correct solution as n gets larger, although it never gets there. As much accuracy as is needed can be obtained by performing enough iterations.

The iterative process described above (which is called functional iteration) can be described graphically as follows: Figure 11 shows a simple function $f(x)$ with a zero at x^*. The current guess, $x^{(n)}$, is also shown along with $f(x^{(n)})$. The right-hand side of (17) can be constructed by drawing a 45° right-angle triangle as shown to get $x - f(x^{(n)})$ on the x-axis. This is the next iterate. In Figure 11 the graph of $f(x)$ has been approximated by a 45° line. This view suggests that a better approximation to the curve could be used. This is the basis of many methods, of which two will be described, the chord method and Newton's method.

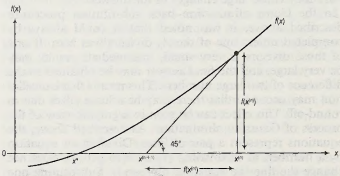


Figure 11: Functional iteration for successive approximations to x^*, the value of x for which $f(x)$ is zero (see text).

Chord method

In the chord method, two different values of x, $x^{(0)}$, and $x^{(1)}$, are selected, and a straight line is drawn through them to approximate the curve. The next point, $x^{(2)}$, is given by the intersection of this line with the x-axis, as shown in Figure 12. The process is repeated using $x^{(1)}$ and $x^{(2)}$ to get $x^{(3)}$, and so on. It is not difficult to construct examples for which the chord method will not converge, so many variants of the method have been devised in attempts to improve its performance. A particular variant, the method of false position, always finds an answer if the initial guesses are such that the values of the function at the two initial points, $f(x^{(0)})$ and $f(x^{(1)})$, have opposite signs. The method computes $x^{(2)}$ exactly as in the chord method, and it then uses $x^{(2)}$ and one of $x^{(0)}$ and $x^{(1)}$ as the starting point for the next iteration so that they also satisfy the opposite-sign criterion.

Newton's method

In Newton's method, the curve is approximated by the straight line tangent to the curve at the current point $x^{(n)}$, as shown in Figure 13. If s is the slope of the curve at the point x (s is the derivative $d[f(x)]/dx$) and $s^{(n)}$ is its value at $x^{(n)}$, then the approximation is given by

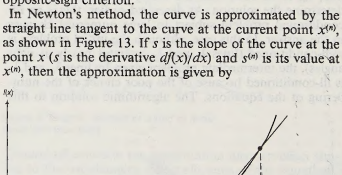


Figure 12: Chord method of iteration, in which each successive approximation is based on two previous approximations (see text).

$$f(x) \approx f(x^{(n)}) + s^{(n)}(x - x^{(n)}).$$

The zero of the approximation is computed to get the Newton formula

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{s^{(n)}}. \tag{19}$$

This is one of the fastest methods for finding zeros when the initial guess is fairly close to the answer, but $x^{(n)}$ may not converge to the root if the initial guess is far away.

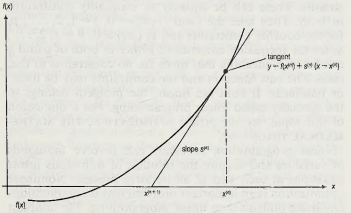


Figure 13: Newton's method of iteration, in which each successive approximation is based on the preceding one and the slope of the tangent line (see text).

Newton's methods and other methods can be applied to systems of nonlinear equations in n unknowns of the form

$$f_i(x_1, x_2, \ldots, x_n) = 0, \; i = 1, 2, \ldots, n, \tag{20}$$

by forming a linear approximation to the functions of many variables and using the techniques for solving linear equations discussed above.

The first example of the iterative solution was devoted to a linear equation. Iterative methods are an efficient way to solve large, sparse linear systems.

General nonlinear equations can be very difficult to solve computationally, in part because it is not possible to know where a solution may exist from local knowledge of the functions. (Local knowledge means information derived from the values of the function at a few points near a particular x.) Methods that use a local approximation to estimate the solution may compute an approximant that is in error by more than the current best approximant, so that the approximants diverge. For this reason, many iterative methods work only if the initial approximant is near the solution. Continuation methods are ones that construct a class of problems based on a parameter such that problems with nearby values of the parameter have solutions that are close. Hence, if the solution for one value of the parameter is known, it forms a good initial guess to the solution for a nearby value of the parameter. After the problem has been solved for the nearby parameter value, the process can be repeated. If the process starts with a value of the parameter for which the solution is known, the solution can be computed for a sequence of other values of the parameter until the original problem has been solved. For example, if the class of problems $f(x) = tf(x_0)$ is solved, where t is the parameter, a function $x(t)$ is defined. For $t = 1$ the solution is x_0. For $t = 0$, the solution is the solution of the original problem, $f(x) = 0$. Thus, $x(t)$ is computed for a sequence of values $t_0 = 1$, $t_1 = 1 - h, \ldots, t_n = 1 - nh, \ldots$, until the solution for $t_N = 0$ has been computed. The step size, h, is chosen to be small enough such that $x(t_n)$ forms a close first approximation to $x(t_{n+1})$.

Differential equations. Models of systems that change involve differential equations. A derivative describes the rate of change, or slope, of a function, and a differential equation is a relation between the rate of change of variables and expressions involving those and other variables. For example, the equations of motion for a space rocket relate the rate of change of the velocity of the rocket (called its acceleration) and its mass, to the forces on the rocket due to the thrust from its engines, the drag due to air resistance, and the force of gravity. This is Newton's second law, written as

$$a = \frac{dv}{dt} = \frac{\textit{total force on body}}{\textit{mass of body}},$$

where a is the acceleration, v the velocity, and t the time. This is called an ordinary differential equation: the unknowns (v, etc.) are functions of a single independent variable, t, only. The equations describing air flow around an airplane wing describe the rate of change, in time of the velocity of the air at every point in terms of the rates of change, with respect to spatial position, of pressure, density, and velocity of the air at that point. These are partial differential equations: the unknowns are functions of both space and time.

Initial-value and boundary-value problems

There are two different types of problems to be solved in differential equations, initial value problems and boundary value problems. The former type arises when the values of the unknowns are given at a particular point, for example, at a given time, and the values at future times are to be computed. For example, the position of the rocket before launch may be known, and future values of its position and velocity are to be computed. The latter type arises when the values of the unknowns are given on the boundary of a region and the values in the interior of the region are to be computed. For example, the flow of air a great distance from the airplane wing may be known to be uniform with a specified velocity and pressure, and the flow near the wing is to be computed. The methods used for these two types of problems are somewhat different, but they share the most important feature: a differential equation describes an infinite-dimensional object, a finite-dimensional object must be chosen to represent the solution, and the value of that approximation must be computed.

Finite difference methods

There are two main forms of approximation: finite differences and finite elements. In finite difference methods, the solution is represented by its values at a finite number of points, called mesh points. For example, if the velocity of the space rocket is to be computed, its velocities v_n at the set of times $t_n = nh$, $n = 0, 1, \ldots, N$, where h is the mesh spacing, could be computed as an approximation. If a differential equation of the form

$$s = \frac{dy}{dt} = f(y, t)$$

is given, where s is the slope, or derivative, of the unknown function $y(t)$, s can be replaced with the slope of a polynomial through a few neighbouring points to get

$$\textit{slope of polynomial through } [y_{n+i}] = f(y_n, t_n), \tag{21}$$

where y_n is the computed approximation to $y(t_n)$. For example, if a straight-line approximation through points y_n and y_{n+1} is used, the approximate equation becomes

$$\frac{y_{n+1} - y_n}{t_{n+1} - t_n} = f(y_n, t_n).$$

This method is called Euler's method. It is usually written $y_{n+1} = y_n + hf(y_n, t_n)$. If y_0 is given as an initial value, this formula can be used to compute successive values of y_1, y_2, etc. Euler's method is the simplest of all methods but is seldom the most efficient.

Finite element methods

In the finite element method, the region over which the solution is to be computed is divided up into a number of small subregions called elements. The unknown solution is expressed as a sum of basis functions $\varphi_i(x)$, which are polynomials in each of the elements, continuous across the boundaries of the elements (and possibly having continuous derivatives), and such that each is zero except in a very small number of elements. The simplest basis functions are the linear functions, which will be illustrated in one dimension. In this case, an interval (say, [0, 1]) of the real line is divided into a series of subintervals. These are called $[x_i, x_{i+1}]$ for $i = 0, 1, \ldots, N$, where $x_0 = 0$, $x_i < x_{i+1}$, and $x_{N+1} = 1$. The ith basis function is as shown in Figure 14. It is zero except in the intervals from x_{i-1} to x_{i+1}, where it rises to 1 linearly and then falls back to zero. It is known

as a hat function. (The zeroth hat function is 1 at x_0 and falls to zero at x_1, while the $(N+1)$st rises from zero at x_N to 1 at x_{N+1}.) The unknown y is expressed as

$$y = a_0\varphi_0(x) + a_1\varphi_1(x) + \cdots + a_{N+1}\varphi_{N+1}(x). \qquad (22)$$

This is a piecewise-linear function.

The next step in the finite element method is to use the differential equation to compute the unknown parameters a_i in (22). This function cannot satisfy the differential equation in general because the differential equation implies relations at all points in the interval. There are an infinite number of points, but there are only a finite number of parameters to be chosen. There are several ways to proceed. One is to apply the differential equation at a finite number of points only. This is called collocation and is applicable in some situations. Another way, valuable when the original problem satisfies a minimum-energy principle, is to seek the minimum-energy form of the approximation. This is an important method for many physical problems because they frequently satisfy a minimum-energy principle. For example, the shape assumed by a soap bubble is the one that has the least energy.

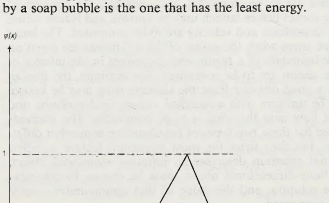


Figure 14: Example of a hat function used as a basis function in the finite-element method of finding approximate solutions of differential equations (see text).

As more points or elements are used, the mesh spacing diminishes, and the truncation error or residual in methods such as Euler's method becomes smaller. Methods for which these errors diminish as the mesh spacing is reduced are called consistent. Unfortunately, consistency alone is not enough to guarantee accuracy, because the number of mesh points or elements, and hence the number of equations such as (22), is increasing at the same time. Each of these approximate equations contributes a small error, and it is possible that the contribution of an increasingly large number of increasingly small errors could have a large effect. A critical property of methods is their stability. The technical definition of stability depends on the type of problem, but the essential idea is that as the number of mesh points (and approximate equations) increases, the effect of a single small error remains bounded. The so-called equivalency theorem for properly posed initial value problems in differential equations states that the consistency and stability of a method guarantee the convergence of the answer—that is, the errors in the answer shrink to zero as the mesh spacing is reduced.

Stability of methods

APPLICATIONS AND IMPLEMENTATION

Methods for the numerical approximation and solution of equations are applied to complex problems that arise in the utilization of computers for a variety of tasks. The numerical solution of these problems requires their decomposition into sequences of subtasks that can be solved by standard methods. Many of these methods are made available to users through mathematical software that consists of programs organized to provide reliable solutions to large classes of problems on various types of computers. This section discusses two important applications and the implementation of methods in software for high-speed computation.

Numerical optimization. Many problems involve finding the best—in some defined respect—of many possible solutions. The best solution might be the one leading to the lowest cost, the largest profit, or the shortest journey. Such problems are ones of optimization. Because of their economic importance, their efficient computational solution is extremely important. The mathematical problem is to find a set of values x_i such that $\varphi(|x_i|)$ is as small (or as large) as possible, where $|x_i|$ means the set of all x_i, $i = 1, 2, \ldots, n$. The function φ represents the cost, the profit, or other value to be optimized. It is called the cost function, or objective function, and the problem is usually defined so that the cost is to be minimized. In many problems, the variables, x_i, are subject to constraints. These can be equality or inequality constraints, or both. They take the form $f_j(|x_i|) = 0$, $j = 1, 2, \ldots, p$, for the equality constraints and $g_k(|x_i|) \geq 0$, $k = 1, 2, \ldots, q$, for the inequality constraints. Either or both of p and q can be zero, meaning that there are no constraints in that class. The cost function and the constraints may be linear or nonlinear. If both are linear, the problem belongs to the specialty called linear programming. For a discussion of this topic, see the article OPTIMIZATION, THE MATHEMATICAL THEORY OF.

Linear programming

Linear programming problems can involve thousands of variables and require the solution of numerous linear equations at each step of an iterative process. Nonlinear optimization is an important problem that is computationally more difficult than linear programming. The simplest case is the minimization of a function of a single variable, $\varphi(x)$. If the function is as shown in Figure 15, the problem is to find the point marked P. At that point the curve has a horizontal tangent; that is, its derivative is zero. One method is to solve the nonlinear equation $d\varphi(x)/dx = 0$ using the methods discussed for nonlinear equations. This method, however, requires a technique for evaluating the derivative, which may not be convenient or possible in some applications. In such cases, the function $\varphi(x)$ may be evaluated at a number of points and this information used to construct an approximation to the function locally, usually as a quadratic function that has a minimum. This approximation, however, is not a reliable indicator of the minimum, and it is normally used to indicate in which direction to look for a smaller value of the function. In the case of a function of several variables ($n > 1$), the derivative or the function approximation indicates a direction in which to move, that is, a line in the space of x_i variables. The users of most optimization algorithms, in effect, search along that line for a smaller value of the cost function.

Because of the nonlinearities, these problems are very difficult, and it is not possible to develop methods that always work. Even in the simple example shown in Figure 15, a method might find the local minimum at Q instead of the global minimum at P.

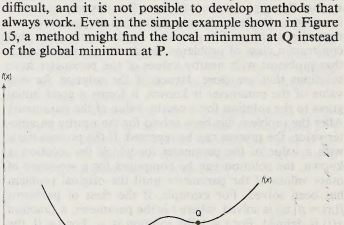


Figure 15: *Nonlinear optimization.* In seeking the global minimum P, computational difficulties may lead the procedure to the local minimum Q (see text).

Inverse problems. An inverse problem takes the form: "Here is the answer, what was the question?" These problems arise when a person needs to determine some information, such as the density of different regions in the interior of a human head to check for medical problems, but it is not possible to gather all of the information needed by direct methods, as it clearly is not in this

example. What the person can determine is information concerning the amount of X rays that can be transmitted along various line segments through the head. This information will give an estimate of the average density along the whole line but not at any specific interior point. The inverse problem in this example is to start with measurements along many different lines and with these data to calculate the density at all interior points.

Inverse problems are ill-posed because there are usually many different questions that could have yielded the same answer. This is simply illustrated in the interpolation and approximation problems discussed above. A function is infinite dimensional, so that if its values are given at a finite set of points, its values cannot be determined everywhere; in fact, there are infinitely many functions passing through any given set of points (as many different functions can be drawn as desired). In computing an approximation, assumptions are made concerning the nature of the function—for example, that it is a polynomial or has some other form. The same approach is taken to inverse problems: assumptions are made about the nature of the unknown function so that the finite amount of data in the observations is sufficient to determine that function. This is called regularization and converts the ill-posed problem into a properly posed one.

Large-scale scientific computation. The increasing speed of computers and advances in numerical methods have made it possible to solve most small problems rapidly by means of readily available software, and the attention of many numerical analysts has turned to the solution of problems so large that they require inordinate amounts of computer time. The use of the most efficient method has much greater importance in large problems than it does in small ones. In the latter the flexibility of human use of the program may be of much greater importance. One measure of the efficiency of a numerical method is the number of arithmetic operations needed to solve a problem. (Others include the amount of memory used and the order in which it is used, but these issues are more properly subjects of computer science.) Computational complexity is the study of the inherent computational cost (measured by number of operations) of solving a problem. The result of a complexity analysis is an estimate of how rapidly the solution time increases as the problem size increases. It can be used to analyze problems and assist in the design of algorithms for their solution.

Computational complexity

On a serial computer, which is one with a single processor capable of performing only one arithmetic operation at a time, the time needed to solve a problem is proportional to the number of operations, so that this is the number that the numerical analyst tries to reduce in algorithm design. Because parallel computers have more than one processor, they are capable of performing several arithmetic operations simultaneously. To make effective use of a parallel computer, an algorithm must have parallelism; that is, it must require the simultaneous performance of certain operations. For example, evaluation of the expression $a \times b + c$ requires two operations, but the multiplication must be completed before the addition can be started: they must be done serially. On the other hand, in evaluating the expression $w + x + y + z$, two of the additions can be performed in parallel, for example $w + x$ and $y + z$, and then their results can be added. In this case there are three arithmetic operations, but a suitably organized parallel computer can complete the sum in the time it takes to do two additions. In the design of numerical algorithms for parallel computers, one relevant criterion is the number of nonoverlapped operations needed.

Mathematical software. An important product of the numerical analyst's activity is mathematical software. It consists of computer programs designed to solve classes of numerical problems both automatically and reliably. Such software can be used by people with little knowledge of numerical techniques, and the programs themselves determine, on the basis of evaluations of the mathematical expressions provided by the user, the best method to use and the parameters of that method. For example, in software for differential equations, the program might decide which of several methods to use and what mesh spacing to use for the method chosen. Since there are infinitely many problems and a program can evaluate a function only at a finite number of points, it is possible for such software to make mistakes (but so can people who try to solve problems). Robust software is software that is not only reliable but that also attempts to determine when it has encountered an intractable problem and informs the user that an answer cannot be found, rather than returning an incorrect one. The oft-heard statement that "any answer is better than none" is not true for most computational problems, as any passenger on an airplane—pondering the amount of numerical computation used in its safe and efficient design—realizes. The challenge to numerical analysts is to develop methods that compute accurate results efficiently or to recognize that they cannot do so.

BIBLIOGRAPHY

General works: HERMAN H. GOLDSTINE, *A History of Numerical Analysis from the 16th Through the 19th Century* (1977); KENDALL ATKINSON, *Elementary Numerical Analysis* (1985); S.D. CONTE and CARL DE BOOR, *Elementary Numerical Analysis: An Algorithmic Approach*, 3rd ed. (1980); F.B. HILDEBRAND, *Introduction to Numerical Analysis*, 2nd ed. (1973); GERMUND DAHLQUIST and ÅKE BJÖRCK, *Numerical Methods*, trans. from Swedish (1974); ROBERT W. HORNBECK, *Numerical Methods* (1975); PETER A. STARK, *Introduction to Numerical Methods* (1970).

Numerical methods: ROBERT F. CHURCHHOUSE (ed.), *Numerical Methods* (1981); EUGENE ISAACSON and HERBERT BISHOP KELLER, *Analysis of Numerical Methods* (1966); GEORGE BOOLE, *Calculus of Finite Differences*, 5th ed., edited by J.F. MOULTON (1970); GILBERT STRANG and GEORGE J. FIX, *An Analysis of the Finite Element Method* (1973); J.H. WILKINSON, *Rounding Errors in Algebraic Processes* (1964); J.F. STEFFENSEN, *Interpolation*, 2nd ed. (1950); JOHN R. RICE, *The Approximation of Functions*, 2 vol. (1964–69); L. COLLATZ, *Functional Analysis and Numerical Mathematics* (1966; originally published in German, 1964); and E.W. CHENEY, *Introduction to Approximation Theory*, 2nd ed. (1982).

Solution of equations: J.H. WILKINSON and C. REINSCH, *Linear Algebra* (1971); BEN NOBLE and JAMES W. DANIEL, *Applied Linear Algebra*, 2nd ed. (1977); A.S. HOUSEHOLDER, *The Numerical Treatment of a Single Nonlinear Equation* (1970); J.E. DENNIS, JR., and ROBERT B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (1983); GEORGE E. FORSYTHE, "What Is a Satisfactory Quadratic Equation Solver?", pp. 53–61 in BRUNO DEJON and PETER HENRICI (eds.), *Constructive Aspects of the Fundamental Theorem of Algebra* (1969); J.M. ORTEGA and W.C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables* (1970); J.F. TRAUB, *Iterative Methods for the Solution of Equations*, 2nd ed. (1982); A.M. OSTROWSKI, *Solution of Equations in Euclidean and Banach Spaces*, 3rd ed. (1973); L.M. MILNE-THOMSON, *The Calculus of Finite Differences*, 2nd ed. (1981); PETER HENRICI, *Discrete Variable Methods in Ordinary Differential Equations* (1962); G.D. SMITH, *Numerical Solution of Partial Differential Equations: Finite Difference Methods*, 3rd ed. (1985); C. WILLIAM GEAR, *Numerical Initial Value Problems in Ordinary Differential Equations* (1971); L.G. CHAMBERS, *Integral Equations: A Short Course* (1976); and GENE H. GOLUB and CHARLES F. VAN LOAN, *Matrix Computations* (1983).

Applications and implementations: G.M. PHILLIPS and P.J. TAYLOR, *Theory and Applications of Numerical Analysis* (1973); R.W. HAMMING, *Numerical Methods for Scientists and Engineers*, 2nd ed. (1973, reprinted 1987); CURTIS F. GERALD and PATRICK O. WHEATLEY, *Applied Numerical Analysis*, 3rd ed. (1984); CARL-ERIC FRÖBERG, *Numerical Mathematics: Theory and Computer Applications*, rev. ed. (1985; originally published in Swedish, 1962); GEORGE E. FORSYTHE, MICHAEL A. MALCOLM, and CLEVE B. MOLER, *Computer Methods for Mathematical Computations* (1977); BRUCE W. ARDEN (ed.), *What Can Be Automated? The Computer Science and Engineering Research Study* (1980); and C. WILLIAM GEAR, "Numerical Software: Science or Alchemy?", *Advances in Computers*, 19:229–248 (1980).

(C.W.Ge.)

Nutrition

Nutrition is the process by which an organism absorbs and utilizes food substances. The study of nutrition involves the identification of individual nutrients essential for growth and for the maintenance of individual organisms; it includes the determination of interrelationships among nutrients within individual organisms, as well as the evaluation of the quantitative requirements of organisms for specific nutrients under various environmental conditions. Treated in this article are the identification and determination of the requirements of nutrients, the ways in which they vary from one organism to another, and the ways in which requirements for certain of them arose. Human nutritional requirements are then discussed, followed by a review of the diseases and disorders related to nutrition.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 421, 422, and 423.

This article is divided into the following sections:

General features of nutrients and nutrition 50
  Functions of food 50
    Nutritional patterns
    Nutritional evolution of organisms
    Methods of ingestion or penetration of nutrients
  Determination of essential requirements 51
  Essential inorganic nutrients 52
  Essential organic nutrients 52
    Amino acid requirements
    Purine and pyrimidine bases and related compounds
    Lipids and related compounds
    Vitamins
    Interdependency of nutritional requirements
  Comparison of nutrient requirements 54
  Syntrophism 54
Human nutrition and diet 54
  Utilization of food by the body 54
    The supply of energy
    Building and maintenance
    Essential nutrients
  Classes of foods 56
    Cereals
    Starchy roots
    Legumes (pulses)
    Vegetables and fruits
    Sugars, preserves, and syrups
    Meat, fish, and eggs
    Milk and milk products
    Fats and oils
    Beverages
  Nutrient recommendations 59
    Setting DRIs and other dietary goals
    Energy
    Protein
    Vitamins
    Minerals
Diet and diseases 63
  Sources of information 63
    Animal experiments
    Comparative and evolutionary studies
    Experiments of nature and travelers' tales
    Epidemiological studies
    Clinical records
    Food composition analysis
    Human experiments and trials
    Classification of nutritional disorders
  Deficiency diseases 65
    Water
    Energy (calories)
    Protein
    Vitamins
    Fats and carbohydrates
    Inorganic elements
    Dietary fibre
  Obesity and weight control 69
    Obesity
    Eating disorders
  Nutrient toxicities 70
    Alcohol
    Vitamins
    Minerals
  Food sensitivity 71
  Habitual dietary patterns and disease 72
    Diet and coronary heart disease
    Diet and hypertension
    Diet and cancer
    Diabetes mellitus
    Dental caries
  Dietary guidelines 74
Bibliography 75

General features of nutrients and nutrition

FUNCTIONS OF FOOD

Food serves three functions in most living organisms. First, it provides materials that are metabolized either by oxidative or by fermentative processes to supply the energy required for the absorption and translocation of nutrients, for the synthesis of cell materials, for motility and locomotion, for excretion of waste products, and for all other activities of the organism. Second, food supplies the electron donors (reducing agents) required for the formation of the reduced coenzymes necessary for the synthetic processes that occur within the cell. Third, food provides the materials from which all of the structural and catalytic components of the living cell can be assembled by processes sometimes called anabolism. The three roles of food are not mutually exclusive; energy-yielding substances in many organisms may function in all three ways.

Anabolism

The essential precursors (*i.e.,* the substances from which other substances are formed) of cell materials can be divided into two groups—nonessential nutrients, which can be synthesized by the cell from other materials, and essential nutrients, which, because they cannot be synthesized by the cell, must be supplied in foods. All of the inorganic materials required for growth, together with an assortment of organic compounds whose number may vary from one to 30 or more, depending on the organism, fall into the latter category. Although organisms are able to synthesize nonessential nutrients, such nutrients are frequently utilized directly if present in food, thereby saving the organism the need to expend the energy required to synthesize them.

Nutritional patterns. One method of classifying living organisms on a nutritional basis centres on the way in which the functions of food are carried out. Thus, organisms such as green plants and some bacteria that need only inorganic compounds for growth are called autotrophic organisms; and organisms, including all animals, fungi, and most bacteria, that require both inorganic and organic compounds for growth are called heterotrophic. Although these general terms are widely employed, the classification has been broadened to describe various nutritional patterns that have been observed. In one scheme, organisms are classified according to the energy source they utilize. Phototrophic, or photosynthetic, organisms trap light energy and convert it to chemical energy in the form of an energy-rich compound, adenosine triphosphate (ATP); chemoautotrophic, or chemosynthetic, organisms utilize inorganic or organic compounds to supply their ATP requirements. An additional method of differentiation is based on the type of electron-donor material utilized to form the reduced coenzymes required for synthesis of cell

Autotrophic and heterotrophic organisms

constituents. If the electron-donor material consists of inorganic compounds, the organism is said to be lithotrophic; if organic, the organism is organotrophic. Combinations of these patterns may also be used to describe organisms; higher plants, for example, are photolithotrophic; *i.e.*, they utilize light energy, with the inorganic compound water serving as the ultimate electron donor. Certain photosynthetic bacteria that cannot utilize water as the electron donor and require organic compounds for this purpose are called photo-organotrophs. Animals, by this classification, are chemoorganotrophs; *i.e.*, they utilize chemical compounds to supply ATP and organic compounds as electron donors.

Despite wide variations in the nature of the external energy source utilized by various organisms, all organisms form from the external energy source the immediate intracellular source of energy, ATP, which is common to all cells. Through the breaking of the high-energy phosphate bonds of ATP and thus by its conversion to a less energy-rich compound, adenosine diphosphate (ADP), ATP provides the energy for the chemical and mechanical work required by an organism. The energy requirements of the organism can be measured in either joules or calories; the need for cell-precursor materials cannot be.

Chemo-organotrophs

All animals, fungi, and most bacteria are chemoorganotrophs (*i.e.*, organic compounds supply both ATP and serve as electron donors). In these organisms the multiple roles played by foodstuffs are partially obscured because the carbon-containing compounds that serve as energy sources also serve as electron donors and provide intermediate compounds from which many cell materials are synthesized. Similarly, essential organic nutrients, when present in excess, may be utilized in part to provide the energy needs of the organisms.

Nutritional evolution of organisms. Little is known about the nutritional evolution of living organisms. Nucleic acids, proteins, carbohydrates, and fats, which are present in all living cells, are formed by specific reaction sequences from a limited number of smaller compounds, most of which are common to all living organisms and, according to current theories, were available on Earth before life arose. Since less complex metabolic organization and less energy are required to synthesize cellular proteins from preformed amino acids than from carbon dioxide and other precursors, it is assumed that the simplest early forms of life were heterotrophic organisms requiring many organic nutrients for growth and that they selected such nutrients from their surroundings. As the supply of these preformed substances was exhausted the organisms presumably developed the capacity to synthesize these preformed substances from simpler (precursor) materials present in the environment; in some organisms, this synthesizing capacity eventually evolved to the extent that carbon from carbon dioxide could be utilized to synthesize organic compounds.

At this point, autotrophy, as it now is known, became possible; autotrophy, in fact, may have evolved as a result of the exhaustion of the supply of preformed organic materials in the environment and the consequent necessity of organisms to synthesize the requirements themselves in order to survive. Implicit in this theory is the demonstrable assumption that autotrophic cells contain the most complex biosynthetic organization found in living things and that heterotrophic cells are simpler in that certain biosynthetic pathways do not occur. After the evolution of photosynthesis, a constantly renewable source of the organic compounds necessary for heterotrophic cell growth became available. It became feasible that those organisms whose environments provided a constantly available supply of a given compound could lose, through changes in their genetic material (mutations), the ability to synthesize that compound and still survive. Entire biosynthetic pathways may have been lost in this way; as long as such mutant organisms remained in an environment that supplied the necessary compound, the simplification in cellular organization and the energy saved by using preformed cell components would have given them a competitive advantage over the more complex parents from which they were derived and permitted stabilization of the mutation within the cell type. A theory that the requirements of modern organisms for essential organic nutrients arose through the loss of synthetic abilities present in more complex parent organisms was confirmed between 1940 and 1950 by the discovery that artificially produced mutant offspring of microorganisms can be readily obtained and may require the presence of one or more preformed organic compounds, which the parent microorganisms could synthesize.

Effects of mutation

Methods of ingestion or penetration of nutrients. Foodstuffs of chemoorganotrophs are composed chiefly of large molecules, such as nucleic acids, proteins, lipids, and carbohydrates, which are derived from other organisms. When these materials are eaten as food, they usually must be broken down into the low-molecular-weight substances from which they were formed (*e.g.*, purine and pyrimidine bases from nucleic acids, amino acids from proteins, fatty acids from fats, and simple sugars from carbohydrates) before they can be absorbed and utilized. The breakdown process is generally accomplished outside the cells (extracellularly) of both unicellular and multicellular organisms through hydrolysis (breakdown involving water), catalyzed by specific enzymes secreted for this purpose. Some unicellular organisms (*e.g.*, the protozoan genus *Amoeba*), however, engulf particulate matter and hydrolyze it intracellularly.

Absorption of nutrients

The organic products of digestion produced extracellularly, together with essential inorganic nutrients, must cross one or more cell membranes before they can be utilized for growth. When nutrient concentrations are high, some materials may diffuse through the cell membrane to enter the cell. Absorption of some materials can occur from an extracellular solution containing small quantities of the nutrients to an intracellular solution containing larger quantities (active transport) and is catalyzed within the cell membrane by systems that are specific for the type of compound being transported. Those catalysts best characterized involve either specific proteins with high affinity for the nutrient substance being transferred or enzymes that chemically modify the nutrient during its passage through the membrane or both.

DETERMINATION OF ESSENTIAL REQUIREMENTS

Trace elements

It has been estimated that the approximate quantities of chemical elements in the human body (in percent of wet weight) are oxygen, 65; carbon, 18; hydrogen, 10; nitrogen, 3; calcium, 2; phosphorus, 1.1; sulfur, 0.25; potassium, 0.20; sodium, 0.15; chlorine, 0.15; magnesium, 0.05; iron, 0.004; copper, 0.0015; manganese, 0.0013; and iodine, 0.00004. Also present are traces of about 20 other elements—*e.g.*, zinc, cobalt, silicon, molybdenum, selenium, aluminum, arsenic, barium, boron, bromine, cadmium, chromium (Table 1). Analytical figures such as those for humans given above serve to distinguish between chemical elements that occur in relatively large amounts—oxygen, carbon, nitrogen, sulfur, and phosphorus—and the so-called trace elements, an obsolescent term applied to elements that are present in small amounts and that in the past could be detected in tissues but not analyzed accurately.

The methods used to determine the pure substances necessary for growth are similar in principle for all organisms. Occasionally, crude diets lacking only one or a few nutrients result in some naturally occurring deficiency disease; the substance that cures such a condition can be determined directly by adding it to the diet. Certain vitamins were recognized in this way. In the absence of a deficiency condition, a crude nutrient medium (or diet) adequate for growth is selected for the organism to be studied, and attempts are then made to replace the various ingredients of the medium with purified materials and, finally, with substances of known composition. If replacement with substances of known composition permits growth, the added materials can be determined with relative ease; if the pure substances do not allow growth, however, chemical testing must be carried out to isolate and identify the substances in crude materials that permit growth. Detailed nutritional requirements of many organisms are known, and essentially all of the organic nutrients are available as pure compounds, so determination of the nutrient require-

ments of a previously unstudied organism is a comparatively easy task. Elucidation of trace-element requirements remains troublesome, however, because these substances occur commonly as contaminants in other materials.

ESSENTIAL INORGANIC NUTRIENTS

Table 1 is a list of inorganic elements (minerals) that are essential for the growth of living things, together with one or more examples of organisms that exhibit a requirement for them. The effects of certain mineral deficiencies on the development of higher plants and animals are illustrated by several photographs. Naturally occurring deficiencies of a number of different trace elements occur in various parts of the world in special types of soils and are therefore of tremendous economic and agricultural importance.

Specialized mineral requirements

Several examples of specialized mineral requirements are present in Table 1. Boron, for example, has been demonstrated to be required for the growth of many—perhaps all—higher plants but has not been implicated as an essential element in the nutrition of either microorganisms or animals. Trace amounts of fluorine (as fluoride) are certainly beneficial, and perhaps essential, for proper tooth formation in higher animals, but no essential role for fluorine in other organisms has been found. Similarly, iodine (as iodide) is required in animals for formation of thyroxine, the active component of an important regulatory hormone; it plays no known role in other organisms. Silicon (as silicate) is a prominent component of the outer skeletons of diatomaceous protozoans and similar organisms and is required in them for normal growth. In higher animals, the requirement for silicon is much smaller. A less obvious example of a specialized mineral requirement is provided by calcium, which is required by higher animals in comparatively large amounts because it plays a role in the formation of bone and eggshells (in birds); for other organisms, calcium is an essential nutrient but only as a trace element. Several additional examples of specialized requirements are evident from Table 1. Mineral elements in wide variety are present in trace amounts in almost all foodstuffs. It cannot be assumed that, because a given mineral element is nonessential, it plays no useful role in metabolism.

Antagonistic mineral nutrients

Important antagonistic relationships between certain mineral nutrients also are known. A large excess of rubidium, for example, interferes with the utilization of potassium in some lactic-acid bacteria; zinc can also interfere with manganese utilization in the same organism. In animal nutrition, excessive molybdenum or zinc (both of which are essential minerals) interferes with the utilization of copper, another essential mineral, and, in higher plants, excessive zinc can lead to a disorder that is known as iron chlorosis. Proper nutrient growth media for microorganisms and plants or diets for animals, therefore, require not only that the essential mineral elements be provided in sufficient amounts but also that they be used in the proper ratios to each other.

Table 1: Essential Inorganic Nutrients for Living Organisms*

element	utilizable ionic form	representative organisms exhibiting the requirement
Boron	$B_4O_7^{2-}$	certain vascular plants and algae; no evidence of an animal requirement
Calcium	Ca^{2+}	plants, animals, most microorganisms
Chlorine	Cl^-	higher animals; no evidence for requirement in plants
Chromium	Cr^{3+}	probably essential in higher animals
Cobalt	Co^{2+}	essential in ruminants; probably functions chiefly through microbial incorporation into vitamin B_{12}
Copper	Cu^{2+}	plants, animals, most microorganisms
Fluorine	F^-	highly beneficial to bone and tooth formation in animals, including humans
Iodine	I^-	higher animals; no evidence for requirement in plants or microorganisms
Iron	Fe^{2+}	animals, higher plants, most microorganisms
Magnesium	Mg^{2+}	animals, plants, microorganisms
Manganese	Mn^{2+}	animals, plants, microorganisms
Molybdenum	MoO_4^{2-}	animals, plants, nitrogen-fixing bacteria
Nitrogen	NO_3^-, NH_4^+	plants, microorganisms (animals derive nitrogen mostly from organic sources and utilize limited amounts of NH_4^+, but not NO_3^-)
Phosphorus	PO_4^{3-}	animals, plants, microorganisms
Potassium	K^+	animals, plants, microorganisms
Selenium	SeO_4^{2-}	higher animals
Silicon	SiO_4^{4-}	certain Protozoa and Porifera
Sodium	Na^+	animals, some plants, some marine bacteria
Sulfur	SO_4^{2-}	plants, many bacteria (animals derive sulfur mostly from organic sources)
Vanadium	VO_4^{2-}	various tunicates and holothurian echinoderms; some algae; possibly higher animals
Zinc	Zn^{2+}	all animals, plants, most microorganisms

*Carbon dioxide (CO_2) and water (H_2O), although not formally tabulated, are important nutrients for all organisms; external sources of carbon dioxide, however, are required for only a few heterotrophic organisms since it is a prominent waste product of energy metabolism in these organisms.

ESSENTIAL ORGANIC NUTRIENTS

Considered in this section is the nature of the essential organic nutrients, necessary building blocks of various cell components that certain organisms can no longer synthesize and therefore must be supplied to them preformed.

Essential compounds

These essential compounds include certain amino acids, which are the precursors of protein; purine and pyrimidine bases or their derivatives, which are the precursors of nucleic acids; fatty acids, which are precursors of fats and phospholipids; and, occasionally (in mutant organisms), simple sugars (*e.g.,* glucose). Other essential organic nutrients include the vitamins, which are required in small amounts (relative to, for example, the amino acids), because of either the catalytic role or the regulatory role they play in metabolism.

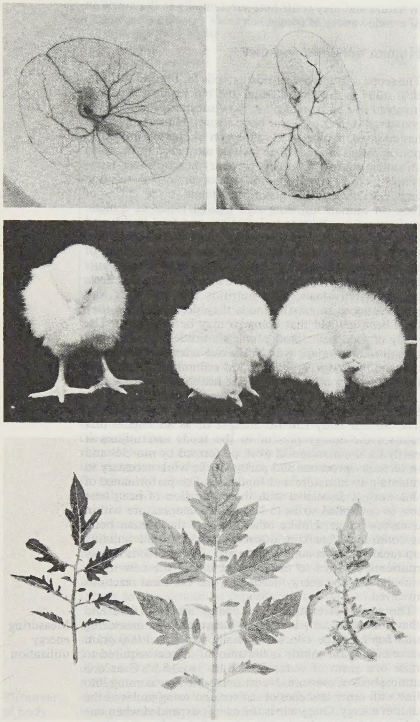
Amino acid requirements. Proteins usually contain about 20 different amino acids or their derivatives, and there are organisms that require none, several, or most of them for growth. Many yeasts and bacteria (*e.g., Escherichia coli*) are able to synthesize all of their amino acids from other compounds (glucose and ammonium salts). Animals may require up to nine or 10 different amino acids in their diets. Species of lactic-acid bacteria have been identified as requiring none (*Streptococcus bovis*), a few (*Lactobacillus plantarum*), or as many as 17 amino acids (*Streptococcus equinus*).

Purine and pyrimidine bases and related compounds. A large number of species of microorganisms require one or more of these substances as precursors for the synthesis of nucleic acids. All higher animals examined have retained the capacity to synthesize purine and pyrimidine bases. Certain compounds (putrescine, spermidine, spermine) that are closely associated with nucleic acids in living cells, although synthesized by most organisms, are required preformed by certain bacteria; *e.g., Haemophilus parainfluenzae, Neisseria perflava.*

Lipids and related compounds. Phospholipids are essential components of all cell membranes. Those organisms that have lost the ability to synthesize certain lipids (since they are supplied in their diets), require them as essential nutrients. Essential fatty acids for rats (and most vertebrates) include linoleic and arachidonic acids and, perhaps, linolenic acid as well. For many bacteria (*e.g., Erysipelothrix rhusiopathiae,* several species of lactic-acid bacteria), oleic acid is essential for growth, although it can be replaced in most cases by linoleic acid and by an unusual fatty acid called lactobacillic acid.

Choline, inositol, and cholesterol

Choline, a nitrogen-containing compound found in the phospholipid lecithin, is usually listed among the nutrients required by rats, although it can be synthesized by them if sufficient vitamin B_{12} is present. Choline is also required for growth of certain pneumococcal bacteria. Inositol, another structural component of certain lipids, is an essential nutrient for many yeasts but appears to be synthesized by most animals. Finally, the steroid cholesterol is an essential nutrient for a wide variety of protozoans and for the bacteria-like organisms known as mycoplasmas. Higher animals are able to synthesize this important cell constituent; it is not present in most bacteria.



Effects of mineral deficiencies on the development of higher plants and animals.
(Top left) Normal 72-hour embryo from hen fed 10 parts per million of supplementary copper. (Top right) Anemic 72-hour embryo from hen receiving copper-deficient diet. (Centre) A normal chick, left, from hen receiving an adequate amount of zinc in its diet and two chicks from hens given zinc-deficient diets. (Bottom) Tomato leaves from plant grown in (left to right) copper-deficient soil, complete nutrient solution, and zinc-deficient soil.
By courtesy of (bottom) D. Arnon; from (top left, top right, centre) J.E. Savage, *Federation Proceedings*, vol. 27, p. 907 (1968)

Vitamins. Vitamins may be defined as nutritionally essential organic substances that play a catalytic role within the cell, usually as components of coenzymes or other groups associated with enzymes. More than half of the water-soluble vitamins required by humans were discovered when they were found to be growth factors for microorganisms. Vitamin requirements are specific for each organism, and their deficiency may cause disease. Vitamin deficiencies in young animals usually result in growth failure, various symptoms whose nature depends on the vitamin, and eventual death. Certain vitamins (retinol, calciferol, tocopherol, and ascorbic acid) do not appear to play an essential role in unicellular organisms. Para-aminobenzoic acid is a vitamin for several bacteria only because it is an essential precursor of the vitamin folic acid, which, since it is unable to cross the bacterial cell membrane, must be synthesized from para-aminobenzoic acid. Heme and lipoic acid are typical vitamins for organisms that cannot synthesize them, but they have a catalytic role in all organisms; they are synthesized by higher animals. A few bacteria have lost the ability to convert certain vitamins to their functional forms, the coenzymes; the preformed coenzymes (or intermediates in their synthesis) thus are vitamins for such organisms.

Para-aminobenzoic acid

Interdependency of nutritional requirements. The effects of one mineral nutrient in reducing or increasing the requirement for another have been mentioned previously (see above *Essential inorganic nutrients*). Similar relationships occur among organic nutrients and originate for several reasons, the most common of which are discussed briefly below.

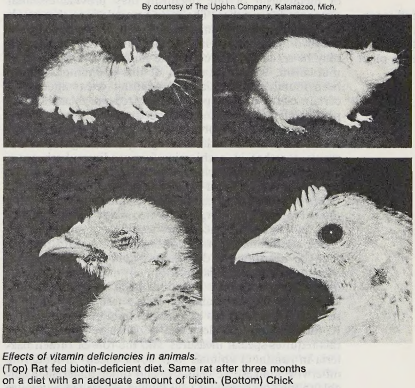
Competition for sites of absorption by the cell. Since absorption of nutrients frequently occurs by way of active transport within cell membranes, an excess of one nutrient (A) may inhibit absorption of a second nutrient (B), if they share the same absorption pathway. In such cases, the apparent requirement for nutrient B increases; B, however, can sometimes be supplied in an alternate form that is able to enter the cell by a different route. Many examples of amino acid antagonism, in which inhibition of growth by one amino acid is counteracted by another amino acid, are best explained by this mechanism. For example, under some conditions *Lactobacillus casei* requires both D- and L-alanine, which differ from each other only in the position of the amino, or NH_2, group in the molecule, and the two forms of this amino acid share the same absorption pathway. Excess D-alanine inhibits growth of this species, but the inhibition can be alleviated either by supplying additional L-alanine or, more effectively, by supplying peptides of L-alanine. The peptides enter the cell by a pathway different from that of the two forms of alanine and, after they are in the cell, can be broken down to form L-alanine. Relationships of this type provide one explanation for the fact that peptides are frequently more effective than amino acids in promoting growth of bacteria.

Amino acid antagonism

Competition for sites of utilization within the cell. This phenomenon is similar to that regarding competition for absorption sites, but it occurs inside the cell and only between structurally similar nutrients (*e.g.,* leucine and valine; serine and threonine).

Precursor-product relationships. The requirement of rats and humans for the essential amino acids phenylalanine and methionine is substantially reduced if tyrosine, which is formed from phenylalanine, or cysteine, which is formed from methionine, is added to the diet. These relationships are explained by the fact that tyrosine and cysteine are synthesized in animals from phenylalanine and methio-

By courtesy of The Upjohn Company, Kalamazoo, Mich.



Effects of vitamin deficiencies in animals.
(Top) Rat fed biotin-deficient diet. Same rat after three months on a diet with an adequate amount of biotin. (Bottom) Chick fed a diet deficient in pantothenic acid. Same chick after three weeks on a diet sufficient in pantothenic acid.

nine, respectively. When the former (product) amino acids are supplied preformed, the latter (precursor) amino acids are required in smaller amounts. Several instances of the sparing of one nutrient by another because they have similar precursor-product relationships have been identified in other organisms.

Changes in metabolic pathways within the cell. Rats fed diets containing large amounts of fat require substantially less thiamine than do those fed diets high in carbohydrate. The utilization of carbohydrate as an energy source (*i.e.,* for ATP formation) is known to involve an important thiamine-dependent step, which is bypassed when fat is used as an energy source, and it is assumed that the lessened requirement for thiamine results from the change in metabolic pathways.

COMPARISON OF NUTRIENT REQUIREMENTS

Trace elements and vitamins are distinguished from other nutrients on the basis of the relatively smaller amounts required. For example, potassium, a mineral element present in comparatively large amounts in all organisms, is required by *Lactobacillus casei* in amounts more than 200 times greater than the requirement for manganese. The magnitude of the requirement for potassium is similar to that for the amino acid leucine. Nutrients such as manganese and the vitamins, which are required in small amounts, are sometimes referred to as micronutrients because of these quantitative relationships. For *Lactobacillus,* glucose is the energy source, and, to permit the amount of growth described, approximately 1,000 times more glucose is required than the nutrients potassium and leucine. These ratios are fairly representative of organoheterotrophs in general. Lactic-acid bacteria, however, form lactic acid from glucose; other organisms, which oxidize this substance to carbon dioxide and water, require only about 5 percent of the amount of glucose required by lactic-acid bacteria for equivalent ATP production and, hence, for equivalent growth.

Micronutrients

SYNTROPHISM

Since the nutritional requirements and metabolic activities of organisms differ, it is clear that two or more different organisms growing relatedly may produce different overall changes in the environment. A rough example is provided by a balanced aquarium, in which aquatic plants utilize light and the waste products of animals—*e.g.,* carbon dioxide, water, ammonia—to synthesize cell materials and generate oxygen, which in turn provide the materials necessary for animal growth. Such relationships are common among microorganisms; *i.e.,* intermediate or end products of metabolism of one organism may provide essential nutrients for another. The mixed populations that result in nature provide examples of this phenomenon, which is called syntrophism; in some instances, the relationship may be so close as to constitute nutritional symbiosis, or mutualism. Several examples of this phenomenon have been found among thiamine-requiring yeasts and fungi, certain of which (group A) synthesized the thiazole component of thiamine molecule but require the pyrimidine portion preformed; for a second group (group B), the relationship is reversed. When group A and group B are grown together in a thiamine-free medium, both types of organisms survive, since each organism synthesizes the growth factor required by its partner; neither organism grows alone under these same conditions. Thus, two or more types of microorganisms frequently grow in situations in which only one species would not.

Influence of populations on each other

Such nutritional interrelationships may explain the fact that the nutritionally demanding lactic-acid bacteria are able to coexist with the nutritionally nondemanding coliform bacteria in the intestinal tracts of animals. It is known that the bacterial flora of the intestinal tract synthesize sufficient amounts of certain vitamins (*e.g.,* vitamin K, folic acid) so that detection of deficiency symptoms in rats requires special measures, and the role of rumen bacteria in ruminant animals (*e.g.,* cows, sheep) in rendering otherwise indigestible cellulose and other materials available to the host animal is well-known. These few examples indicate that syntrophic interrelationships are widespread in nature and may contribute substantially to the nutrition of a wide variety of species. (E.E.Sn./A.S.T.)

Human nutrition and diet

The scope of human nutrition extends far beyond the classical study of the physiological and biochemical processes involved in nourishment; *i.e.,* how substances in food are converted into energy and body tissues. Human nutrition has come to involve all the effects on humans of any component found in food; these include most chronic degenerative diseases (dental decay, coronary heart disease, some cancers, etc.), which are now major targets of research activity. The scope of nutrition extends to the effects of food on human function; *e.g.,* mental function, athletic performance, resistance to infection, and fetal health and development. There is a growing interaction between nutritional science and genetics because of the diversity of human chemical make-up and because food components of which most people are unaware can have marked effects on some individuals. Last, nutrition also considers why people choose to eat the foods they do, even after they have been advised that doing so may be unhealthy. The study of food habits and people's attitudes, beliefs, likes, and dislikes overlaps with the social sciences of physiology, anthropology, sociology, and economics. Dietetics is the application of nutrition in the health sciences.

UTILIZATION OF FOOD BY THE BODY

The human body can be thought of as an engine that releases the energy present in the foods and utilizes it partly for the mechanical work performed by muscles and in secretory processes and partly for the work necessary to maintain its structure and functions. The performance of this work is associated with the production of heat; heat loss is controlled so as to keep body temperature within a narrow range. Unlike other engines, the human body is continually breaking down (catabolism) and building up (anabolism) its component parts. Certain foods supply nutrients essential to the manufacture of the new material and provide energy needed for the chemical reactions involved.

The supply of energy. The energy taken in food and that utilized in daily life can be measured. The measuring unit has been the kilocalorie (kcal), which is 1,000 gram calories; a gram calorie is the amount of heat required to raise one gram of water from 14.5° to 15.5° C at one atmosphere of pressure. Nutritionists are now coming into line with other branches of science and using joules as the unit of energy. One joule is the energy expended when one kilogram is moved one metre by a force of one newton. The amount of energy spent by the whole body and that present in foods are relatively large and are expressed as kilojoules ($kJ = 10^3$ J) or megajoules ($MJ = 10^6$ J). One kilocalorie is equivalent to 4.186 kilojoules.

Measuring energy utilization

The energy present in food can be determined directly by measuring the output of heat when the food is oxidized in a calorimeter. Heats of combustion of individual proteins, fats, and starches are about 5.4, 9.3, and 4.1 kilocalories (23, 39, and 17 kilojoules) per gram, respectively.

Not all of this energy is available to the body because some ingested material is not absorbed from the gut and is lost in the feces; further, the nitrogenous compounds are not completely combusted, and some of the energy in proteins is lost in the urine, mostly as urea. Corrections for these losses give physiological values for dietary protein, fat, and carbohydrate of approximately four, nine, and four kilocalories (17, 38, and 17 kilojoules) per gram, respectively. These are called the Atwater factors, after the American physiologist Wilbur Olin Atwater, who, between 1895 and 1905, calculated the quantitative aspects of energy exchanges. Tables of energy value and nutrient content of common foods provide background data for general dietetic advice (see Table 2).

Proteins, fats, and carbohydrates can, within wide limits, be interchanged as sources of energy. Among the members of prosperous communities most diets provide 12 percent of the energy as protein, about 40 percent as fat, and 48 percent as carbohydrate. In many poor agricultural soci-

Sources of energy

Table 2: The Energy Value and Nutrient Content of Some Common Foods
(values per 100 g edible portion)

	energy		water	carbo-hydrate	protein	fat	alcohol
	kcal	kJ	(g)	(g)	(g)	(g)	(g)
Whole wheat flour	310	1,318	14	63.9	12.7	2.2	—
White bread	235	1,002	37.3	49.3	8.4	1.9	—
White rice, boiled	138	587	68	30.9	2.6	1.3	—
Milk, fresh, whole	66	275	87.8	4.8	3.2	3.9	—
Butter, salted	737	3,031	15.6	trace	0.5	81.7	—
Cheese, Cheddar	412	1,708	36	0.1	25.5	34.4	—
Steak, grilled	218	912	59.3	0	27.3	12.1	—
Tuna, canned in oil, drained	189	794	63.3	0	27.1	9	—
New potatoes, boiled in unsalted water	75	321	80.5	17.8	1.5	0.3	—
Peas, frozen, boiled in unsalted water	69	291	78.3	9.7	6	0.9	—
Cabbage, boiled in salted water	18	75	92.5	2.5	0.8	0.6	—
Orange	37	158	86.1	8.5	1.1	0.1	—
Apple, raw	47	199	84.5	11.8	0.4	0.1	—
White sugar	394	1,680	trace	105	trace	0	—
Beer*, canned	32	132	—	2.3	0.3	trace	3.1
Spirits* (brandy, gin, rum, whiskey)	222	919	—	trace	trace	0	31.7

*Values per 100 ml.
Source: Data from *The Composition of Foods*, 5th ed. (1991) are reproduced with the permission of The Royal Society of Chemistry and the Controller of Her Majesty's Stationery Office.

eties, for whom cereals provide most of the energy, the figures for the individuals' diets are 10 percent for protein, 10 percent for fat, and 80 percent for carbohydrate. Throughout most of the world, protein provides between 8 and 14 percent of the energy ingested. High-fat diets are associated with a high incidence of atherosclerosis (deposits of lipid material in the larger arteries). Wide variations in the proportions of fat and carbohydrate in the diet may, nevertheless, be compatible with good health, at least in the short term. Ethyl alcohol is another source of energy to the body. Only a small part of the alcohol intake, usually less than 5 percent, is excreted in the urine and expired air; most is oxidized in the liver, where it serves as a source of energy with a value of seven kilocalories (29 kilojoules) per gram (almost that of pure fat). Muscle cannot utilize alcohol, and alcohol intake reduces the use of fat and carbohydrate by the liver. In this way alcoholic drinks increase the energy value of a diet and often contribute to obesity and fat deposits in the liver.

Building and maintenance. The body of a young adult male weighing 65 kilograms (143 pounds) consists of some 11 kilograms of protein, nine of fat, one of carbohydrate, four of minerals, and 40 of water. During the first 20 years of life, an average of about 1.5 grams of protein and 150 milligrams of calcium must be retained from the diet every day in order to build soft tissues and skeleton. Tissues, however, are not static, and their components are being continually catabolized and replaced at varying rates. The inner epithelial lining of the gut is replaced every three or four days, and red blood cells have a lifespan of only 120 days. On the other hand, collagen, a protein constituent of tendons, is turned over at intervals of a few years. Much of the material derived from the breakdown of tissues is reutilized, but new material also is required from food. Probably some 200 grams of tissue protein is replenished daily in an adult, but the minimal dietary requirement of protein is only about 40 grams per day.

Turnover of body stores

Essential nutrients. The principal evidence that an organic compound is essential is that it consistently cures a specific deficiency disease. An inorganic element is essential if it regularly occurs in the body and is demonstrated to have a function (such as being an integral part of an enzyme). Humans require oxygen, water, food energy, protein, 14 other organic compounds (vitamins and essential fatty acids), and some 18 inorganic elements, as well as carbon, hydrogen, and nitrogen.

Protein. Growing children need more protein per kilogram of body weight than do adults. Protein requirement at all ages is increased by infections not only because there is an increased utilization of protein but also because illness usually impairs the appetite and thus reduces dietary intake of all substances, including protein. In many countries children are weaned on a diet of cereal paps with little or no supplement of milk or other protein-containing foods. Such a diet at the least retards growth and development. If a child on such a diet suffers from an acute infection, notably measles or gastroenteritis, a severe illness, which may take several forms—known variously as protein-calorie malnutrition, kwashiorkor, or nutritional marasmus—may ensue. The death rate from protein-calorie malnutrition in many poverty-stricken communities is high. (See below *Energy (calories)*, under the heading *Deficiency diseases*.) Primary protein deficiency is not common among adults, for whom cereals in general satisfy the protein requirement.

Protein deficiency

Foods vary in the quantity of proteins they contain per 100 grams (or per typical serving). They also vary in the nutritional value, or quality, of their protein. This is because of differences in the pattern of essential amino acids that they contain (see Table 3). Eight amino acids are essential (*i.e.,* required, but not produced, by the body) for humans: isoleucine, leucine, lysine, methionine (which can be partly replaced by cystine), phenylalanine (which can be partly replaced by tyrosine), threonine, tryptophan, and valine. Histidine appears to be essential as well in some situations. These amino acids are not all required in the same amount each day; for example, leucine is required most, and tryptophan least. Vegetable food proteins usually have lower nutritive values because they have suboptimal proportions of one or more essential amino acids, called the limiting amino acids—*e.g.,* lysine in wheat (see Table 3). When vegetable protein sources are combined, however, the amino acid pattern of the combination is usually improved.

Minerals. Iron is required for the synthesis of hemoglobin, the oxygen-binding pigment in red blood cells. Normally the iron liberated from old cells is retained and can be reutilized. When, however, there is chronic bleeding from wounds or there is severe and prolonged menstruation, the normal amount of dietary iron may be insufficient to replenish the body's supply. Losses of iron in the menses, the needs of a fetus, and the inevitable loss at labour and in the milk of a lactating woman increase the iron requirements of women during their reproductive life.

Iron and calcium

Most dietary iron is in a form that is poorly absorbed from the gut. Many diets contain about 12 milligrams of iron, of which less than 10 percent need be absorbed by a normal adult male. Absorption of iron in plant foods is usually less than 10 percent. The iron in meat is absorbed better, and so is that in human milk. As body stores of iron become depleted, the efficiency of intestinal absorption is enhanced. If menstrual losses are large, iron-deficiency anemia follows unless the diet is supplemented with absorbable iron compounds.

Calcium is the most obvious and persistent of the mineral nutrients, yet it is more difficult to measure the adequacy of its intake than for other nutrients. More than 99 percent of the body's calcium is in the skeleton, where it not only provides structural support but also is a large reservoir for maintaining a constant calcium concentration in plasma. Several hormones, including parathormone, cal-

Table 3: Essential Amino Acid Patterns in Proteins of Some Common Foods*

	human milk	cow's milk	beef muscle	hen's egg	wheat flour	cornmeal (maize)	soy-bean
Histidine	24	27	41	25	25	28	30
Isoleucine	47	53	48	55	39	39	51
Leucine	90	97	86	88	75	128	82
Lysine	60	81	95	72	(25)	(28)	68
Methionine + cystine	36	35	41	59	46	41	33
Phenylalanine + tyrosine	75	84	76	93	84	104	95
Threonine	39	41	49	46	30	40	41
Tryptophan	13	13	11	15	11	(6)	14
Valine	51	64	50	68	44	52	52

*Milligrams of amino acid per one gram of protein. In wheat flour, lysine (ringed) is the limiting amino acid; in cornmeal, lysine and tryptophan (ringed) are both limiting.
Source (except for cornmeal): Japanese Standard Tables of Amino Acid Composition of Foods, 1986, published by Resources Council, Science and Technology Agency.

citonin, and a derivative of vitamin D, are all involved in this regulation. The amount of calcium in the bones is nearly 30 grams at birth and builds up to about 1,200 grams in an adult. Hence, an average of 180 milligrams of calcium must be retained in the body throughout childhood, and the individual daily amounts should reach 400 milligrams during the adolescent growth spurt. Absorption of calcium, like that of many other metallic elements, is inefficient. The diet must therefore supply more calcium than the amount theoretically retained for skeletal growth. Three major questions about calcium are as yet not fully answered by nutritional science: (1) whether generous intakes of calcium during childhood and adolescence will lead to taller adults or heavier bones, (2) whether a generous intake of calcium from about 45 years of age will delay the progress of osteoporosis (thinning of the bones), which occurs in older people and is more likely to cause symptoms in women after menopause, and (3) what mechanisms enable people in Africa, for example, who grow up on cereal diets low in calcium, to retain enough calcium to achieve a skeleton of similar height to those in northern countries where milk and cheese (both rich in calcium) are staples in the diet.

Sodium and potassium

Sodium is present only in small quantities in most natural foods, but salt is added, often in large amounts, in food processing and by cooks to enhance flavour. Sodium is the predominant ion in extracellular fluid; an excess can cause edema (an accumulation of extracellular fluid), especially in conditions such as congestive heart failure. A low sodium intake leads to a lowering of the blood pressure and brings about diuresis, ridding the body of the excess extracellular fluid. There is now much evidence that excess dietary salt may contribute to high blood pressure in some individuals, but other factors are also responsible.

Potassium, present in all natural foods, is the predominant intracellular ion. Deficiency of it does not occur as a result of a primary dietary lack, but it may arise when there is chronic diarrhea or from repeated diuresis. All wasting diseases are associated with loss of potassium from the tissues. Potassium deficiency disturbs the excitability of tissues and leads to paralysis of muscle, including cardiac muscle. Sodium and potassium are two of the most important ions in maintaining the homeostatic equilibrium of the body fluids.

Other elements required by humans are magnesium, iodine, zinc, phosphorus (in the form of phosphate), sulfur (usually in the forms of the amino acids methionine and cystine), and chlorine (in the form of chloride, often combined with sodium). Other elements required in trace amounts are copper, chromium, molybdenum, selenium, cobalt (in the form of vitamin B_{12}), probably fluorine (in the form of fluoride), and possibly vanadium.

Vitamins. For normal nutrition and metabolism the body requires certain organic substances, called vitamins, which it cannot make for itself, at least in sufficient quantities. Many of them function as components of enzyme systems. Because they are required in such small quantities, vitamins do not contribute significantly to the energy needs of the body. In their absence, however, the carbohydrates, fats, and proteins required for energy production and tissue maintenance cannot be properly metabolized. Vitamins were originally identified by letters as each new one was discovered, but this method of nomenclature is being replaced as the vitamins become known by their chemical names (Table 4). For a detailed treatment of deficiency diseases, see below *Deficiency diseases: Vitamins.*

Vitamin deficiency diseases

The evidence that vitamins promote health in any way apart from preventing deficiency diseases is controversial and not established scientifically. The recommended intakes or allowances of vitamins (Table 5) carry a safety margin that allows for a wide range of individual needs, so there are usually no deficiency effects if the intake is only a small percentage below the recommended intake. High intakes are of no benefit and may be dangerous in the cases of the fat-soluble vitamins A and D and the water-soluble vitamin B_6.

CLASSES OF FOODS

Grouping of food

The following set of 17 groups covers most of the nutritional differences between foods: (1) cereals and cereal products, (2) starchy roots, (3) legumes (pulses), (4) other vegetables, (5) fruits, (6) nuts and seeds, (7) sugars, syrups, sweets, and preserves, (8) meat, including poultry, and meat products, (9) seafood—finfish and shellfish, (10) eggs and roe, (11) milk, cream, and cheese, (12) fats and oils, (13) herbs and spices, (14) nonalcoholic, nondairy beverages, (15) alcoholic beverages, (16) dietetic preparations, (17) miscellaneous (*e.g.,* salt and vinegar).

Cereals. Of the present world population, half lives in Asia and three-quarters in developing countries. In these areas cereals play the leading dietary role. On the whole, the carbohydrate-rich cereals compare favourably with the protein-rich foods in energy value; in addition they are significantly less expensive to produce. Therefore most of the world's diets are arranged to meet main calorie re-

Table 4: Nomenclature of Vitamins

name recommended by IUNS*	alternative name	usual pharmaceutical preparation
Vitamin A	Retinol	Retinol
Thiamin	Vitamin B_1	Thiamine hydrochloride
Riboflavin	Vitamin B_2	Riboflavin
Niacin	Nicotinic acid and nicotinamide	Nicotinamide
Vitamin B_6	Pyroxidine	Pyroxidine hydrochloride
Pantothenic acid		Calcium pantothenate
Biotin		Biotin
Folate	Folacin	Folic acid
Vitamin B_{12}	Cobalamin	Cyanocobalamin or hydroxocobalamin
Vitamin C	Ascorbic acid	Ascorbic acid
Vitamin D	Vitamins D_2 and D_3	(Ergo) calciferol
Vitamin E		α-Tocopherol
Vitamin K		Vitamin K_1

*International Union of Nutritional Sciences. Note that in some instances IUNS spelling differs from standard usage.
Source: A. Stewart Truswell *et al., ABC of Nutrition* (1986).

Table 5: Size of Adult Requirements for Essential Nutrients	
adult daily requirement*	in foods
2–10 μg	vitamin D, vitamin B_{12}, vitamin K, chromium
100 μg	biotin, iodine, selenium
200 μg	folic acid, molybdenum
1–2 mg	vitamin A, thiamine, riboflavin, vitamin B_6, fluoride, copper
5–10 mg	pantothenate, manganese
15 mg	niacin, vitamin E, iron, zinc
50 mg	vitamin C
300 mg	magnesium
1 g	calcium, phosphorus (as phosphate)
1–5 g	sodium, chloride, potassium, essential fatty acids
50 g	protein (8–10 essential amino acids)
50–100 g	available carbohydrate
1 kg (litre)	water

*Figures are approximate and in places rounded. The range of size of requirements is about 10⁶.
Source: A. Stewart Truswell *et al., ABC of Nutrition* (1986).

quirements from the cheaper carbohydrate foods, and the protein-rich foods are used sparingly to provide essential amino acids. Cereal products generally contribute a large proportion of the dietary protein of vegetable origin, in addition to their carbohydrate content.

Forms of cereals

Cereals are largely consumed as bread and to a lesser extent as cakes, crackers, and breakfast cereals. They are also consumed in other forms in some countries; *e.g.*, boiled, crushed, or rolled, made into pasta, etc. There has been controversy as to the relative merits of white bread and bread made from whole wheat flour of much longer extraction, which is darker in colour. White flour consists of about 72 percent of the grain but contains little of the germ, or embryo, and of the outer coverings (bran). Since the B vitamins are concentrated mainly in the scutellum (covering of the germ), and to a lesser extent in the bran, the vitamin B content of white flour, unless artificially enriched, is much less than that of brown flour. The dietary fibre is mostly in the bran, and white flour contains only about one-third of that in whole wheat flour. White flour is routinely fortified with synthetic vitamins in a number of countries, including the United States and United Kingdom, so that the vitamin content is similar to that of the darker flours. In the United Kingdom white bread is also fortified with calcium carbonate.

Nutritional value

Table 6 shows the contribution of cereals to the intakes of a number of key nutrients in the United Kingdom. As a group they provide almost half the carbohydrate, one-quarter of the protein and calcium, more than one-third of the iron, and nearly half the thiamine. The protein of cereals does not have as high a nutritive value as that of many animal foods because of the low lysine content. It is possible to fortify bread with synthetic lysine. Corn has been altered to produce a higher lysine and tryptophan content than standard corn (see Table 3), but the yield is lower. The usual solution to the suboptimal amino acid pattern of cereals is to eat them mixed with other foods (legumes, animal foods) that have more than enough lysine. This is the traditional, intuitive way in which cereals have been eaten in settled communities in many different cultures. The amino acids of mixed protein foods tend to complement one another and increase the percentage of the limiting amino acid. In less developed countries, the contribution of cereals to the intake of nutrients is greater than in the United Kingdom, which is an example of an industrial country in which the cereal consumed by humans is predominantly wheat.

Wheat and rice are the two major cereals for human consumption; wheat in Europe, North America, and northern India, and rice in East and Southeast Asia. The nutritional weakness of rice is its low thiamine and riboflavin content when it is milled to polished rice. Beriberi was formerly common in poor Asian communities in which a large proportion of the diet consisted of polished rice. Beriberi has almost completely disappeared from Asia with the advent of greater availability of other foods, of better ways of handling the rice (*e.g.*, parboiling), and, in some areas, of fortification with thiamine.

Corn differs from other cereals in that the yellow form contains carotenoids with provitamin A activity. Its other special feature is its low content of tryptophan. The niacin in corn is in a bound form that cannot be digested and absorbed by humans unless pretreated with lime (calcium hydroxide) or by roasting. Niacin is also formed in the body as a metabolite of the amino acid tryptophan, but this alternative source is not available when the tryptophan content is too low.

Niacin

Most cereals contain little fat, although oats contain 7 percent. Wheat contains a type of dietary fibre (arabinoxylan hemicellulose) that is particularly valuable for treatment and prevention of constipation. It increases the bulk of the feces and shortens the mouth-to-anus transit time. Bran is a concentrated source of the same dietary fibre. (D.W.K.-J./A.S.T.)

Starchy roots. Potatoes, cassava, and yams are easily cultivated and are valuable as cheap sources of energy. Their nutritive value, in general, resembles that of cereals, but their protein content is lower. Protein deficiency may be common in tropical communities in which the staple food is cassava or yams. The potato, however, provides some protein—less than cereal does—but also contains some vitamin C, and sweet potatoes contain the pigment beta-carotene, convertible in the body into vitamin A. Although the quantity of protein in potatoes is only small (2 percent), its nutritive value appears to be high. Cassava contains half the protein of potatoes. There are toxic substances in, or adjacent to, some of the starchy root foods that have to be avoided or dealt with by careful preparation; *e.g.*, solanine in the green sprouts of potatoes and a cyanide compound in the leaves and outer parts of the roots of cassava.

Legumes (pulses). Peas and beans, the seeds of Fabaceae, resemble the cereals in nutritive value but have a higher protein content and, since they are not subject to milling, are a good source of B vitamins. They are thus a valuable supplement to a cereal diet, especially in tropical or subtropical countries; moreover, they, particularly the soybean, are also valued for their taste. The soybean is especially rich in protein (38 percent) and is a major commercial source of edible oil.

Table 6: Percentage of Several Major Nutrients Supplied by Food Groups in the United Kingdom*

	energy	protein	fat	carbo-hydrate	iron	cal-cium	thi-amine	vitamin C	vitamin A
Milk, cream, and cheese	14	22	19	7	3	59	12	7	13.5
Meat	16	32	27	2	24	3	15	2	37
Fish	1	4.5	1	0.3	2	1	1	—	—
Eggs	2	4.5	3	—	5	1.5	2	—	3
Butter, margarine, and other fats	15.5	—	37	—	1	—	—	—	20
Sugar, syrups, preserves	9	—	—	20	1	—	—	—	—
Total vegetables	9	9.5	2	15	19	6	19	48	24
Total fruits	3	1	—	5	4	2	4	41	1
Total cereals (Breads)	29 (13.5)	24 (14)	10 (2)	48 (24)	37 (17)	24 (12)	42 (21)	— (—)	1 (—)

*From the Household Food Consumption Survey (1981). The figures for breads are in parentheses because they are part of the total for cereals.

Vegetables and fruits. Vegetables and fruits all have similar nutritive properties. Because 70 percent or more of their weight is water, they provide comparatively little energy or protein, but many contain vitamin C and carotene, two nutrients not found in cereals. Fresh fruits, particularly the citrus variety, and their juices are usually rich in vitamin C, but cooked vegetables are an uncertain source, as this vitamin is easily destroyed. Vegetables and fruits contain fibre, which adds bulk to the intestinal content and is useful in preventing constipation. Vegetables also provide calcium and iron but often in a form that is poorly absorbed.

Sugars, preserves, and syrups. In adults in the United States, the average intake of added sugars is 60 grams per day in males, or one pound per week. The sugars, mostly sucrose and high-fructose corn syrup, are added both in processing and at the table. Together sucrose and fructose provide 12 percent of the average total calories in adults and a little more in children. There are also naturally occurring sugars in foods (fructose, glucose, sucrose in fruits and some vegetables, and lactose in milk). The intake of these in the United States is about 8 percent of total caloric intake in adults and much more in young children (due to the lactose). Sugar, however, contains no protein, no minerals, and no vitamins and thus has been called a source of "empty calories."

Advantages and disadvantages

Sugar is an excellent preservative because it adsorbs water and prevents the growth of microorganisms. Jams contain from 30 to 60 percent sugar, and honey and natural syrups (*e.g.,* maple) are composed of more than 75 percent sugar.

Meat, fish, and eggs. Meats generally consist of about 20 percent protein, 20 percent fat, and 60 percent water; the amount of fat present in a particular portion of meat varies greatly, not only with the kind of meat—pork, beef, lamb, etc.—but also with its quality; the "energy value" varies in direct proportion with the fat content. Meat is valuable for its protein, which is of high biological value. Meat also is an excellent source of B vitamins, including thiamine, one of the most important. Pork is the best source of thiamine, liver is next, and skeletal muscle, from any meat source, is third. Meat is also a good source of niacin, vitamin B_6, vitamin B_{12}, other vitamins of the B group, and the mineral nutrients iron, zinc, phosphorus, potassium, and magnesium. Liver is the storage organ for, and very rich in, vitamins A and B_{12}. It is also an excellent source of riboflavin and of folic acid.

Fish

The muscular tissue of fishes consists of 13–20 percent protein; varying amounts of fat, ranging from less than 1 to more than 20 percent; and 60–82 percent water, varying inversely with fat content. Exact proportions vary among different species, and seasonal and feeding variations result in differences among individuals of the same species. The fatty acids in fish are largely polyunsaturated. One of the major fatty acids, eicosapentaenoic acid, in large amounts reduces the tendency to thrombosis.

Eggs have a deservedly high reputation as a food. The white is protein and the yolk is rich in both protein and vitamin A. Eggs also provide calcium and iron; egg yolk, however, has a high cholesterol content.

Milk and milk products. The milk of each species of animal is a complete food for its young. One pint of cow's summer milk contributes about 90 percent of the calcium, 30 to 40 percent of the riboflavin, 25 to 30 percent of the protein, 10 to 20 percent of the calories and vitamins A and B, and up to 10 percent of the iron and vitamin D needed by an adult. Although human milk supplies a higher proportion of the daily needs of a five-year-old child for calories, protein, vitamin A, thiamine, and riboflavin, the contribution of calcium is reduced to about 70 percent because of the higher calcium requirements of a child.

Human versus cow's milk

The milks of mammals contain all essential nutrients. Human milk (Table 7) is the perfect food for infants, provided it comes from a healthy, well-nourished mother and the infant is full-term. Breast milk contains important antibodies, white cells, and nutrients, and, in communities where hygiene is poor, breast-fed babies have fewer infections than formula-fed babies.

Cow's milk (Table 7) is good food for human adults, but the cream (*i.e.,* fat) contains 52 percent saturated fatty acids (longer than 10 carbon atoms in length) as compared to only 3 percent polyunsaturated fat. This fat raises the plasma cholesterol and is thought to be one of the dietary components that contributes to coronary heart disease, along with the same fat in concentrated forms (cream and butter). To circumvent this, the dairy industry has developed low-fat cow's milk (with 2 percent instead of almost 4 percent fat), very low-fat skim milk, or skim milk with extra nonfat milk solids (lactose, protein, and calcium), which gives more body to the milk.

Table 7: Some Nutrients in Milks

	human* milk	whole cow's milk (full cream)	modified infant milk formula† (powder diluted as directed)
Energy (kilocalories)	70	67	69
Protein (total; g)	1.1	3.5	1.5
Casein (% protein)	40%	80%	40%
Lactose (g)	7.3	5.0	7.2
Fat (total; g)	4.0	3.7	3.6
Saturated fat (% fat)	46%	66%	44%
Linoleic acid (% fat)	7–11%	3%	17%
Sodium (millimoles)	0.7	2.2	0.71
Calcium (mg)	35	120	49
Phosphorus (mg)	15	95	30
Iron (mg)	0.075	0.050	0.9
Vitamin C (mg)	3.8	1.5	6.9
Vitamin D (μg)	0.8	0.15	1.1

*The composition of breast milk varies considerably with stage of lactation, between individuals, and with maternal nutrition.
†Mean of Cow and Gate Premium and SMA-S26 infant formula brands.
Source: A. Stewart Truswell *et al., ABC of Nutrition* (1986).

Lactose

Lactose, the characteristic sugar of milk, is a disaccharide made of the monosaccharides glucose and galactose. While these monosaccharides are easily absorbed, lactose is not. Lactose passes to the large intestine, where it is fermented by the resident bacteria to produce gas and sometimes diarrhea. Galactose is incorporated into the complex fatty substance ceramide, which is the major component of the myelin sheaths of nerves (the white matter of the central nervous system).

In the late 1960s it was discovered that the adults of many ethnic groups cannot break down the lactose of large quantities of milk into galactose and glucose; they lose most of their intestinal lactase enzyme activity as they grow up (lactose intolerance). It is now recognized that such individuals share this deficiency with adults throughout the world, such as in Asia and Africa, and with most of the animal kingdom. People originating in northern Europe are the exception from the global viewpoint: they usually retain full intestinal lactase activity into adult life. People who have little of the enzyme lactase in their bodies can still take large amounts of milk if it has been allowed to go sour, if lactobacilli have split most of the lactose into lactic acid, or if the lactose has been treated with commercially available lactase.

The vitamin C present in milk is destroyed by heating (pasteurization), which in many countries is required to prevent the milk from spreading bacterial and other infections.

Cheese making

Cheese making is an ancient art formerly used on farms to convert surplus milk into a food that could be stored without refrigeration. Cheese is rich in protein and calcium and is a good source of vitamin A and riboflavin. Most cheeses, however, contain about 25 to 30 percent fat (constituting about 70 percent of the calories of the cheese), which is mostly saturated, and they are usually high in salt (sodium).

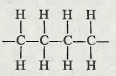
Fats and oils. The animal fats used by humans are butter, suet from beef, lard from pork, and fish oils. Important vegetable oils include olive oil, peanut (groundnut) oil, coconut oil, cottonseed oil, sunflower seed oil, soybean oil, safflower oil, rape oil, sesame (gingelly) oil, mustard oil, red palm oil, and corn oil. All these are high in calories. Only butter (other than the previously mentioned fish-liver oils) contains any of the vitamins A and D, but red palm oil does contain carotene, which is converted to vitamin A in the body. Vitamins A and D in controlled amounts are added to margarines.

All natural fats and oils contain variable amounts of vitamin E, the fat-soluble vitamin antioxidant. They provide more calories per gram than any other food, and they contain no protein and few micronutrients.

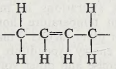
Fatty acids

The predominant substances in fats and oils are triglycerides, chemical compounds containing any three fatty acids combined with a molecule of glycerol. The fatty acids consist of a chain of carbon atoms with a carboxylic acid group (—COOH) at one end. The number of carbon atoms ranges from four to more than 22, but the most common chain length is 16 or 18. Because they are synthesized in the body from two-carbon units (acetyl coenzyme A), chain lengths are nearly always even numbers. Each carbon has four valences, or potential sites for bonding—two valences for the adjacent carbons and one for each hydrogen.

Butyric acid is an example of a saturated fatty acid. The four valences of each carbon atom are attached to two hydrogen atoms and to the two adjacent carbons:



When adjacent carbon atoms are linked by a double bond, each has only one valence available for hydrogen:



When no double bonds are present, the fatty acid is said to be saturated; with the presence of one (monounsaturated) or more (polyunsaturated) double bonds, the fatty acid is said to be unsaturated.

Fats with a high percentage of saturated fatty acids tend to be solid at room temperature; *e.g.,* butter and lard. Those with a high percentage of unsaturated fatty acids are usually liquid oils; *e.g.,* sunflower, safflower, and corn oils. The names and shorthand numbers for the principal fatty acids in foods are shown in Table 8. In this shorthand notation, the number to the left of the colon is the number of carbon atoms, and the number to the right of the colon represents the number of double bonds; *e.g.,* 4:0 has four carbon atoms and no double bonds (*i.e.,* is saturated). This is the shorthand notation for butyric acid.

Essential fatty acids

A small group of fatty acids are essential in the diet. They occur in body structures, especially the different membranes inside and around cells, and cannot be synthesized in the body from other fats. Linoleic acid (18:2) is the most important of these fatty acids because it is convertible to the other essential fatty acids. Linoleic acid has two double bonds and is a polyunsaturated fatty acid. As well as being an essential fatty acid it tends to lower the plasma cholesterol. Linoleic acid occurs in moderate to high proportions in many of the seed oils; *e.g.,* corn, sunflower, cottonseed, and safflower oils. Some margarines (polyunsaturated margarines) use a blend of oils selected so that they have a moderately high linoleic acid content.

Edible fats and oils contain smaller amounts of other lipids as well as triglycerides (see below *Habitual dietary patterns and disease: Diet and coronary heart disease*).

Table 8: Important Natural Fatty Acids

acids	shorthand
Saturated acids	
Butyric	4:0
Caproic	6:0
Caprylic	8:0
Capric	10:0
Lauric	12:0
Myristic	14:0
Palmitic	16:0
Stearic	18:0
Arachidic	20:0
Behenic	22.0
Monounsaturated acids	
Palmitoleic	16:1
Oleic	18:1
Erucic	22:1
Polyunsaturated acids	
Linoleic	18:2
Linolenic	18:3
Arachnidonic	20.4
Eicosapentaenoic	20.5
Docosahexaenoic	22:6

Beverages. Although most adults drink one to two litres (about one to two quarts) of water a day, much of this is in the form of liquids such as coffee, tea, fruit juices, soft drinks, beer, wines, or spirits. In general, these are appreciated more for their taste or for their effects than for their nutritive value. Fruit juices are, of course, useful for their vitamin C content and good sources of potassium while low in sodium. Coffee and tea by themselves are of no nutritive value, except that coffee contains some niacin and tea contains fluoride and manganese; these beverages also contain natural caffeine, which has a stimulating effect. Caffeine is added to colas, and so-called diet soft drinks contain small quantities of artificial sweeteners in place of sugars so that their overall caloric value is reduced. Beer contains 2 to 6 percent alcohol, natural wines 10–13 percent, and most spirits up to 40 percent. Since ethyl alcohol has an energy value of seven kilocalories per gram, very significant amounts of energy can be obtained from alcoholic drinks, and beer and wine contain natural sugars as well. With one or two exceptions, they contain no nutrients and are only a source of "empty calories." The only vitamin present in significant amounts in beer from a brewery is riboflavin. Wines are devoid of vitamins but sometimes contain large amounts of iron, probably acquired from iron vessels used in preparation, especially of cheap wine. It is possible for excess iron to be absorbed and stored in the liver where it may contribute to toxic manifestations. (A.J.T.)

Alcohol

NUTRIENT RECOMMENDATIONS

Notions of what constitutes a healthful diet vary with geography and custom as well as with changing times and an evolving understanding of nutrition. With industrialization and globalization, food can now be transported over long distances, but in the past people had to live almost entirely on what was locally produced. Researchers must be careful in generalizing a national diet from a relatively small sample of the population; the poor cannot afford to eat the same diet as the rich, and many countries have large immigrant groups with their own distinctive food patterns. Even within a culture, some people abstain from certain foods on moral or religious grounds.

Regional differences in diets

In general, persons living in more affluent countries eat more meat and other animal products. On the other hand, the diets of those living in poorer, agricultural countries rely primarily on cereals in the form of wheat flour, white rice, or corn, with animal products providing less than 10 percent of the energy. Another big difference between cultures is the extent to which dairy products are consumed. The Chinese, for example, obtain less than 1 percent of their energy from dairy products. In contrast, in Pakistan dairy products contribute almost 10 percent of energy. Among Western diets, the lowest in saturated fat is the so-called Mediterranean diet. In the 1950s it was realized that Europeans living in rural areas near the Mediterranean Sea had a greater life expectancy than those living elsewhere in Europe, despite poor medical services and a lower standard of living. The traditional diet of Mediterranean peoples is low in animal products; instead, olive oil is a major source of monounsaturated fat. Also, tomatoes and green leafy vegetables are regularly consumed in large quantities, and these contain a variety of antioxidant compounds that are now thought to be healthful. (K.J.C.)

Setting DRIs and other dietary goals. Following the publication of dietary goals for the Nordic countries in 1968 and for the United States in 1977, dietary goals and guidelines have been set forth by a number of countries and revised periodically as a way of translating scientific nutrition recommendations into simple and practical dietary suggestions. These authoritative statements—some published by scientific bodies and some by government

agencies—aim to promote long-term health and to prevent or reduce the chances of developing chronic and degenerative diseases. Although the guidelines of different countries may vary in important ways, most recent dietary recommendations include variations on the following fundamental themes: eat a variety of foods; perform regular physical activity and maintain a healthy weight; limit consumption of fat, saturated fat, cholesterol, sugar, salt, and alcohol; and eat plenty of grain products (especially whole grains), vegetables, and fruits.

Nutrient recommendations are usually determined by scientific bodies within a country at the behest of government agencies. The World Health Organization and other agencies of the United Nations have also issued reports on nutrients and food components.

RDAs

The Recommended Dietary Allowances (RDAs), first published by the U.S. National Academy of Sciences in 1941 and revised every few years until 1989, established dietary standards for evaluating nutritional intakes of populations and for planning food supplies. The RDAs reflected the best scientific judgment of the time in setting amounts for different nutrients that would be adequate to meet the nutritional needs of most healthy people.

During the 1990s a paradigm shift took place as scientists from the United States and Canada joined forces in an ambitious multiyear project to reframe dietary standards for the two countries. In the new approach, known as the Dietary Reference Intakes (DRIs), classical indicators of deficiency, such as scurvy and beriberi, were considered an insufficient basis for recommendations. Where warranted by a sufficient research base, indicators with broader significance are now used, those that might reflect a decreased risk of chronic diseases such as osteoporosis, heart disease, hypertension, diabetes mellitus, or cancer. DRIs are intended to help individuals plan a healthful diet as well as avoid consuming too much of a nutrient. The new and comprehensive approach of the DRIs is serving as a model for other countries. DRI reports have been published every year or two, starting in 1997. The expectation is that DRIs will eventually be released for all established nutrients, and for some food components such as flavonoids that are not considered nutrients but have an impact on health.

The collective term Dietary Reference Intakes encompasses four categories of reference values. The Estimated Average Requirement (EAR) is the intake level for a nutrient at which the needs of 50 percent of the population will be met. Because the needs of the other half of the population will not be met by this amount, the EAR is increased by about 20 percent to arrive at the RDA. The RDA is the average daily dietary intake level that is sufficient to meet the nutrient requirement of nearly all (97 to 98 percent) healthy people in a particular life stage. When the EAR, and thus the RDA, cannot be set due to insufficient scientific evidence, another parameter, the Adequate Intake (AI), is given, based on estimates of intake levels of healthy populations. Lastly, the Tolerable Upper Intake Level (UL) is the highest level of a daily nutrient intake that will most likely present no risk of adverse health effects in almost all individuals in the general population.

Nutrition information is commonly displayed on food labels, but this information is generally simplified to avoid confusion. Because only one nutrient reference value is listed, and because sex and age categories usually are not usually taken into consideration, the amount chosen is generally the highest RDA value. In the United States, for example, the Daily Values, determined by the Food and Drug Administration, are generally based on RDA values published in 1968. Different food components are listed on the food label as a percentage of their Daily Values.

Confidence that a desirable level of intake is reasonable for a particular group of people can be bolstered by multiple lines of evidence pointing in the same direction, an understanding of the function of the nutrient and how it is handled by the body, and a comprehensive theoretical model with strong statistical underpinnings. Of critical importance in estimating nutrient requirements is explicitly defining exactly the criterion that the specified level of intake is intended to satisfy. Approaches that use different definitions of adequacy are not comparable. For example, it is one thing to prevent clinical impairment of bodily function (basal requirement), which does not necessarily require any reserves of the nutrient, but it is another to consider an amount that will provide desirable reserves (normative requirement) in the body. Yet another approach attempts to evaluate a nutrient intake conducive to optimal health, even if an amount is required beyond that normally obtainable in food—possibly necessitating the use of supplements. Furthermore, determining upper levels of safe intake requires evidence of a different sort. These issues are extremely complex, and the scientists who collaborate to set nutrient recommendations face exceptional challenges in their attempts to reach consensus.

Energy. Energy is needed not only when a person is physically active but even when the body is lying motionless. Depending on an individual's level of physical activity, between 50 and 80 percent of the energy expended each day is devoted to basic metabolic processes (basal metabolism), which enable the body to stay warm, breathe, pump blood, and conduct numerous physiological and biosynthetic activities, including synthesis of new tissue in growing children and in pregnant and lactating women. Digestion and subsequent processing of food by the body also uses energy and produces heat. This phenomenon, known as the thermic effect of food (or diet-induced thermogenesis), accounts for about 10 percent of daily energy expenditure, varying somewhat with the composition of the diet and prior dietary practices. Adaptive thermogenesis, another small but important component of energy expenditure, reflects alterations in metabolism due to changes in ambient temperature, hormone production, emotional stress, or other factors. Finally, the most variable component in energy expenditure is physical activity, which includes exercise and other voluntary activities as well as involuntary activities such as fidgeting, shivering, and maintaining posture. Physical activity accounts for 20 to 40 percent of the total energy expenditure, even less in a very sedentary person and more in someone who is extremely active.

Basal metabolism

Basal or resting energy expenditure is correlated primarily with lean body mass (fat-free mass and essential fat, excluding storage fat), which is the metabolically active tissue in the body. At rest, organs such as the liver, brain, heart, and kidney have the highest metabolic activity and, therefore, the highest need for energy, while muscle and bone require less energy, and body fat even less. Besides body composition, other factors affecting basal metabolism include age, sex, body temperature, and thyroid hormone levels.

The basal metabolic rate (BMR), a precisely defined measure of the energy expenditure necessary to support life, is determined under controlled and standardized conditions—shortly after awakening in the morning, at least 12 hours after the last meal, and with a comfortable room temperature. Because of practical considerations, the BMR is rarely measured; the resting energy expenditure (REE) is determined under less-stringent conditions, with the individual resting comfortably about 2 to 4 hours after a meal. In practice, the BMR and REE differ by no more than 10 percent—the REE is usually slightly higher—and the terms are used interchangeably.

Energy expenditure can be assessed by direct calorimetry, which employs apparatuses, such as water-cooled garments or insulated chambers large enough to accommodate a person, to measure heat dissipated from the body. However, it is usually measured by the less-cumbersome techniques of indirect calorimetry, in which heat produced by the body is calculated from measurements of oxygen inhaled, carbon dioxide exhaled, and urinary nitrogen excreted. BMR (in kilocalories per day) can be roughly estimated based on the following formula: $BMR = 70 \times (\text{body weight in kilograms})^{3/4}$.

The energy costs of various activities have been measured. While resting may require as little as 1 kilocalorie per minute, strenuous work may demand 10 times that much. Mental activity, though it may seem taxing, has no appreciable effect on energy requirement. A 70-kilogram (154-pound) man, whose REE over the course of a day might be 1,750 kilocalories, could expend a total of 2,400 kilocalo-

Average daily energy expenditure

ries on a very sedentary day and up to 4,000 kilocalories on a very active day. A 55-kilogram (121-pound) woman, whose daily resting energy expenditure might be 1,350 kilocalories, could use from 1,850 to more than 3,000 total kilocalories, depending on level of activity.

The law of conservation of energy applies: If one takes in more energy than is expended, over time one will gain weight; insufficient energy intake results in weight loss, as the body taps its energy stores to provide for immediate needs. Excess food energy is stored in small amounts as glycogen, a short-term storage form of carbohydrate in muscle and liver, and as fat, the body's main energy reserve found in adipose tissue. Adipose tissue is mostly fat (about 87 percent), but it also contains some protein and water. In order to lose 454 grams (just over one pound) of adipose tissue, an energy deficit of about 3,500 kilocalories (14.6 megajoules) is required.

Protein. Proteins, like carbohydrates and fats, contain carbon, hydrogen, and oxygen, but they also contain nitrogen, a component of the amino chemical group (NH_2), and in some cases sulfur. Proteins serve as the basic structural material of the body, as well as being biochemical catalysts and regulators of genes. Aside from water, protein constitutes the major part of muscles, bones, internal organs, and the skin, nails, and hair. Protein is also an important part of cell membranes and blood (*e.g.,* hemoglobin). Enzymes, which catalyze chemical reactions in the body, are also protein, as are antibodies, collagen in connective tissue, and many hormones, such as insulin.

Tissues throughout the body require ongoing repair and replacement, and thus the body's protein is turning over constantly, being broken down and then resynthesized as needed. Tissue proteins are in a dynamic equilibrium with proteins in the blood, with input from proteins in the diet and losses through urine, feces, and skin (see Figure 1). In a healthy adult, adjustments are made so that the amount of protein lost is in balance with the amount of protein ingested. However, during periods of rapid growth, pregnancy and lactation, or recuperation after illness or depletion, the body is in positive nitrogen balance, as more protein is being retained than excreted. The opposite is true, and there is negative nitrogen balance, during illness or wasting, when more tissue is being broken down than synthesized.

The proteins in food—such as albumin in egg white, casein in dairy products, and gluten in wheat—are broken down during digestion into constituent amino acids, which, when absorbed, contribute to the body's metabolic pool. Amino acids are then joined via peptide linkages to assemble specific proteins, as directed by the genetic material and in response to the body's needs at the time. Each gene makes one or more proteins, each with a unique sequence of amino acids and precise three-dimensional configuration. Amino acids are also required for the synthesis of other important nonprotein compounds, such as peptide hormones, some neurotransmitters, and creatine.

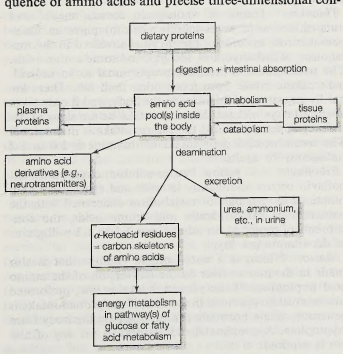


Figure 1. General scheme of protein and amino acid metabolism

Food contains approximately 20 common amino acids, 9 of which are considered essential, or indispensable, for humans, *i.e.,* they cannot be synthesized by the body or cannot be synthesized in sufficient quantities and therefore must be taken in the diet. The essential amino acids for humans are histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine. Conditionally indispensable amino acids include arginine, cysteine, and tyrosine, which may need to be provided under special circumstances, such as in premature infants or people with liver disease, because of impaired conversion from precursors.

Essential amino acids

The relative proportions of different amino acids vary from food to food (see Table 3). Foods of animal origin—meat, fish, eggs, and dairy products—are sources of good-quality, or complete, protein; *i.e.,* their essential amino acid patterns are similar to human needs for protein. (Gelatin, which lacks the amino acid tryptophan, is an exception.) Individual foods of plant origin, with the exception of soybeans, are lower quality, or incomplete, protein sources. Lysine, methionine, and tryptophan are the primary limiting amino acids; *i.e.,* they are in smallest supply and therefore limit the amount of protein that can be synthesized. However, a varied vegetarian diet can readily fulfill human protein requirements if the protein-containing foods are balanced such that their essential amino acids complement each other. For example, legumes such as beans are high in lysine and low in methionine, while grains have complementary strengths and weaknesses. Thus, if beans and rice are eaten over the course of a day, their joint amino acid patterns will supplement each other and provide a higher quality protein than would either food alone. However, careful balancing of plant proteins is necessary only for those whose protein intake is marginal or inadequate. In affluent populations, where protein intake is greatly in excess of needs, obtaining sufficient good-quality protein is usually of concern only for young children who are not provided with animal proteins.

The World Health Organization recommends a daily intake of 0.75 gram of good-quality protein per kilogram of body weight for adults of both sexes. Thus, a 70-kilogram (154-pound) man would need 52.5 grams of protein, and a 55-kilogram (121-pound) woman would need about 41 grams of protein. This recommendation, based on nitrogen balance studies, assumes an adequate energy intake. Infants, children, and pregnant and lactating women have additional protein needs to support synthesis of new tissue or milk production. Protein requirements of endurance athletes and bodybuilders may be slightly higher than those of sedentary individuals, but this has no practical significance because athletes typically consume much more protein than they need.

Protein consumed in excess of the body's needs is degraded; the nitrogen is excreted as urea, and the remaining keto acids are used for energy, providing four kilocalories per gram, or are converted to carbohydrate or fat. During conditions of fasting, starvation, or insufficient dietary intake of protein, lean tissue is broken down to supply amino acids for vital body functions. Persistent protein inadequacy results in suboptimal metabolic function with increased risk of infection and disease.

Vitamins. Vitamins are organic compounds found in very small amounts in food and required for normal functioning—indeed, for survival. Humans are able to synthesize certain vitamins to some extent. For example, vitamin D is produced when the skin is exposed to sunlight; niacin can be synthesized from the amino acid tryptophan; and vitamin K and biotin are synthesized by bacteria living in the gut. However, in general, humans depend on their diet to supply vitamins. When a vitamin is in short supply or is not able to be utilized properly, a specific deficiency syndrome results. When the deficient vitamin is resupplied before irreversible damage occurs, the signs and symptoms are reversed. The amounts of vitamins in foods and the amounts required on a daily basis are measured in milligrams and micrograms (see Table 5).

Vitamins assist in energy-yielding reactions and facilitate metabolic and physiologic processes throughout the body. Vitamin A, for example, is required for embryonic development, growth, reproduction, proper immune function, and the integrity of epithelial cells, in addition to its role in vision. The B vitamins function as coenzymes that assist in energy metabolism; folic acid (folate), one of the B vitamins, helps protect against birth defects in the early stages of pregnancy. Vitamin C plays a role in building connective tissue as well as being an antioxidant that helps protect against damage by reactive molecules (free radicals). Vitamin D is involved in calcium and phosphorus homeostasis and bone metabolism. Vitamin E, another antioxidant, protects against free-radical damage in lipid systems, and vitamin K plays a key role in blood clotting. Although vitamins are often discussed individually, many of their functions are interrelated, and a deficiency of one can influence the function of another.

Fat-soluble and water-soluble vitamins

The 13 vitamins known to be required by human beings are categorized into two groups according to their solubility. The four fat-soluble vitamins (soluble in nonpolar solvents) are vitamins A, D, E, and K. Although now known to behave as a hormone, the activated form of vitamin D, vitamin D hormone (calcitriol), is still grouped with the vitamins as well. The nine water-soluble vitamins (soluble in polar solvents) are vitamin C and the eight B-complex vitamins: thiamine, riboflavin, niacin, vitamin B_6, folic acid, vitamin B_{12}, pantothenic acid, and biotin. Choline is a vitamin-like dietary component that is clearly required for normal metabolism but that can be synthesized by the body. Although choline may be necessary in the diet of premature infants and possibly of those with certain medical conditions, it has not been established as essential in the human diet throughout life.

Different vitamins are more or less susceptible to destruction by environmental conditions and chemical agents. For example, thiamine is especially vulnerable to prolonged heating, riboflavin to ultraviolet or fluorescent light, and vitamin C to oxidation (as when a piece of fruit is cut open and the vitamin is exposed to air). In general, water-soluble vitamins are more easily destroyed during cooking than are fat-soluble vitamins.

The solubility of a vitamin influences the way it is absorbed, transported, stored, and excreted by the body, as well as where it is found in foods. With the exception of vitamin B_{12}, which is supplied only by foods of animal origin, the water-soluble vitamins are synthesized by plants and found in both plant and animal foods. Strict vegetarians (vegans), who eat no foods of animal origin, are therefore at risk of vitamin B_{12} deficiency. Fat-soluble vitamins, on the other hand, are found in association with fats and oils in foods and in the body and typically require protein carriers for transport through the water-filled compartments of the body.

Water-soluble vitamins are not appreciably stored in the body (except for vitamin B_{12}) and thus must be consumed regularly in the diet. If taken in excess, they are readily excreted in the urine, although there is potential toxicity even with water-soluble vitamins; especially noteworthy in this regard is vitamin B_6. Because fat-soluble vitamins are stored in the liver and fatty tissue, they do not necessarily have to be taken in daily, so long as average intakes over time—weeks, months, or even years—meet the body's needs. However, the fact that these vitamins can be stored increases the possibility of toxicity if very large doses are taken. This is particularly of concern with vitamins A and D, which can be toxic if taken in excess. Under certain circumstances, pharmacological ("megadose") levels of some vitamins—many times higher than the amount typically found in food—have accepted medical uses. Niacin, for example, is used to lower blood cholesterol levels; vitamin D is used to treat psoriasis; and pharmacological derivatives of vitamin A are used to treat acne and other skin conditions, as well as to diminish skin wrinkling. However, consumption of vitamins or other dietary supplements in amounts significantly in excess of recommended levels is not advised without medical supervision.

Vitamins synthesized in the laboratory are the same molecules as those extracted from food, and they cannot be distinguished by the body. However, various forms of a vitamin are not necessarily equivalent. In the particular case of vitamin E, supplements labeled *d*-α-tocopherol (or "natural") generally contain more vitamin E activity than those labeled *dl*-α-tocopherol. Vitamins in food have a distinct advantage over vitamins in supplement form because they come associated with other substances that may be beneficial, and there is also less potential for toxicity. Nutritional supplements cannot substitute for a healthful diet. (J.W.)

Vitamin A. A fat-soluble compound, vitamin A (retinol) is found in animal foods, especially in the liver of land animals or fish, where it is stored and concentrated. Most people, however, derive most or all of their vitamin A from plant foods, many of which contain the yellow-orange pigment beta-, or β-, carotene. One molecule of β-carotene can be cleaved by an intestinal enzyme into two molecules of vitamin A. Beta-carotene occurs in fruits such as peaches, melons, and mangoes. It is also a companion of chorophyll, the green pigment in leaves, so that green vegetables are good sources. Enough vitamin A is stored in the liver of a well-nourished adult to last about two years of deprivation. Vitamin A is required by humans in very small amounts; the recommended intakes for adult men and women are 900 and 700 micrograms (μg), respectively, with an extra allowance for women who are pregnant or lactating. It is assumed that 6 μg of β-carotene has the activity of 1 μg of retinol.

Beta-carotene

Vitamin D. Vitamin D is fat-soluble, and cholecalciferol is the major active substance. Its main function is to stimulate the absorption of calcium. The natural source of vitamin D is from the action of short-wave ultraviolet light from the Sun on a derivative of cholesterol in the skin. Vitamin D also occurs in some foods. It is only when people live in high latitudes, consistently wear concealing clothes, and spend nearly all their time indoors, and when the sky is usually cloudy, that there is insufficient skin exposure to ultraviolet light to make the required amount of vitamin D. Under these conditions dietary intake becomes critical and cholecalciferol assumes the role of a vitamin. As little as five micrograms in the daily diet is considered adequate for children and most adults.

Vitamin E. Eight chemically similar substances have vitamin E activity; α-tocopherol is the most active of these and the most abundant in most foods. Vitamin E is fat-soluble and an antioxidant that is present in cell membranes. Its requirement is increased by high intakes of polyunsaturated fats.

Vitamin K. Vitamin K is a fat-soluble vitamin involved in the process of coagulation. This activity is found in several chemically different substances. One (vitamin K_1) occurs in some plant foods; others (the vitamin K_2 group) are formed by bacteria in the large intestine. Some of this is absorbed by the host.

Thiamine. Found in whole-grain cereals, meats, and nuts, thiamine (a water-soluble vitamin) plays an indispensable role as cofactor for enzymes involved in the metabolism of carbohydrates, alcohol, and some amino acids. The requirement is therefore proportional to an individual's calorie intake from foods other than fats. There are hardly any stores of thiamine in the body, and it only takes about 30 days without thiamine intake before signs of deficiency appear, provided the calorie intake is maintained. The recommended daily intake of thiamine is 1.0 to 1.2 milligrams for adults.

Riboflavin. A yellow, water-soluble compound, riboflavin occurs abundantly in milk and eggs. Riboflavin functions in that part of metabolism concerned with the oxidation of carbohydrates and amino acids, the constituents of proteins. An adult needs 1.1 to 1.3 milligrams of the vitamin per day.

Niacin. Niacin is a water-soluble vitamin that is also made in the human liver by the conversion of the amino acid tryptophan. If the protein intake is low, preformed niacin must be provided in the diet; if the protein intake is generous, niacin normally can be made by the body from tryptophan. Approximately 15 milligrams per day of niacin is required.

Vitamin B_6. Vitamin B_6 is a water-soluble vitamin of the B group. The five closely related substances with B_6 ac-

tivity, of which the best-known is pyridoxine, act as cofactors for the many enzymes involved in amino acid metabolism. Vitamin B_6 requirements, therefore, are proportional to protein intake.

Folic acid. A number of related water-soluble compounds have folic acid activity. The main function of folic acid is the synthesis of DNA. The vitamin is found in animal organs and whole-grain cereals, as well as in vegetables. The requirement for folic acid is notably increased in pregnancy.

Vitamin B_{12}. The requirement for vitamin B_{12}, a water-soluble compound, is the smallest of all the vitamins, only two micrograms per day, with liver stores usually enough to last five years of deprivation. Vitamin B_{12} participates with folic acid in DNA synthesis.

Pantothenic acid. A water-soluble vitamin, pantothenic acid promotes a large number of metabolic reactions essential for growth and well-being. It is found in liver, eggs, whole grains, legumes, and yeast.

Biotin. Biotin, a water-soluble compound, is a member of the B complex of vitamins. It functions in the formation and metabolism of fats and carbohydrates. It is especially abundant in eggs, beef liver, and yeast.

Vitamin C. Also water-soluble, vitamin C is found in almost all plant foods but not in meat. It is a powerful antioxidant and is required for the formation of collagen: *i.e.,* in wound healing. Relatively large amounts of vitamin C are required—for instance, an adult male is said to need about 90 mg per day.

Minerals. Unlike the complex organic compounds (carbohydrates, lipids, proteins, vitamins) discussed in previous sections, minerals are simple inorganic elements—often in the form of salts in the body—that are not themselves metabolized, nor are they a source of energy. Minerals constitute about 4 to 6 percent of body weight—about one-half as calcium, one-quarter as phosphorus (phosphates), and the remainder being made up of the other essential minerals that must be derived from the diet. Minerals not only impart hardness to bones and teeth but also function broadly in metabolism, *e.g.,* as electrolytes controlling the movement of water in and out of cells, as components of enzyme systems, and as constituents of many organic molecules.

As nutrients, minerals are traditionally divided into two groups according to the amounts present in and needed by the body. The major minerals (macrominerals), those required in large amounts of 100 milligrams or more per day, are calcium, phosphorus (phosphates), magnesium, sulfur, sodium, chloride, and potassium. The trace elements (microminerals or trace minerals), required in much smaller amounts of about 15 milligrams per day or less, include iron, zinc, copper, manganese, iodine (iodide), selenium, fluoride, molybdenum, chromium, and cobalt (as part of the vitamin B_{12} molecule). Fluoride is considered a beneficial nutrient because of its role in protecting against dental caries, although an essential function in the strict sense has not been established in human nutrition.

Ultratrace elements

The term "ultratrace elements" is sometimes used to describe minerals that are found in the diet in extremely small quantities (micrograms each day) and are present in human tissue as well; these include arsenic, boron, nickel, silicon, and vanadium. Despite demonstrated roles for these elements in experimental animals, the exact functions of these and other ultratrace elements (*e.g.,* tin, lithium, aluminum) in human tissue, and indeed their importance for human health, are uncertain.

Minerals have diverse functions, including muscle contraction, nerve transmission, blood clotting, immunity, maintenance of blood pressure, and growth and development. The major minerals, with the exception of sulfur, typically occur in the body in ionic (charged) form: sodium, potassium, magnesium, and calcium as positive ions (cations) and chloride and phosphates as negative ions (anions). Mineral salts dissolved in body fluids help regulate fluid balance, osmotic pressure, and acid-base balance.

Sulfur, too, has important functions in ionic forms (such as sulfate), but much of the body's sulfur is nonionic, serving as an integral part of certain organic molecules, such as the B vitamins thiamine, biotin, and pantothenic acid, and the amino acids methionine, cysteine, and cystine. Other mineral elements that are constituents of organic compounds include iron, which is part of hemoglobin, (the oxygen-carrying protein in red blood cells), and iodine, a component of thyroid hormones, which help regulate body metabolism. Additionally, many organic molecules contain phosphate groups, such as phospholipids in cell membranes, the genetic material (DNA and RNA), and the high-energy molecule adenosine triphosphate (ATP).

The levels of different minerals in foods are influenced by growing conditions (*e.g.,* soil and water composition) as well as by how the food is processed. Minerals are not destroyed during food preparation; in fact, a food can be burned completely and the minerals (ash) will remain unchanged. However, minerals can be lost by leaching into cooking water that is subsequently discarded.

Many factors influence mineral absorption and thus availability to the body. In general, minerals are better absorbed from animal than from plant foods, which contain fibre and other substances that interfere with absorption. Phytic acid, found principally in cereal grains and legumes, can form complexes with some minerals and make them insoluble and thereby indigestible. Only a small percentage of the calcium in spinach is absorbed because spinach also contains large amounts of oxalic acid, which binds calcium. Some minerals, particularly those of a similar size and charge, compete with each other for absorption. For example, iron supplementation may reduce zinc absorption, while excessive intakes of zinc can interfere with copper absorption. On the other hand, the absorption of iron from plants (nonheme iron) is enhanced when vitamin C is simultaneously present in the diet, and calcium absorption is improved by adequate amounts of vitamin D. Another key factor that influences mineral absorption is the physiological need for the mineral at the time.

Minerals can be toxic if taken in doses not far above recommended levels. This is particularly true for the trace elements, such as iron and copper. Accidental ingestion of iron supplements has been a major cause of fatal poisoning in young children. (J.W.)

Diet and diseases

SOURCES OF INFORMATION

Animal experiments. The results of studies on animals provide suggestive associations to humans, but for each nutritional biochemical process some or many animals may not be suitable models for humans. For example, rats do not need vitamin C in their diet because they synthesize it in their bodies from glucose. For a model of human scurvy, therefore, rats cannot be used. Only a very small number of higher animals are similar to humans in that they lack the enzyme to complete the synthesis of vitamin C. In 1907 it was discovered in Norway that guinea pigs must have vitamin C in their diets; research on vitamin C made rapid progress after this.

Animal models

Comparative and evolutionary studies. Humans and their predecessors have been on the Earth for at least 1,000,000 years. For 99 percent of this time they lived as hunter-gatherers; agriculture did not begin until 10,000 years ago. There has not been time for humans to evolve biochemical mechanisms to deal with the mixture of foods that has become usual in Western communities. Presumably human bodies are well adapted to the habits and diet of their hunter-gatherer forebears. There is some information about these diets (and a little about their effects) from archaeological records and some information from studies of the few isolated groups of hunter-gatherers remaining today.

Experiments of nature and travelers' tales. People in the United States and Britain cannot be persuaded to take an unfamiliar and unpleasant diet long-term in order to follow its effect on body functions. In other parts of the world, however, groups of people may be on restrictive diets for reasons such as culture and the availability of food. Diets may consist of a lot of fatty fish, or of blood and sour milk, or of mostly starchy low-protein roots, or of food cooked in salt water. It may be no coincidence that many of the leading human nutritionists have worked in

several different countries, returning with their experiences (the so-called travelers' tales). The trouble with such experiences is that the data are not very reliable.

Epidemiological studies. Epidemiology is the study of the distribution of disease in different sections of the population. Modern epidemiology is a more important tool for studying human nutrition than is animal nutrition. It has been particularly useful in finding which diets are associated with each of the chronic degenerative diseases—dental caries, coronary heart disease, high blood pressure, diverticulosis, and various cancers. Association, however, does not prove cause and effect.

Prospective studies

Prospective (or cohort) studies are more reliable than retrospective or cross-sectional ones. In a prospective study a large group of several thousand people in one place (a town or large firm) at risk of later developing a common disease have their medical histories taken and are given thorough examinations, including laboratory tests (blood biochemistry, etc.). Their health is then monitored, usually for five years or longer. If one variable at the initial examination is significantly associated with early development of disease X, this is considered evidence that a high value of the variable is a "risk factor" for disease X. If the same association should hold up in other prospective studies made in different places, the evidence becomes very strong.

Clinical records. Measurements of the food intake of patients with deficiency diseases are an important part of the classic evidence about these diseases. Precise information has come from experiences with patients fed entirely with solutions given by vein (total parenteral nutrition). Patients with extensive disease of the small intestine can be kept alive in this way. When patients rely on this route for all their nutrients for weeks, months, or years, the essentiality of trace elements becomes obvious. If a particular nutrient is not included in the fluid for total parenteral nutrition, the patient develops specific symptoms, which are then corrected when that nutrient is supplied. The role of diet in treatment of diseases such as diabetes mellitus or kidney failure is important. The dietary changes that will give the lowest blood glucose or blood urea, respectively, are explored by the nutritionist.

Food composition analysis. The independent variables in nutritional epidemiology and in dietary treatment of diseases are food constituents (protein, vitamin C, iron, etc.). Foods keep changing because of new methods of production, different formulations of ingredients, and changing consumption patterns. To facilitate international sharing of the available food composition data, Infoods (International Network of Food Data Systems) was set up in 1983, with its secretariat at the Massachusetts Institute of Technology in Cambridge.

Human experiments and trials. There are many different types of human experiments and trials. They range in duration from hours to years and in number of subjects from a handful to thousands. The variables measured can be clinical symptoms; biochemical substances in any available body fluid or accessible tissue, including expiratory air; body composition (estimates of size of the different compartments of the body); or the chemical transformations of radioactively labeled nutrients. Such studies are usually performed on healthy subjects. Some examples illustrate the scope.

Depletion diets

The first type of study involves the depletion of a single nutrient in human volunteers under experimental conditions (with all other essential nutrients provided, either in the food or as supplements). Ethical considerations are important, so that subjects often have been the scientists themselves and numbers usually have been very small. Such experiments, however, have provided important information. Some experiments have been continued (under close supervision) until the first clinical signs of deficiency appear; this takes three or four months for vitamin C and folic acid and much longer for vitamin A (because of large stores in the liver). Other experiments have used indexes such as low blood levels as measures and have stopped before any symptoms or signs appear.

A second type of study, involving absorption and uptake studies using test meals, takes only a few hours to carry out. Many are needed, however, because humans digest and absorb nutrients in different amounts and at different rates, depending on the food. For example, measures of iron uptake vary considerably, and a number of tests have been done with radioactive isotopes of iron (short half-life) to find the effects of different foods on uptake of this nutrient element.

In a third type of trial, the metabolic study, the diet is usually changed in one way only, and the result is measured in a change in blood or excreta. One type of study is the balance experiment, where the intake of a nutrient is controlled and measured and its output (in urine, feces, etc.) is measured. Because of variability in defecation, measurements have to be made for a metabolic "period" of about a week. The amount of the nutrient measured can be changed (with time allowed for adjustment) and the effect of this on the balance determined. When, for example, the output of nitrogen exceeds intake, there is a negative nitrogen balance; when output is less than intake, the balance is positive.

Another type of metabolic experiment measures the effect of a change of diet on plasma cholesterol. Such an effect takes 10 to 14 days to appear and requires control periods for comparison before and after the dietary change being examined. Whenever one component is added to, or removed from, a person's diet, another food must be removed or added to keep the calories and other important nutrients constant. That substitution may itself affect the variable under study.

Double-blind trials

In a fourth type of study, the double-blind trial, control groups receive identical-looking placebo tablets, and individuals are randomly allocated to an experimental or control group by someone who is not going to assess the clinical results. The researchers also do not know who took which tablets. Controlled double-blind trials of vitamin C tablets during winter months to see if they prevent colds measured incidence, severity, duration, and type of respiratory symptoms. Most of the trials have shown no statistically significant reduction in the incidence of colds with the use of vitamin C.

A fifth type of trial, the prevention trial, measures the effects of one combination of foods versus another. It is not possible to achieve double-blind conditions, but the control group is kept separate from the experimental group and the emphasis is on objective rather than subjective changes, such as biochemical measurements on blood samples and electrocardiographic changes. For example, the effects of a low saturated-fat, increased polyunsaturated-fat diet were measured in half of 850 middle-aged male war veterans in Los Angeles. The other half (the control group) consumed their usual foods. Variables measured were blood lipids, tissue fatty acid patterns (reflecting compliance with the prescribed increase of polyunsaturated fat), electrocardiographic changes, and the incidences of coronary and other diseases.

Classification of nutritional disorders. *Undernutrition.* Undernutrition is a condition in which there is not enough food energy as measured in calories or kilojoules. Undernutrition usually means that not enough total food is consumed. The main characteristics are loss of body weight and wasting of body fat and later muscle. Severe forms are starvation in adults and marasmus in young children. Undernutrition may be due to a failure of the food supply or to deliberate fasting, or it may be secondary to prolonged disease. Epidemic failure of the food supply is called famine.

Malnutrition. Malnutrition covers at least 25 different deficiency diseases resulting from lack of one of the essential nutrients: protein, vitamins, essential fats, or nutrient elements. Kwashiorkor, rickets, and iron-deficiency anemia are examples of the results of malnutrition. Multiple deficiencies are more likely than single, though one type often predominates. The deficiency disorder may be subclinical (recognizable only by laboratory tests) and not show characteristic features.

Obesity. Obesity is an excess of fat in the body (subcutaneous or internal) that results from a prolonged net positive energy balance; *i.e.,* the energy (calories or joules) consumed has been greater than the energy expended. Obesity is the opposite of undernutrition. Any or all of the

Positive energy balance

proximate energy-yielding nutrients can be responsible: usually fat, carbohydrate, and/or alcohol.

Nutrient excess. Nutrient excess is due to too much of one nutrient. For many nutrients a very high intake in the short term or a high intake over a long period (chronic nutrient toxicity) can be harmful. The nutrient may be an essential one (*e.g.,* as in vitamin D intoxication and iron overload), or it may be a nonessential nutrient (*e.g.,* alcohol or saturated fat, resulting in raised blood cholesterol).

Effects of toxicants. Unless very refined, foods contain hundreds of substances other than the nutrients, most in tiny amounts. Some naturally occurring toxicants can cause illness if people rely too heavily on a single foodstuff—*e.g.,* lathyrism, a paralysis of both legs (which occurs in India) from eating too much of the pulse *Lathyrus sativus* (grass pea). Potatoes contain alkaloids, including solanine, mostly in the green sprouts, and occasionally cases are reported of vomiting and other symptoms attributed to solanine poisoning. Other substances affect only a few individuals who are hypersensitive to them—*e.g.,* celiac disease (from sensitivity to wheat gluten) and urticaria (from sensitivity to such substances as naturally occurring salicylates). (A.S.T.)

DEFICIENCY DISEASES

Deficiency diseases may best be considered under six main headings, according to the principal deficiency: water, energy (calories), proteins, vitamins, inorganic elements, and fats and carbohydrates. There is considerable overlap among these groups.

Water. Some 65 percent of a normal body consists of water distributed between three compartments: within the cells, between the cells, and within the circulatory system. A great variety of diseases disturb the water balance, but in healthy people dehydration and overhydration result from wrong intake of water. The prominent symptom of dehydration is thirst, which develops quickly when water is unavailable. Dehydration becomes rapidly worse, and death may ensue within two or three days. Water is lost from the body by evaporation from the lungs and skin, through the urine, and through the stools, a loss that can become serious in diarrhea. Lung and skin losses are increased by work that causes sweating, particularly in a hot, dry climate. The kidneys usually excrete between one and two litres of urine per day, but when dehydration threatens they can conserve water through concentration of the urine. Overhydration from overdrinking is not a common occurrence in healthy people because the combined effects of excretion by the kidney and satiation, the opposite of thirst, are usually adequate even for the most enthusiastic drinker.

Dehydration and overhydration

The distribution of water between the three compartments is largely determined by the ions sodium and potassium. The important positively charged ion, or cation, in the interstitial (between the cells) fluid is sodium. Mammals, which have acquired an energy-dependent mechanism for removing sodium from the individual cells, the sodium pump, have potassium as their principal cation within the cells. (The cell membrane is permeable to sodium ions, and the sodium concentration outside the cells is stronger than within, so that one would expect sodium to pass from the tissue fluids into the cells; its movement in the opposite direction requires energy. The process that involves this expenditure of energy is called the sodium pump.) Ion and fluid balance are both normally maintained by the regulatory power of the sodium pump and of the kidney tubules. Upset of the delicate ion and fluid balance may be aggravated by deficient or excess intake of water.

The most severe dehydration occurs from diarrhea, with or without vomiting, in undernourished infants in poor communities who develop gastroenteritis, and in adults with cholera. The most efficient lifesaving treatment is intravenous infusion of sterile water containing glucose, sodium chloride, and potassium chloride. Water is poorly absorbed when drunk, but a good first aid formula consists of sodium chloride, 3.5 grams; sodium bicarbonate, 2.5 grams; potassium chloride, 1.5 grams; glucose 20 grams (or sucrose 40 grams); and clean water up to one litre. The glucose assists absorption of the sodium and vice versa, and as they are absorbed water is absorbed with them.

Energy (calories). Chemical energy in food, ultimately dependent on solar energy, occurs in carbohydrates, fats, and proteins (and alcohol). It is expended in mechanical work, in the work of the internal organs, in maintaining body temperature, and in growth. Undernutrition (calorie deficit) can result from (1) failure of the food supply, (2) loss of appetite, (3) fasting and anorexia nervosa, (4) persistent vomiting or inability to swallow, (5) incomplete absorption, comprising a group of diseases (malabsorption syndrome) in which digestion and intestinal absorption are impaired and there is excess loss of nutrients in the feces, (6) increased basal metabolic rate, as in prolonged fever, overactivity of the thyroid gland, or some cancers, and (7) loss of calories from the body; *e.g.,* glucose in the urine in diabetes.

Calorie deficit

Cessation of growth is an early feature of undernutrition in children. There is otherwise little difference between starving adults and marasmic infants. Body tissues are used as energy reserves. Body fat (adipose tissue) is the largest and most concentrated energy store and it is used first and most. As energy deficit continues, the proteins in muscles are used as well. Typically there is a large loss (wasting) of subcutaneous fat and a smaller loss (wasting) of the muscles. Some compensatory changes reduce the rate of energy expenditure: reduced voluntary activity, thyroid secretion, basal metabolic rate, and heart rate.

Severity of undernutrition is usually graded by comparing actual weight to the standard weight for a certain height. In adults undernutrition is defined as mild when weight is 81 to 90 percent of standard weight for the height; moderate when weight is 71 to 80 percent of standard weight for the height; and severe when weight is 70 percent or less of standard weight for the height. During famine people with mild starvation are in no danger; those with moderate starvation need extra feeding; those severely underweight need hospital care. Table 9 gives moderate (80 percent of standard weights) and severe (70 percent of standard) underweights for adults of different heights, and it can be used for men or women.

Undernutrition defined

Table 9: Underweights in Adults by Height

height		weight			
		80% of standard		70% of standard	
(m)	(in)	(kg)	(lb)	(kg)	(lb)
1.45	57.1	38	84	33	73
1.48	58.3	39.5	87	34.5	76
1.50	59.1	40.5	89	35	77
1.52	59.8	41.5	91.5	36	79
1.54	60.6	42.5	94	37	82
1.56	61.4	44	97	38.5	85
1.58	62.2	45	99	39	86
1.60	63.0	46	101	40	88
1.62	63.8	47	104	41	90
1.64	64.6	48.5	107	42	93
1.66	65.4	49.5	109	43	95
1.68	66.1	50.5	111	44	97
1.70	66.9	52	115	45.5	100
1.72	67.7	53	117	47	104
1.74	68.5	54.5	120	48	106
1.76	69.3	56	123.5	49	108
1.78	70.1	57	126	50	110
1.80	70.9	58	128	51	112.5
1.82	71.7	60	132	52	115
1.84	72.4	61	134.5	53.5	118
1.86	73.2	62	137	54	119
1.88	74.0	63.5	140	55.5	122
1.90	74.8	65	143	57	126

Protein. Because children have higher protein requirements per calorie and are more at risk of being given a low-protein diet, protein deficiency is more prominent among them. In most malnourished children, calorie deficiency and protein deficiency are combined, and features of both are detectable. There is a spectrum that begins with marasmus at one end, showing severe wasting but no signs of protein deficiency, and kwashiorkor at the other end, showing edema, severe reduction of the major plasma protein (albumin), and usually little weight loss.

The most common severe form of protein-energy malnutrition, nutritional marasmus usually occurs in the first

Nutritional marasmus

year of life. The affected child is extremely underweight and has lost all subcutaneous fat. The body is wizened, the face gaunt, and the limbs are like matchsticks. The cause is a diet very low in calories from all sources (including protein), often from early weaning to bottled formula diluted because of poverty. Poor hygiene leads to gastroenteritis and a vicious cycle of poor appetite and even more diluted formula. In turn further depletion leads to thinning of the inner lining of the intestine and reduction of its resistance to infection. In marasmus, the child's weight is 60 percent or less of standard weight for age (which is the median weight for healthy children in the United States). With such a large loss of tissue it takes about three months of good nutrition before the child's weight is normal, but the general condition is usually much improved after the first two weeks.

Kwashiorkor

In its fullest form, kwashiorkor is less common than marasmus. It was first described in Ghana in the 1930s, and the word kwashiorkor comes from the Ga language of that country. There is edema, with puffy feet or face, and patches of excess skin pigmentation and peeling. There may be little wasting or loss of weight. The hair is pale and thinned; the child is miserable and apathetic and will not eat. The liver is infiltrated with fat and the concentration of albumin in blood plasma is about one-quarter of the normal. While all agree that marasmus results from not enough food, the classical hypothesis of the pathogenesis of kwashiorkor—insufficient protein with adequate starchy food and sugar—is not fully accepted by all researchers. A major reason is that the dietary histories of children with kwashiorkor are inaccurate.

Treatment of malnutrition

Treatment of protein-energy malnutrition is essentially the same whether it is marasmus or kwashiorkor. There are three phases to treatment. (1) Resuscitation is the correction of dehydration, deficiency of potassium, low blood sugar, and low body temperature and the treatment of infections. (2) The cure is begun with mixed foods, usually including milk, providing increasing calories and protein per kilogram of body weight. Appetite may be poor and the child must be fed by hand. A multivitamin preparation and supplements of potassium, magnesium, and zinc are needed. (3) Rehabilitation takes place after about three weeks when the child usually has improved and has a good appetite, but is still underweight. It will take many weeks of good nutrition before catch-up growth is complete. If the home background is unfavourable because of poverty or social disorganization, the mother will need support as well as advice.

Other nutrients are often deficient in malnourished children: potassium, magnesium, zinc, folic acid, vitamin A, iron, and others that differ from one region (and staple diet) to another. Children malnourished for a long period in early childhood are likely to grow up to be short adults. A more serious and complex question is whether severe malnutrition impairs a child's mental ability. Malnourished children often come from an emotionally and educationally deprived home; however, the balance of evidence from psychological follow-up studies shows that children who have been malnourished have the potential to reach normal intellectual development, though they may not. The outcome is likely to be favourable if the child returns to an encouraging and stimulating home and unfavourable if the malnutrition was prolonged and occurred in the first year of life.

Mild malnutrition

In a developing country, for every child with marasmus, marasmic kwashiorkor, or kwashiorkor (the severe forms), there are likely to be about 10 with mild to moderate forms, the most useful sign of which is a low weight for age (between 61 and 80 percent of the normal standard). It has been estimated that 2 percent of young children in Third World countries have severe malnutrition, while about 20 percent show mild to moderate forms. Such children are not growing at their genetic potential, and they have an increased risk of severer attacks of common infections of childhood (gastroenteritis, respiratory infections, measles, etc.). Two keys to recognizing mild to moderate protein-energy malnutrition are widespread inexpensive maternal and child health clinics, each with a weighing machine, and a clear and simple growth chart (Figure 2). Another simple

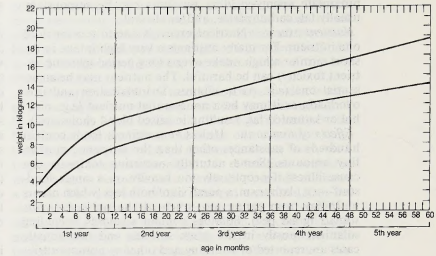


Figure 2: *Standard weights for children.*
The top line is the World Health Organization standard weight for age (the 50th percentile of the National Center for Health Statistics reference) and the bottom line is the 3rd percentile of the reference, close to 80 percent of the standard. In prosperous communities children's weights should be close to the top line. In poor communities they should not fall too far below the bottom line.
By courtesy of the World Health Organization

measurement is the mid-upper-arm circumference: from 12 to 60 months of age a measurement of 13.5 centimetres (5.3 inches) or above is normal, which remains the same for these four years. A circumference of 12.5 to 13.5 centimetres suggests mild malnutrition and under 12.5 centimetres indicates definite malnutrition.

Vitamins. Each of the vitamin-deficiency diseases is commonly attributed to a deficiency of a single vitamin, and many have been produced in animals and in human volunteers by diets apparently complete except for the vitamin under trial. In human experience, these supposed single-vitamin diseases usually arise from unhealthy and unbalanced diets that are also deficient in other vitamins or in nonvitamin nutrients. In these complex nutrient-deficiency states, the nutrient whose deficiency plays the largest part in the causation of symptoms is often referred to as the most limiting nutrient; *e.g.,* niacin is the most limiting nutrient in pellagra, but there are often deficiencies of other B vitamins, and protein deficiency is commonly present.

Scurvy

A good example of a single-nutrient disease is scurvy, the scourge of mariners in the days of long ocean voyages. In 1754 a Scottish naval surgeon, James Lind, established the concept of deficiency diseases by showing that scurvy can be both cured and prevented by the use of orange and lemon juice. Scurvy is an eminently preventable disease, and it is now seen mainly at the extremes of life and where there is disregard for the principles of nutrition. At an early stage scurvy results in swelling and bleeding of the gums and bleeding into the skin. Bleeding into the muscles and under the periosteum, the membrane that covers the bones, in small children may also occur. It is probable that in the classical epidemics among mariners and in prisons the diet was unsatisfactory in many ways and that other deficiency states contributed to the more serious complications and to death, but experiments on human volunteers fed diets lacking in vitamin C but otherwise satisfactory have shown that scurvy can be and often is a single-nutrient-deficiency disease.

Vitamin A deficiency

The earliest manifestation of vitamin A deficiency is night blindness, because the retinal pigment rhodopsin (visual purple) in the retina of the eye requires retinol (vitamin A) for its functioning. Retinol deficiency manifests itself as inability of the eyes to adapt in dim light. Moreover, epithelial cells throughout the body—the various cells covering or lining the body and its organs—degenerate in vitamin A deficiency. In the eyes, the conjunctiva and cornea undergo changes (xerophthalmia) that can lead eventually to breakdown and complete blindness. Other epithelial cells, including the skin and the mucous lining of the respiratory and urinary tracts, degenerate, with characteristic results such as hardening and drying of the skin, respiratory in-

fections, and kidney stones. (All of these changes can be simulated to more or less degree by diseases that have nothing to do with vitamin A deficiency.)

Xerophthalmia is a late manifestation of vitamin A deficiency. It has been estimated that there may be as many as 500,000 new cases of xerophthalmia each year, half of which lead to loss of vision. The highest incidence of xerophthalmia is found in East Asia—*e.g.,* Bangladesh, India, Indonesia, Philippines. It is virtually confined to infants and young children and usually is associated with protein malnutrition. In the early stages of xerophthalmia, the conjunctiva is dry, wrinkled, and thickened. If not treated with vitamin A at this stage the cornea becomes involved (keratomalacia). The result is a scar, but there is a danger the cornea will perforate, resulting in total blindness in the eye. Where xerophthalmia occurs in a community, other children may be found to have low plasma levels of vitamin A. Children with mild eye signs (night blindness or early xerophthalmia) have an increased risk of acquiring severe measles and respiratory infections. These children develop vitamin A deficiency first because the mother was undernourished, so that they are born with small liver stores of vitamin A and their mother's milk has a low vitamin A content; second because they are given little or no dietary preformed vitamin A or β-carotene; and third because absorption of vitamin A is impaired in protein malnutrition. The World Health Organization has made prevention of xerophthalmia one of its priorities in nutrition. In the short term, the incidence of the disease is being reduced by giving young children a high-dose vitamin A capsule every six months. The long-term solution is education, so that mothers and babies regularly eat foods rich in β-carotene, such as dark green leafy vegetables or pumpkin, squash, mangoes, pawpaw, even sweet potatoes and yellow corn, which all provide β-carotene.

(J.F.B./A.S.T.)

Vitamin D (also known as vitamin D hormone) is synthesized in the body in a series of steps, starting in the skin by the action of sunlight's ultraviolet rays on a precursor compound; thus, without adequate food sources of vitamin D, a deficiency of the vitamin can occur when exposure to sunlight is limited. Lack of vitamin D in children causes rickets, a disease characterized by inadequate mineralization of bone, growth retardation, and skeletal deformities, such as bowed legs. The adult form of rickets, known as osteomalacia, results in weak muscles as well as weak bones. Individuals with limited sun exposure (including women who completely cover their bodies for religious reasons) and elderly or homebound persons, and those with dark skin, particularly those who live in northern latitudes, are at risk of vitamin D deficiency. Fortification of milk and other foods (*e.g.,* margarine, cereals, and breads) with the vitamin has helped protect those populations in which sun exposure is inadequate. Supplemental vitamin D also may help protect against bone fractures in the elderly, who make and activate vitamin D less efficiently.

Rickets and osteomalacia

Vitamin E deficiency is rare in humans, although it may develop in premature infants and in people with impaired fat absorption or metabolism. In the former, fragility of blood cells (hemolysis) is seen; in the latter, where deficiency is more prolonged, neuromuscular dysfunction involving the spinal cord and the retina may result in loss of reflexes, impaired balance and coordination, muscle weakness, and visual disturbances. No specific metabolic function has been established for vitamin E; however, it is an important part of the antioxidant system that inhibits lipid peroxidation; *i.e.,* it protects cells and their membranes against the damaging effects of free radicals (reactive oxygen and nitrogen species) that are produced metabolically or enter the body from the environment. The requirement for vitamin E is increased with increasing consumption of polyunsaturated fatty acids. People who smoke or are subjected to air pollution may also need more of the vitamin to protect against oxidative damage to the lungs.

Vitamin K is necessary for the formation of prothrombin and other blood-clotting factors in the liver, and it also plays a role in bone metabolism. A form of the vitamin is produced by bacteria in the colon and can be utilized to some degree. Vitamin K deficiency causes impaired clotting of the blood and internal bleeding, even without injury. Due to poor transport of vitamin K across the placenta, newborn infants in developed countries are routinely given the vitamin within six hours of birth to protect against a condition known as hemorrhagic disease of the newborn. Vitamin K deficiency is rare in adults, except in syndromes with poor fat absorption, in liver disease, or during treatment with certain anticoagulant drugs, which interfere with vitamin K metabolism. Bleeding due to vitamin K deficiency may be seen in patients whose gut bacteria have been killed by antibiotics.

Prolonged deficiency of thiamine (vitamin B_1) results in beriberi, a disease that has been endemic in populations where white rice is a staple. Thiamine deficiency is still seen in areas where white rice or flour constitute the bulk of the diet and thiamine lost in milling is not replaced through enrichment. Symptoms of the form known as dry beriberi include loss of appetite, confusion and other mental symptoms, muscle weakness, painful calf muscles, poor coordination, tingling, and paralysis. In wet beriberi there is edema and the possibility of an enlarged heart and heart failure. Thiamine deficiency can also occur in populations eating large quantities of raw fish harbouring intestinal microbes that contain the enzyme thiaminase. In the developed world, thiamine deficiency is linked primarily to chronic alcoholism with poor diet, manifesting as Wernicke-Korsakoff's syndrome, a condition with rapid eye movements, loss of muscle coordination, mental confusion, and memory loss.

Beriberi

Vitamin B_6 (pyridoxine and related compounds) is essential in protein metabolism, the synthesis of neurotransmitters, and other critical functions in the body. Deficiency symptoms include dermatitis, anemia, impaired immune function, depression, confusion, and convulsions. Although full-blown vitamin B_6 deficiency is rare, marginal inadequacy is more widespread, especially among the elderly, who may have a reduced ability to absorb the vitamin. People with alcoholism, especially those with the liver diseases cirrhosis and hepatitis, are at risk of deficiency. A number of drugs, including the drug isoniazid, used in the treatment of tuberculosis, interfere with vitamin B_6 metabolism.

Riboflavin (vitamin B_2) deficiency, known as ariboflavinosis, is unlikely without the simultaneous deficiency of other nutrients. After several months of riboflavin deprivation, symptoms include cracks in the skin at the corners of the mouth, fissures of the lips, and an inflamed, magenta-coloured tongue. Because riboflavin is readily destroyed by ultraviolet light, jaundiced infants who are treated with light therapy are administered the vitamin. Milk, milk products, and cereals, major sources of riboflavin in the diet, are packaged to prevent exposure to light.

Deficiency of biotin is rare, and this may be due in part to synthesis of the vitamin by bacteria in the colon, although the importance of this source is unclear. Biotin deficiency has been observed in people who regularly eat large quantities of raw egg white, which contains a glycoprotein (avidin) that binds biotin and prevents its absorption. A rare genetic defect that renders some infants unable to absorb a form of biotin in food can be treated with a supplement of the vitamin. Long-term use of certain anticonvulsant drugs may also impair biotin absorption. Symptoms include skin rash, hair loss, and eventually neurological abnormalities.

(J.W.)

Pellagra, before the 20th century, occurred in Italy (the name is Italian for "rough skin") and in the southern United States. It has now disappeared from Europe and North America but still occurs in Egypt and southern Africa. It affects poor people who subsist primarily on corn and little other food. In industrial countries it is occasionally seen in alcoholics. The principal nutritional deficiency in pellagra is of niacin. In corn the niacin is in bound form, not able to be broken down by humans unless prepared in a special way. In addition, unlike other cereals, corn is low in the amino acid tryptophan, a precursor of niacin. Pellagra is the disease of dermatosis, diarrhea, and dementia. The first clinical features of pellagra are demarcated, symmetrical areas of pigmentation and thickening and crack-

Pellagra

ing of the skin where it is exposed to sunlight. The collar area is usually affected and provides a useful diagnostic sign ("Casal's collar"). In severer cases there are mental disturbances such as fatigue, depression, and progressive confusion. Gastrointestinal effects include diarrhea. Other B vitamins, especially riboflavin and vitamin B_6 are usually deficient, and complete treatment calls for multiple B vitamins (in diet or tablets), not just niacin.

Folic acid and vitamin B_{12} deficiencies

The major effect of deficiency of either folic acid or vitamin B_{12} is the failure of cell division because one of the family of folic acid compounds (5, 10-methyltetrahydrofolate) is required for addition of a methyl group to uracil to form thymine, one of the four bases of DNA. An important function of vitamin B_{12} involves interconversion of folic acid compounds. (When it is lacking, the methyltetrahydrofolate compound accumulates and the compound required for DNA synthesis is not made.)

Cells that turn over rapidly are most affected by deficiencies of folic acid and vitamin B: red blood cells, white blood cells, and platelets all decrease in number. Replacement of the lining cells of the intestine and formation of germ cells are also affected. In the bone marrow the red cell precursors are larger than normal, with an immature nucleus, called megaloblastic change. Thus folic acid deficiency and vitamin B_{12} deficiency each causes megaloblastic anemia.

Vitamin B_{12} plays another biochemical role, unrelated to folic acid, in the maintenance of myelin (the sheath that forms around a nerve fibre) in the nervous system. In vitamin B_{12} deficiency, unusual odd-numbered fatty acids appear in myelin, apparently because of the accumulation of three-carbon compounds instead of the usual two-carbon raw material for fatty acid synthesis. The usual symptom is subacute degeneration of both sensory and motor tracts in the spinal cord.

Folic acid deficiency can occur in late pregnancy, when the requirement is double that in other adults. It is also fairly common in patients with diseases in which cell division is increased; *e.g.,* in blood diseases or cancer. The most common cause of vitamin B_{12} deficiency is pernicious anemia, in which absorption of the vitamin is defective because of the failure of the stomach to secrete a special protein (intrinsic factor) that assists absorption of vitamin B_{12} in the lower intestine. Dietary deficiency occurs in vegans, who eat no animal food. It takes five or more years of such a diet before symptoms appear because the stores of vitamin B_{12} in the liver, although only about 1.5 milligrams, are enough to last about five years with the daily requirement as small as one microgram.

Fats and carbohydrates. Most human diets contain between 10 and 15 percent of the total calories as protein. In poor communities this is mostly from vegetable foods; in rich communities it is more from animal foods. The rest of the dietary energy comes from carbohydrates, fats, and, in some people, alcohol. The proportion of calories from fats is only about 10 percent in poor communities, but it is 40 percent or even more in rich communities. Carbohydrates vary reciprocally: from 75 to 80 percent in poor communities to 40 percent in rich communities (Figure 3).

Humans are adaptable omnivores and survive on a greater variety of foods and metabolic fuels than do many other animals. Adaptations occur to a low or high carbohydrate/fat ratio. Dietary carbohydrates can be considered for nutritional purposes in three groups: (1) metabolically available sugars, (2) starches, and (3) carbohydrates not absorbed in the small intestine (unavailable carbohydrates)—mostly cellulose, hemicelluloses, and pectin. These are the main types of dietary fibre. The minimum requirement for available carbohydrate is low, only about 50 grams for an adult. Below this intake ketosis occurs. In practice, few diets contain as few carbohydrates as this, except for discredited very-low-carbohydrate weight-loss diets.

Essential fatty acids

There is also a minimum requirement for fat—not for total fat, but only for a small group of fatty acids, of which linoleic acid is the most abundant and arachidonic acid and α-linolenic acid are the only other important ones (see Table 8). The basic requirement for essential fatty acids to prevent deficiency is 1 to 2 percent of dietary energy; *i.e.,* two to six grams of linoleic acid. The characteristics of essential fatty acid deficiency include a scaly dermatitis, poor wound healing, anemia, reduced platelet count, and fatty infiltration of the liver. Deficiency of essential fatty acids as a result of an inadequate diet is rare. Even if people do not eat separated fats (butter, oils, etc.), they usually obtain enough linoleic acid from the cell membranes of plant foods and meats. With total parenteral nutrition, however, essential fatty acid deficiency is well described. It will occur if this form of feeding is prolonged unless a lipid emulsion rich in linoleic acid (*e.g.,* Intralipid) is added to the administered fluids. There is good evidence that amounts of linoleic acid several times the basic requirement can contribute to reducing the risk of coronary heart disease by different mechanisms, such as lowering the plasma cholesterol and reducing the tendency to thrombosis.

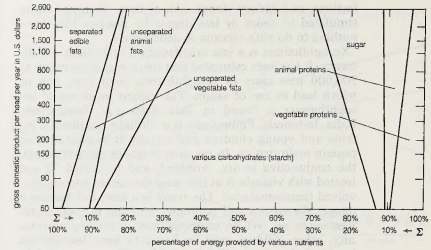


Figure 3: Dietary energy derived from fats, carbohydrates, and proteins as a percentage of total energy, related to the income or gross domestic product of 85 countries at 1962 values of money on a logarithmic scale.

From J. Persse, F. Sizaret, and P. Francois, *Nutrition Newsletter* (1969); Food and Agriculture Organization

Requirements

Inorganic elements. Twenty-two of the 103 chemical elements are required as nutrients by humans. Another three or four might yet be shown to be needed in minute amounts. Carbon, nitrogen, hydrogen, and oxygen make up the organic nutrients, and the latter two are also consumed as water. Cobalt is required only as vitamin B_{12} and sulfur only as part of the essential amino acids methionine and cystine. Requirements for the rest range from 0.1 milligram or less for iodine, selenium, and chromium to up to one to five grams per day for sodium, chloride, and potassium. Elements required in minute amounts, usually present in food and body tissues in small amounts, were formerly called trace elements because analysts could only report presence of an amount too small to quantify. With modern methods, such as advanced atomic absorption spectrometry, extremely low concentrations can be accurately determined, and there is no reason to separate trace elements from minerals. Table 5 shows that there is no grouping of requirements for the inorganic elements.

Sodium depletion

Sodium depletion occurs, usually with accompanying water loss, as a result of massive loss of fluids; *e.g.,* diarrhea, vomiting, and excessive urination. Sodium is the major ion in the extracellular fluid, the total volume of which contracts as does the blood volume; the blood pressure falls and the heart rate increases. The patient is weak and exhausted. Sodium depletion occurs in a number of disease states, and chloride depletion usually accompanies it.

Potassium depletion occurs in similar situations and may become obvious if sodium and water, but not potassium, are replaced. Potassium loss is especially likely to occur with diarrhea or overuse of purgatives, with regular use of diuretic drugs, with corticosteroid treatment, and with wasting diseases. In potassium depletion the serum potassium level is low, there is weakness of voluntary muscles, intestinal peristalsis may stop, and an electrocardiograph shows low T waves.

Calcium depletion, unlike that of sodium and potassium, is difficult to recognize because 99 percent of the calcium in the body is in the bones. A 1 percent reduction of bone calcium is impossible to detect, but it is quite a large

amount of the element (10 to 12 grams). The major questions about calcium are: whether a generous intake during childhood and adolescence will contribute to taller adult height or heavier bones (less likely to fracture or become osteoporotic in old age), and whether a generous intake of calcium from about 45 years onward will delay the development of osteoporosis. Osteoporosis is reduced bone mass (*i.e.*, for a particular length or volume of bone, there is less mineral [calcium phosphate] and less bone protein). Osteoporosis is common in postmenopausal women in industrial societies. It causes bone pain and a tendency to fractures. It clearly has multiple causes, including lack of exercise as well as the possibility of insufficient dietary calcium. (A.S.T.)

Iron deficiency is the most common of all nutritional deficiencies, with much of the world's population being deficient in the mineral to some degree. Young children and premenopausal women are the most vulnerable. The main function of iron is in the formation of hemoglobin, the red pigment of the blood that carries oxygen from the lungs to other tissues. Since each millilitre of blood contains 0.5 milligram of iron (as a component of hemoglobin), bleeding can drain the body's iron reserves. When iron stores are depleted, a condition arises known as microcytic, hypochromic anemia, characterized by small red blood cells that contain less hemoglobin than normal. Symptoms of severe iron deficiency anemia include fatigue, weakness, apathy, pale skin, difficulty breathing on exertion, and low resistance to cold temperatures. During childhood, iron deficiency can affect behaviour and learning ability as well as growth and development. Severe anemia increases the risk of pregnancy complications and maternal death. Iron deficiency anemia is most common during late infancy and early childhood, when iron stores present from birth are exhausted and milk, which is poor in iron, is a primary food; during the adolescent growth spurt; and in women during the childbearing years, because of blood loss during menstruation and the additional iron needs of pregnancy. Intestinal blood loss and subsequent iron deficiency anemia in adults may also stem from ulcers, hemorrhoids, tumours, or chronic use of certain drugs such as aspirin. In developing countries, blood loss due to hookworm and other infections, coupled with inadequate dietary iron intake, exacerbates iron deficiency.

Anemia

Iodine deficiency disorders are the most common cause of preventable brain damage, affecting an estimated 50 million people worldwide. During pregnancy, severe iodine deficiency may impair fetal development, resulting in cretinism (irreversible mental retardation with short stature and developmental abnormalities) as well as in miscarriage and stillbirth. Other more pervasive consequences of chronic iodine deficiency include lesser cognitive and neuromuscular deficits. The ocean is a dependable source of iodine, but away from coastal areas iodine in food is variable and largely reflects the amount in the soil. In chronic iodine deficiency the thyroid gland enlarges as it attempts to trap more iodide (the form in which iodine functions in the body) from the blood for synthesis of thyroid hormones, and it eventually becomes a visible lump at the front of the neck known as a goitre. Some foods, such as cassava, millet, sweet potato, certain beans, and members of the cabbage family, contain substances known as goitrogens that interfere with thyroid hormone synthesis; these substances, which are destroyed by cooking, can be a significant factor in persons with coexisting iodine deficiency who rely on goitrogenic foods as staples. Since a strategy of universal iodization of salt was adopted in 1993, there has been remarkable progress in improving iodine status worldwide. Nonetheless, millions of people living in iodine-deficient areas, primarily in Central Africa, Southeast and Central Asia, and even in central and eastern Europe, remain at risk.

Zinc deficiency in humans was first reported in the 1960s in Egypt and Iran, where children and adolescent boys with stunted growth and undeveloped genitalia responded to treatment with zinc. Deficiency of the mineral was attributed to the regional diet, which was low in meat and high in legumes, unleavened breads, and whole-grain foods that contain fibre, phytic acid, and other factors that inhibit zinc absorption. Also contributing to zinc deficiency was the practice of clay eating, which interferes with the absorption of zinc, iron, and other minerals. Severe zinc deficiency has also been described in patients fed intravenous solutions inadequate in zinc and in those with the inherited zinc-responsive syndrome known as acrodermatitis enteropathica. Symptoms of zinc deficiency may include skin lesions, diarrhea, increased susceptibility to infections, night blindness, poor appetite, hair loss, slow wound healing, low sperm count, and impotence. Zinc is highest in protein-rich foods, especially red meat and shellfish, and zinc status may be low in protein-energy malnutrition. Even in developed countries, young children, pregnant women, the elderly, strict vegetarians, alcoholics, and people with malabsorption syndromes are particularly vulnerable to zinc deficiency.

Fluoride contributes to the mineralization of bones and teeth and protects against tooth decay. Epidemiological studies in the United States in the 1930s and 1940s revealed an inverse relationship between the natural fluoride content of waters and the rate of dental caries. Whenever fluoridation has been introduced, the number of decayed, missing, and filled teeth in children has decreased, and where fluoridation has been subsequently stopped, the rate of caries has increased again. In communities and countries where the drinking water is not fluoridated, many obtain some fluoride from toothpastes and mouthrinses, a few give fluoride supplements to children, and dentists periodically apply fluoride solution directly to their patients' teeth.

Fluoridation

Dietary fibre. Dietary fibre, the structural parts of plants, cannot be digested by the human intestine because the necessary enzymes are lacking. Even though these nondigestible compounds pass through the gut unchanged (except for a small percentage that is fermented by bacteria in the large intestine), they nevertheless contribute to good health. Insoluble fibre, which does not dissolve in water, provides bulk, or roughage, that helps with bowel function (regularity) and accelerates the exit from the body of potentially carcinogenic or otherwise harmful substances in food. Types of insoluble fibre are cellulose, most hemicelluloses, and lignin. Major food sources of insoluble fibre are whole-grain breads and cereals, wheat bran, and vegetables. Soluble fibre, which dissolves or swells in water, slows down the transit time of food through the gut (an undesirable effect) but also helps lower blood cholesterol levels (a desirable effect). Types of soluble fibre are gums, pectins, some hemicelluloses, and mucilages; fruits (especially citrus fruits and apples), oats, barley, and legumes are major food sources. Both soluble and insoluble fibre help delay glucose absorption, thus ensuring a slower and more even supply of blood glucose. Dietary fibre is thought to provide important protection against some gastrointestinal diseases and to reduce the risk of other chronic diseases as well. (J.W.)

OBESITY AND WEIGHT CONTROL

Obesity. The World Health Organization (WHO) has recognized obesity as a worldwide epidemic, affecting an estimated 300 million adults—and millions of children—and paradoxically coexisting with undernutrition in both developing and industrialized countries. Obesity (excess body fat for stature) contributes to adverse health consequences such as high blood pressure, blood lipid abnormalities, coronary heart disease, congestive heart failure, ischemic stroke, type 2 diabetes, osteoarthritis, several common cancers, and reduced life expectancy. Genes play a significant role in the regulation of body weight. Nevertheless, environmental factors such as calorie-rich diets and a sedentary lifestyle can be instrumental in determining how an individual's genetic heritage will unfold.

Dietary carbohydrates are not the problem in obesity. In some Asian cultures, for example, where carbohydrate foods such as rice predominate, people are relatively thin, and heart disease and diabetes rates are lower than they are in Western cultures. What matters in weight control is the ratio of food energy (calories) consumed to energy expended, over time.

Height-weight tables as a reference for healthy weights

Body mass index

have been supplanted by the parameter known as the body mass index (BMI). The BMI estimates total body fat, although it is less sensitive than using a skinfold caliper or other method to measure body fat indirectly. The BMI is defined as weight in kilograms divided by the square of the height in metres: weight ÷ height^2 = BMI. In 1997 the WHO recommended international adoption of the definition of a healthy BMI for adult women and men as between 18.5 and 24.9. A BMI lower than 18.5 is considered underweight; a BMI of 25 or higher denotes overweight and, within that category, 30 or higher indicates obesity. Definitions of overweight and obesity are more difficult to quantify for children, whose BMI changes with age.

A healthful eating plan for gradual weight loss in adults will likely contain about 1,200 to 1,500 kilocalories per day, probably accompanied by a balanced vitamin and mineral supplement. A desirable weight loss is about one pound per week from fat stores (as opposed to lean tissue), which requires an energy deficit of 3,500 kilocalories, or about 500 kilocalories per day. Consuming less than 1,000 kilocalories per day is not recommended; a preferred approach would be to increase physical activity, which has the added benefit of helping to maintain lean tissue. Individuals who are severely obese and unable to lose weight may, after medical consultation, consider weight-loss medications that suppress appetite or decrease nutrient absorption or even surgery to reduce the volume of the stomach or bypass it altogether. Carbohydrate-restricted diets, very-low-fat diets, and novelty diets—those in which one food or food group is emphasized—may result in weight loss but generally fail to establish the good dietary and exercise practices necessary to maintain the desired weight, and weight is often regained shortly after the diet is stopped.

A successful approach to long-term weight management requires establishing new patterns: eating healthfully, but eating less; engaging in regular physical activity; and changing behaviour patterns that are counterproductive, such as eating while watching television. Limiting intake of fatty foods (see Table 10), which are more energy-rich, is also helpful, as are eating smaller portions and drinking water instead of calorie-containing drinks. Low-fat foods are not always low in total calories, as the fat may be replaced by sugars, which themselves provide calories.

In the 1990s, research with genetically obese laboratory animals led to the discovery of the *ob* gene in mice and humans. Under the direction of this gene, adipose (fat) tissue cells secrete leptin, a protein hormone. When fat stores increase, leptin sends a signal to the hypothalamus (a regulatory centre in the brain) that stimulates one to eat less and expend more energy. Certain genetic mutations result in insufficient production of functional leptin or in a failure to respond to the leptin signal. Treatment with leptin may prove useful for the small percentage of obese persons who have a defect in the *ob* gene, although it is not yet known whether leptin therapy will induce weight loss in those who are leptin-resistant or who do not have mutations in the *ob* gene.

Anorexia nervosa and bulimia nervosa

Eating disorders. Eating disorders such as anorexia nervosa and bulimia nervosa are serious health problems reflecting an undue concern with body weight. Girls and young women are most vulnerable to the pressures of society to be thin, although boys and men can also fall prey to these disorders, which have lifelong consequences and can even be fatal. The incidence of eating disorders has risen during the last 50 years, particularly in the United States and western Europe.

Anorexia nervosa is characterized by low body weight, propensity for drastic undereating, intense fear of gaining weight or becoming fat (despite being underweight), and a distorted body image. Consequences include impaired immunity, anemia, and diminished digestive function. Without intervention, a state of semi-starvation similar to marasmus may occur, requiring hospitalization and even force-feeding to prevent death. Treatment usually requires a coordinated approach, with the participation of a physician, psychiatrist, and dietitian.

Bulimia nervosa is thought to be more prevalent than anorexia nervosa, and both disorders may even occur in the same person. In bulimia nervosa recurrent episodes of "binge eating" are followed by a form of purging, such as self-induced vomiting, fasting, excessive exercise, or the use of laxatives, enemas, or diuretics. Treatment usually involves a structured eating plan.

Young athletes often restrict energy intakes to meet weight guidelines and body-image expectations of their sport. Females are most affected, but male athletes, such as gymnasts, wrestlers, boxers, and jockeys, are also vulnerable. Intense training among young female athletes, coupled with food-energy restriction, often results in amenorrhea (cessation of menstruation) and bone loss similar to that at menopause. (J.W.)

NUTRIENT TOXICITIES

Alcohol. The acute effects of a large intake of alcohol are well-known. Mental impairment starts when the blood concentration is about 0.05 percent. A concentration of alcohol in the blood of 0.40 percent usually causes unconsciousness, and 0.50 percent can be fatal. Automobile accidents and violence, which are often alcohol-related, are major causes of death for young persons. Women who drink during pregnancy risk physical and mental damage to their babies, a condition called fetal alcohol syndrome. Alcohol also can interact dangerously with a variety of medications, such as tranquilizers, antidepressants, and pain relievers. Although numerous studies have confirmed that light to moderate drinkers have less heart disease and tend to live longer than either nondrinkers or heavy drinkers, increasing chronic alcohol consumption carries with it significant risks as well, including liver disease, pancreatitis, and cancer. In alcoholics, nutritional impairment may result from the displacement of nutrient-rich food as well as from complications of gastrointestinal dysfunction and widespread metabolic alterations. Deficiency of the vitamin thiamine, as seen in the neurological condition Wernicke-Korsakoff's syndrome, is a hallmark of alcoholism and requires urgent treatment.

Tolerable Upper Intake Level

Vitamins. In 1997 the United States Institute of Medicine established a reference value called the Tolerable Upper Intake Level (UL) for selected nutrients, which is also being used as a model for other countries. The UL is the highest level of daily nutrient intake likely to pose no risk of adverse health effects for almost all individuals in the general population and are not meant to apply to people under medical supervision. Discussed below as "safe intakes," most ULs for infants, children, and adolescents are considerably lower.

Because they can be stored in the liver and fatty tissue, fat-soluble vitamins, particularly vitamins A and D, have more potential for toxicity than do water-soluble vitamins, which, with the exception of vitamin B_{12}, are readily excreted in the urine if taken in excess. Nonetheless, water-soluble vitamins can be toxic if taken as supplements or in fortified food.

Symptoms of acute vitamin A poisoning, which usually require a dose of at least 15,000 micrograms (50,000 International Units) in adults, include abdominal pain, nausea, vomiting, headache, dizziness, blurred vision, and lack of muscular coordination. Chronic hypervitaminosis A, usually resulting from a sustained daily intake of 30,000 micrograms (100,000 IU) for months or years, may result in loss of bone density and liver damage. Vitamin A toxicity in young infants may be seen in a swelling of the fontanelles (soft spots of the skull) due to increased intracranial pressure. Large doses of vitamin A taken by a pregnant woman also can cause developmental abnormalities in a fetus, especially if taken during the first trimester; the precise threshold for causing birth defects is unknown, but less than 3,000 micrograms (10,000 IU) daily appears to be a safe intake. Although most vitamins occurring naturally in food do not cause adverse effects, toxic levels of vitamin A may be found in the liver of certain animals. High β-carotene intake, from supplements or from carrots or other foods that are high in β-carotene, may after several weeks impart a yellowish cast to the skin but does not cause the same toxic effects as preformed vitamin A.

Large doses of vitamin B_6 have been taken in hopes of treating conditions such as carpal tunnel syndrome and premenstrual syndrome. The most critical adverse effect

Table 10: Representative Energy Values of Some Common Foods

food	kilocalories	food	kilocalories	food	kilocalories
Cucumber (2 oz)	5	Bread (1 slice, 1 oz)	65	Carbonated soft drink (12-oz can)	155
Lettuce (1 oz)	5	Potatoes (new, boiled, 4 oz)	85	Peanuts (1 oz)	160
Cauliflower (4 oz, raw)	15	Egg (medium, boiled)	90	Cake, sponge (2 oz)	170
Tomatoes (4 oz)	15	Banana (1 fruit)	90	Chicken (meat only, roast, 4 oz)	170
Carrot (3 oz)	20	Wine (4 oz)	90	Cheese, Cheddar (1½ oz)	175
Grapefruit (½ fruit = 6 oz)	20	Beer (½ pint)	90	Avocado (½ fruit)	190
Milk (1 oz, in tea)	20	Butter (½ oz, for slice of bread)	105	Rice (2 oz before cooking)	205
Sugar (1 level teaspoon)	30	Cornflakes (1 oz)	105	Macaroni (2 oz before cooking)	210
Jam (½ oz)	35	Fish (cod, grilled, 4 oz)	110	Fish (fried in batter, 4 oz)	225
Orange juice (unsweetened, 4 oz)	45	Yogurt (low-fat, flavoured, 5 oz)	115	Beefsteak (grilled, 4 oz)	250
Apple (one, 5 oz)	50	Dates (2 oz)	140	Pork chops (7 oz, grilled, fat cut off)	265
Peas (4 oz, boiled)	60	Baked beans (8 oz)	145	Cookies, chocolate (2 oz)	300
Whisky (1 oz) and soda	65	Chocolate, milk (1 oz)	150	French fries (4 oz)	330

Source: A. Stewart Truswell *et al., ABC of Nutrition* (1986).

seen from such supplementation has been a severe sensory neuropathy of the extremities, including inability to walk. A daily intake of up to 100 milligrams is considered safe, although only 1 to 2 milligrams are required for good health.

Vitamin C and the common cold

Use of vitamin C supplements has been widespread since 1970, when chemist and Nobel laureate Linus Pauling suggested that the vitamin was protective against the common cold. Some studies have found a moderate benefit of vitamin C in reducing the duration and severity of common-cold episodes, but numerous studies have failed to find a significant effect on incidence. The most common side effect of high vitamin C intake is diarrhea and other gastrointestinal symptoms, likely due to the unabsorbed vitamin traversing the intestine. The safe intake of 2,000 milligrams a day is based on the avoidance of these gastrointestinal symptoms. Although other possible adverse effects of high vitamin C intake have been investigated, none has been demonstrated in healthy people.

High intake of vitamin D can lead to a variety of debilitating effects, notably calcification of soft tissues and cardiovascular and renal damage. Although not a concern for most people, young children are especially vulnerable.

Because of its function as an antioxidant, supplementation with large doses (several hundred milligrams per day) of vitamin E in hopes of protecting against heart disease and other chronic diseases has become widespread. Such doses—many times the amount normally found in food—appear safe for most people, but their effectiveness in preventing disease or slowing the aging process has not been demonstrated. Daily intakes greater than 1,000 milligrams are not advised because they may interfere with blood clotting, causing hemorrhagic effects.

Large doses of niacin, given for its cholesterol-lowering effect, may produce a reddening of the skin, along with burning, tingling, and itching. Known as a "niacin flush," this is the first indicator of niacin excess, and this symptom is the basis for the safe daily intake of 35 milligrams. Liver toxicity and other adverse effects have also been reported with several grams of niacin a day.

Minerals. A desirable dietary intake of the minerals generally falls in a fairly narrow range. Because of interactions, a high intake of one mineral may adversely affect the absorption or utilization of another. Excessive intake from food alone is unlikely, but consumption of fortified foods or supplements increases the chance of toxicity. Furthermore, environmental or occupational exposure to potentially toxic levels of minerals presents additional risks.

Widespread calcium supplementation, primarily by children who do not drink milk and by women hoping to prevent osteoporosis, has raised concerns about possible adverse consequences of high calcium intake. A major concern has been kidney stones (nephrolithiasis), the majority of which are composed of a calcium oxalate compound. For years, a low-calcium diet was recommended for people at risk of developing kidney stones, despite disappointing effectiveness and a fair amount of research challenging the approach. However, a recent study has provided strong evidence that a diet relatively low in sodium and animal protein with normal amounts of calcium (1,200 milligrams per day) is much more effective in preventing recurrent stone formation than was the traditional low-calcium diet. In fact, dietary calcium may be protective against kidney stones because it helps bind oxalate in the intestine. Constipation is a common side effect of high calcium intake, but daily consumption of up to 2,500 milligrams is considered safe for adults and for children at least one year old.

High-dose iron supplements, commonly used to treat iron deficiency anemia, may cause constipation and other gastrointestinal effects. Acute toxicity and death from ingestion of iron supplements is a major poisoning hazard for young children. In people with hereditary hemochromatosis, a disease characterized by the overabsorption of iron, or in those who have repeated blood transfusions, iron can build up to dangerous levels, leading to severe organ damage, particularly of the liver and heart. It is considered prudent for men and postmenopausal women to avoid iron supplements and high iron intakes from fortified foods.

Hereditary hemochromatosis

Excess zinc has been reported to cause gastrointestinal symptoms such as nausea and vomiting. Chronic intake of large amounts of zinc may interfere with the body's utilization of copper, impair immune response, and reduce the level of high-density lipoprotein cholesterol (so-called good cholesterol). A safe intake of 40 milligrams of zinc daily is unlikely to be exceeded by food alone, although it may be exceeded by zinc lozenges or supplements, which are widely used despite a lack of data about their safety or efficacy.

While the teeth are developing and before they erupt, excess fluoride ingestion can cause mottled tooth enamel; however, this is only a cosmetic effect. In adults, excess fluoride intake is associated with joint pain and stiffness and, in extreme cases, crippling skeletal fluorosis. Even in communities where water supplies naturally provide fluoride levels several times higher than recommended, skeletal fluorosis is extremely rare. (J.W.)

FOOD SENSITIVITY

Some foods and drinks are more likely to provoke sensitivity than others and people exhibit varying sensitivities to foods. The most common reactions are gastrointestinal, in the form of indigestion, vomiting, or diarrhea; urticaria and angioneurotic edema; eczema; asthma; and migraine. Gastrointestinal effects are the result of sensitivity to the irritant effect of one of the ingredients in food. In sensitive people, these ingredients might relax the cardiac sphincter at the lower end of the esophagus so that acid regurgitates from the stomach; they might stimulate gastric acid secretion; or they might increase the motility of the intestine.

Common sensitivity reactions

When the onset of eczema is related to food it is the result of a true allergic mechanism, an abnormal immunological reaction by immunoglobulin E to one of the proteins of the food. Urticaria and asthma sometimes can be truly allergic reactions but it seems likely that other mechanisms are more usual. Urticaria in some people is due to naturally occurring salicylates in plant foods. It is thought that they act by inhibiting the enzyme cyclo-oxygenase and allowing more arachidonic acid to be converted to leukotrienes by the action of the enzyme lipoxygenase. Whether hyperactivity in children (usually boys) is aggravated by naturally occurring food salicylates or by some artificial food colourings is controversial. There is little objective confirmation of this belief.

A minority of sufferers from migraine are able to incriminate particular foods as provoking agents. Cheese, chocolate, citrus fruits, and alcoholic drinks are the most likely agents, possibly via pressor amines like tyramine and histamine, which tend to raise blood pressure.

There are other more specific examples of food sensitivity, the most important of which is celiac disease, or gluten enteropathy. There is malabsorption, fat in the feces (steatorrhea), weight loss, impaired growth in children, and other manifestations of malnutrition. Celiac disease is due to atrophy (flattening of the cells) of the inner lining of the small intestine. This condition was first described in 1888, but it was not until 1953 that doctors in The Netherlands discovered that a fraction of wheat, the gluten, acts as a toxin in these patients. When wheat is meticulously eliminated from the diet, patients recover. This is one therapeutic diet that is literally lifesaving.

Gluten sensitivity

When symptoms of food sensitivity are immediate, the individual makes the diagnosis. There are some people, for instance, whose lips swell the moment their mouth touches a peanut, and if by chance they should swallow some they will have several systemic symptoms—vomiting, urticaria, even asthma. When the symptoms are more delayed, diagnosis of the responsible food or food component requires more detective work. If the symptoms are chronic or recur frequently, they should disappear on an elimination diet, a diet of very few foods (lamb, rice, carrots, lettuce, refined seed oil, sugar, and water) that do not cause food sensitivity. Suspected foods then can be reintroduced one at a time. Skin tests are not very reliable.

Not all people who believe they are sensitive to a particular food really are. They may in fact dislike the food, or they may have eaten it coincidentally with the onset of an illness, or they may have been told by an unorthodox practitioner that their symptoms are due to a particular food.

HABITUAL DIETARY PATTERNS AND DISEASE

A major preoccupation of nutrition research and a theme of the public debate over nutrition in North America, Europe, and Australasia is the relationship between the usual diet and the chances of developing (prematurely) one of the prevalent chronic degenerative diseases—coronary heart disease, high blood pressure, cancer, diabetes mellitus, dental caries, gallstones, urinary tract stones, colonic diverticulosis, cirrhosis of the liver, etc. Only the first five of these will be discussed below.

The causation of each of these degenerative diseases is multifactorial; there are several factors that act over a long period of time and interact with one another. There are protective factors as well as pathogenic factors; there are genetic, dietary, and non-dietary factors. Even in the diet there often appear to be multiple factors operating in different ways. The evidence for each disease comes from a variety of sources and several disciplines; from descriptive epidemiology—cross-sectional and retrospective; from prospective (cohort) epidemiological studies; from basic science (biochemistry, microbiology, pathology); from animal models; from human metabolic studies; and from prevention trials (primary or secondary).

Multiple causes

Diet and coronary heart disease. In a prospective study, a large group of people at moderate risk of developing coronary heart disease (usually middle-aged Caucasian men) is examined and blood tests performed. The group is then followed up for five to 20 years, during which time some develop one of the manifestations of coronary heart disease, while most do not. Coronary heart disease may or may not show signs and symptoms of disease. When symptoms are present there are four possible manifestations: sudden cardiac death, myocardial infarction (the typical heart attack, requiring hospitalization and a period of time off work), angina pectoris (heart pain on exertion), and cardiac failure with an electrocardiogram indicating that the cause is coronary artery insufficiency. Those who develop clinical coronary heart disease are examined with the battery of findings at the initial examination to see which characteristics are "risk factors"; *i.e.,* strongly associated with the disease and appearing to be predictive. Over 20 prospective studies have been made on thousands of people in 14 countries. They have all found that the three biggest risk factors for coronary heart disease are a high plasma total cholesterol, cigarette smoking, and high blood pressure (hypertension).

Being lipid and nonpolar, cholesterol is totally insoluble in an aqueous solution. All cholesterol in blood plasma, therefore, is carried on lipoproteins. Low-density lipoprotein (LDL) cholesterol, which normally carries about three-fourths of the total cholesterol, is the risk factor; it can, in some situations, carry cholesterol into the inner lining of artery walls. High-density lipoprotein (HDL) cholesterol, on the contrary, tends to act as a protective factor; it may be able to clear some of the cholesterol deposited in the periphery and carry it to the liver, where it can be excreted with the bile. Deposition of cholesterol on the inner aspect of the walls of arteries is part of the process of formation of atherosclerosis, the pathological basis of coronary heart disease. Plasma HDL cholesterol only makes up about one-quarter of normal total cholesterol concentration, and its level is higher in women than in men, but never very high. A total cholesterol that is high nearly always results from elevation of LDL cholesterol; when very high it is always due to raised LDL cholesterol. It is a little simpler and cheaper to measure total cholesterol than LDL cholesterol. When discussing elevation of cholesterol and its treatment the two terms are almost synonymous. A change of plasma total cholesterol means the same sort of change in LDL cholesterol.

Lipoproteins

Plasma (total or LDL) cholesterol may be raised as the inherited condition familial hypercholesterolemia, a condition with a tendency toward premature coronary heart disease. Plasma cholesterol is also raised secondary to certain diseases; *e.g.,* hypothyroidism, some types of kidney disease, bile duct obstruction, and diabetes mellitus. Finally, it is moderately raised by a diet rich in saturated fat and cholesterol, a phenomenon first demonstrated in rabbits just before World War I. From a global viewpoint such countries as Great Britain, Finland, Ireland, New Zealand, the Czech Republic, Slovakia, Australia, Hungary, and the United States are high plasma cholesterol countries, consuming high-fat diets and experiencing high incidences of coronary heart disease. In residents of these countries plasma cholesterol (total and LDL) can be consistently reduced by combining the following dietary changes (listed in order of effectiveness): (1) reduce saturated fat as much as possible, (2) reduce energy intake if overweight, (3) increase consumption of vegetables, legumes, fruits, and cereals (preferably whole-grain), (4) partly replace the reduced saturated fat with a moderate increase of polyunsaturated and some monounsaturated oils, and (5) reduce dietary cholesterol. In practical dietetics these translate to avoiding butter, hydrogenated margarine, lard and suet, full-cream milk, cream, meat fat, fried foods, chocolate, and sausages; to restricting egg yolks, most cheeses, cakes, pastries, and cookies; and to eating more vegetables, legumes, fruits, bread, rice, pasta, breakfast cereals, polyunsaturated oils (*e.g.,* sunflower seed, safflower, corn oils), and skim or low-fat milk.

Reduction of plasma cholesterol

Plasma total (and LDL) cholesterol is related to the risk of coronary heart disease in a continuous gradient. The higher the plasma cholesterol above about 4.9 millimoles per litre (190 milligrams per 100 millilitres), the greater the chances of subsequent coronary heart disease (Figure 4). Trials in people starting with moderately elevated plasma cholesterol have shown that the amount of coronary heart disease is less in subsequent years in those given a cholesterol-lowering drug compared with untreated controls. A World Health Organization expert committee concluded in 1982 that there is a causal relationship between plasma total cholesterol and coronary heart disease. In high-incidence countries, the committee recommended lowering the whole population's plasma cholesterol distribution by progressive changes in eating patterns toward the sort of diet described in the preceding paragraph. The United States National Heart, Lung, and Blood Institute and the British Department of Health, as well as heart associations or foundations in other countries, have reached essentially the same conclusions and recommendations. (A.S.T.)

Cholesterol and heart disease

Diet and hypertension. Hypertension, or high blood pressure, is one of the most common health problems in

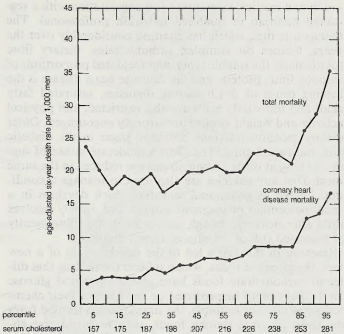


Figure 4: Relation between serum (total) cholesterol and subsequent death rate from coronary heart disease in a prospective study of more than 361,000 men screened.
From M.J. Martin *et al.*, *The Lancet*, vol. 11 for 1986, pp. 933–936

developed countries. It is an important risk factor for other diseases, such as coronary heart disease, congestive heart failure, and stroke. Most people with high blood pressure have essential, or primary, hypertension, for which no specific cause can be determined. Heredity plays a role in the development of the disease, but so do modifiable factors such as excess weight, physical inactivity, high alcohol intake, and diets high in salt. For reasons that are not entirely clear, African Americans have among the highest rates of hypertension in the world.

Hypertension is usually defined as a blood pressure greater than 140/90 mm Hg, *i.e.*, equivalent to the pressure exerted by a column of mercury 140 millimetres high during contraction of the heart (systole) and 90 millimetres high during relaxation (diastole); either systolic or diastolic blood pressure, or both, may be elevated in hypertension. Individuals with hypertension can be asymptomatic for years and then suddenly experience a fatal stroke or heart attack. Prevention and management of hypertension can significantly decrease the chance of complications. Early identification of hypertension is important so that lifestyle modification can begin as early as possible.

Overweight people have a much greater risk of developing hypertension than do lean people. Weight loss alone, sometimes as little as 4.5 kilograms (10 pounds), can be effective in reducing high blood pressure. Increasing physical activity can, of course, help with weight control, but it also appears to lower blood pressure independently.

Sodium intake

Blood pressure is positively related to dietary intake of sodium (salt). Primitive societies in which sodium intake is low have very little hypertension, and the increase in blood pressure that typically occurs with age in industrialized societies fails to occur. On the other hand, in countries with extremely high salt consumption, hypertension is common and stroke is a leading cause of death. Furthermore, experimental studies indicate that decreasing sodium intake can reduce blood pressure.

Some people appear to be genetically sensitive to salt. Although salt restriction may only help lower blood pressure in those who are salt sensitive, many individuals consume more salt than is needed. Dietary recommendations typically encourage the general population to limit sodium intake to no more than 2,400 milligrams daily, which amounts to a little more than a teaspoon of salt. This level can be achieved by restricting salt used in cooking, not adding salt at the table, and limiting highly salted foods, processed foods (many of which have hidden sodium), and so-called fast foods.

Heavy alcohol consumption (more than two drinks a day) is associated with hypertension. Vegetarians, and particularly vegans (who consume no foods of animal origin), tend to have lower blood pressures than do meat eaters. The diet recommended for reducing blood pressure emphasizes fruits, vegetables, and low-fat dairy products; includes whole grains, poultry, fish, and nuts; and contains only small amounts of red meat and sugary foods and beverages. Reducing salt intake should further increase the effectiveness of the diet.

Hypertension is treated by a variety of drugs, some of which have nutritional repercussions. Thiazide diuretics, for example, increase potassium loss from the body, usually necessitating the intake of additional potassium, which is particularly plentiful in foods such as bananas, citrus fruits, vegetables, and potatoes. Use of potassium-based salt substitutes is not advised without medical supervision.

Diet and cancer. Second only to cardiovascular disease as a cause of death in much of the world, cancer is the major killer of adults between ages 45 and 64.

Studies of identical twins reveal that, even for those with an identical genetic makeup, the risk for most cancers is still largely related to environmental factors. Another line of evidence supporting the limited role of heredity in most cancers is found in studies of migrant populations, which show that cancer rates tend to grow more like those of a group's adopted country with each passing generation. For example, rates of breast and colorectal cancers in individuals who migrate from rural Asia to the United States gradually increase to match the higher cancer rates of the United States. By contrast, rates of stomach cancer among Japanese migrants gradually decrease after they have moved to the United States. Nutrition is among the critical environmental and lifestyle factors investigated in migration studies, although identifying specific dietary components that affect the changing disease rates has been more elusive. A number of cancer organizations around the world have estimated that 30 to 40 percent of all cases of cancer could be prevented by dietary means.

Most cancer-causing substances (carcinogens) probably enter the body through the alimentary canal in food and beverages. Although some foodborne toxins, pesticides, and food additives may be carcinogenic if taken in sufficient quantity, it is primarily the foodstuffs themselves that are associated with cancer. Some dietary patterns or components may promote cancer, while others may inhibit it.

Studies attempting to relate total fat or specific types of fat to various cancers have been inconsistent. High intake of dietary fat may promote cancer, but this could be due at least in part to the extra energy (calories) that fat provides. Obesity is associated with several types of cancer, including colorectal, prostate, uterine, pancreatic, and breast cancers. A possible mechanism for this effect is the higher circulating levels of estrogen, insulin, and other hormones that accompany increased body fat. Furthermore, regular exercise has been shown in a number of studies to reduce the risk of breast and colon cancers. In laboratory animals, restricting energy intake is the most effective method for reducing cancer risk; chronic underfeeding inhibits the growth of many spontaneous tumours and most experimentally induced tumours.

Alcohol

High alcohol consumption is another factor that has been implicated in the development of various cancers, especially of the mouth, throat, liver, and esophagus (where it acts synergistically with tobacco) and probably of the breast, colon, and rectum. The fact that moderate use of alcohol has a beneficial effect on cardiovascular disease underscores how complex and potentially confusing is the connection between food and health.

Foods also contain substances that offer some protection against cancer. For example, fresh fruits and vegetables, and specifically vitamins C and E, eaten at the same time as nitrate-containing foods (such as ham, bacon, sausages, and luncheon meats), inhibit nitrosamine production and thus help protect against stomach cancer. More than 200 studies have found a strong association between diets high in vegetables and fruits and lower risk for various cancers, although identifying specific protective factors in such diets has been more difficult. Vitamin C, vitamin E, carotenoids such as β-carotene (a plant precursor of vitamin A), and the trace mineral selenium, act in the body's antioxidant systems to help prevent DNA damage by reactive mole-

cules known as free radicals. Specific vegetables, notably the cruciferous vegetables (broccoli, cauliflower, and other members of the cabbage family), contain sulforaphane and other compounds known as isothiocyanates, which induce enzymes that detoxify carcinogens and have been demonstrated to protect against cancer in animal studies. Dietary fibre in plant foods may also be protective: it dilutes potential carcinogens, binds to them, and speeds up transit time through the gut, thereby limiting exposure. Fruits and vegetables are rich in phytochemicals (biologically active plant substances), which are currently being investigated for potential anticarcinogenic activity. Animal studies suggest that antioxidant compounds known as polyphenols, which are found in both black and green tea, may be protective against the growth of cancer. Regular consumption of tea, especially in Japan and China, where green tea is the preferred type, has been associated with a decreased risk of various cancers, especially stomach cancer, but the evidence has been conflicting.

The dietary approach most likely to reduce cancer risk is one that is rich in foods from plant sources, such as fruits, vegetables, whole grains, beans, and nuts; has a limited intake of fat; includes a balance of energy intake and physical activity to maintain a healthy body weight; and includes alcohol in moderation, if at all. Intake of carcinogenic compounds can also be reduced by trimming fat and removing burned portions from meat before eating.

Diabetes mellitus. Diabetes mellitus is a group of metabolic disorders of carbohydrate metabolism characterized by high blood glucose levels (hyperglycemia) and usually resulting from insufficient production of the hormone insulin (type 1 diabetes) or an ineffective response of cells to insulin (type 2 diabetes). Secreted by the pancreas, insulin is required to transport blood glucose into cells. Diabetes is an important risk factor for cardiovascular disease, as well as a leading cause of adult blindness. Other long-term complications include kidney failure, nerve damage, and lower-limb amputation due to impaired circulation.

Type 1 diabetes (formerly known as juvenile-onset or insulin-dependent diabetes) can occur at any age but often begins in late childhood with the pancreas failing to secrete adequate amounts of insulin. Type 1 diabetes has a strong genetic link, but most cases are the result of an autoimmune disorder, possibly set off by a viral infection, foreign protein, or environmental toxin. Although elevated blood sugar is an important feature of diabetes, sugar or carbohydrate in the diet is not the cause of the disease. Type 1 diabetes is managed by injections of insulin, along with small, regularly spaced meals and snacks that spread glucose intake throughout the day and minimize fluctuations in blood glucose.

Type 2 diabetes

Type 2 diabetes (formerly known as adult-onset or non-insulin-dependent diabetes) is the more common type of diabetes, constituting 90 to 95 percent of cases. With this condition, insulin resistance renders cells unable to admit glucose, which then accumulates in the blood. Although type 2 diabetes generally starts in middle age, it is increasingly reported in childhood, especially in obese children. Genetic susceptibility to this form of diabetes may not be expressed unless a person has excess body fat, especially abdominal obesity. Weight loss often helps to normalize blood glucose regulation, and oral antidiabetic agents may also be used. Lifestyle intervention (*e.g.,* diet and exercise) is highly effective in delaying or preventing type 2 diabetes in high-risk individuals.

Migration studies have shown that urbanization and the adoption of a Western diet and habits can dramatically increase the rate of type 2 diabetes. For example, a high prevalence of the disorder is seen in the Pima Indians of Arizona, who are sedentary and eat a high-fat diet, whereas prevalence is low in a closely related group of Pimas living a traditional lifestyle—physically active with a diet lower in fat, resulting in lower body weight—in a remote, mountainous region of Mexico. Type 2 diabetes is a serious health problem among Native Americans and other ethnic minorities in the United States. Worldwide, the prevalence of type 2 diabetes has increased sharply, along with the rise in obesity.

Specific treatment plans for diabetics are designed after individual medical assessment and consultation with a registered dietitian or qualified nutrition professional. The therapeutic diet, which has changed considerably over the years, focuses on complex carbohydrates, dietary fibre (particularly the soluble type), and regulated proportions of carbohydrate, protein, and fat. Because heart disease is the leading cause of death among diabetics, saturated fatty acids and *trans*-fatty acids are also restricted, and physical activity and weight control are strongly encouraged. Older dietary recommendations restricted sugar in the diabetic diet, but recent guidelines allow a moderate intake of sugars, so long as other carbohydrates are reduced in the same meal. Diet and exercise are also used to manage a condition known as gestational diabetes, which develops in a small percentage of pregnant women and usually resolves itself after delivery, though such women are subsequently at increased risk of developing type 2 diabetes.

Research in the 1990s led to the development of a new tool, the glycemic index, which reflects the finding that different carbohydrate foods have effects on blood glucose levels that cannot be predicted on the basis of their chemical structure. For example, the simple sugars formed from digestion of some starchy foods, such as bread, are absorbed more quickly and cause a faster rise in blood glucose than does table sugar (sucrose), fruit, or milk. In practical terms, however, if a carbohydrate food is eaten as part of a mixed meal, its so-called glycemic effect is less consequential. The glycemic index may prove to be a useful option for planning diabetic diets, but it in no way obviates the need for other established therapeutic practices, such as limiting total carbohydrate intake and managing body weight.

Tooth decay

Dental caries. Dental caries (tooth decay) is an infectious oral disease in which bacteria, primarily *Streptococcus mutans*, in the dental plaque metabolize simple sugars and other fermentable carbohydrates into acids that dissolve tooth enamel. Dental plaque (not to be confused with the lipid-containing plaque found in arteries) is a mass of bacteria and sticky polymers that shield the tooth from saliva and the tongue, thereby facilitating decay. All dietary forms of sugar, including honey and corn syrup, can cause tooth decay; fermentable carbohydrates in breads, cereals, and other grain products, as well as milk, fruits, and fruit juices, also have cariogenic (decay-causing) potential. Eating sugary or starchy foods between meals, especially sticky foods that stay on the teeth longer, increases the time that teeth are exposed to destructive acids. Artificial sweeteners are not cariogenic, and xylitol, a sugar alcohol used in some chewing gums, is even cariostatic, *i.e.,* it reduces new tooth decay by inhibiting plaque and suppressing decay-causing bacteria.

Fluoride is extremely effective at protecting tooth enamel from decay, especially while enamel is being formed in the jaws before the permanent teeth erupt. Fluoridation of water in communities where fluoride is not naturally high is a safe and effective public health measure. In areas without fluoridated water, fluoride supplements are recommended for children. Regular brushing and flossing of the teeth and gums, as well as rinsing the mouth after meals and snacks, are important measures that protect against dental caries.

(J.W.)

DIETARY GUIDELINES

Different formats for dietary goals and guidelines have been developed over the years as educational tools, grouping foods of similar nutrient content together to help facilitate the selection of a balanced diet. In the United States, the four food-group plan of the 1950s—which suggested a milk group, a meat group, a fruit and vegetable group, and a breads and cereals group as a basic diet—was replaced in the 1990s by the five major food groups of the Food Guide Pyramid (see Figure 5). This innovative visual display was introduced by the United States Department of Agriculture as a tool for helping the public cultivate a daily pattern of wise food choices, ranging from liberal consumption of grain products, as represented in the broad base of the pyramid, to sparing use of fats, oils, and sugary foods, as represented in the apex. The pyramid shape, in which the relative quantities of daily servings of the differ-

Food Guide Pyramid

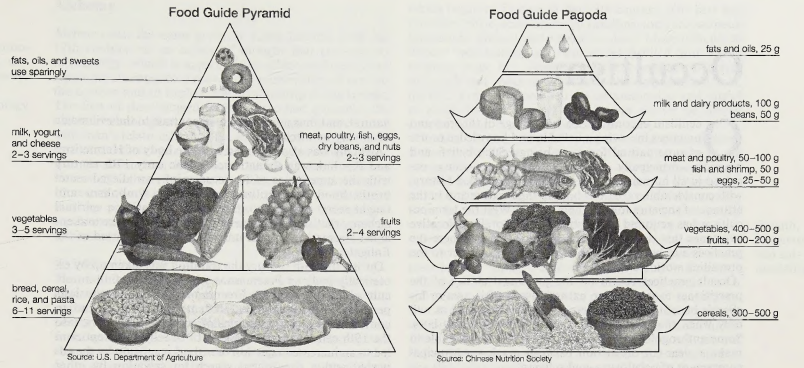


Figure 5: *Comparison of nutrition guidelines in the United States and China.*
(Left) The Food Guide Pyramid, developed by the United States Department of Agriculture, is similar to most other countries' guidelines in that it emphasizes grain, vegetables, and fruits. (Right) In contrast, the Chinese Nutrition Society's Food Guide Pagoda does not include sugar because it is consumed in relatively small amounts by the Chinese.
EB, Inc.

ent food groups can be easily grasped, reflects not only a new geometry of representation but also a new emphasis in nutrition. Subsequently, similar devices have been developed for particular cultural and ethnic food patterns such as Asian, Latin American, Mediterranean, and even vegetarian diets—all emphasizing grains, vegetables, and fruits. While an adaptation of the pyramid is used by Mexico, Chile, the Philippines, and Panama, a rainbow is used by Canada, a square by Zimbabwe, plates by Australia and the United Kingdom, a bean pot by Guatemala, the number 6 by Japan, and a pagoda by South Korea and China.

Dietary guidelines have been largely the province of more affluent countries, where correcting imbalances due to overconsumption and inappropriate food choices has been key. Not until 1989 were proposals for dietary guidelines published from the perspective of low-income countries, such as India, where the primary nutrition problems stem from the lack of opportunity to acquire or consume needed food. But even in such countries, there is a growing risk of obesity and chronic disease among the small but increasing number of affluent people who have adopted some of the dietary habits of the industrialized world. For example, the Chinese Dietary Guidelines, published by the Chinese Nutrition Society in 1997, make recommendations for that part of the population dealing with nutritional deficiency diseases and lack of food in some remote areas, as well as for the urban population that is particularly coping with changing lifestyles, dietary excess, and increasing rates of chronic disease. The Food Guide Pagoda (see Figure 5), a graphic display intended to help Chinese consumers put the dietary recommendations into practice, rests on the traditional cereal-based Chinese diet. Those who cannot tolerate fresh milk are encouraged to consume yogurt or other dairy products as a source of calcium. Unlike dietary recommendations in Western countries, the pagoda does not include sugar, as sugar consumption by the Chinese is quite low; however, children and adolescents in particular are cautioned to limit sugar intake because of the risk of dental caries. (J.W.)

BIBLIOGRAPHY

General works: Sources on animal and plant nutrition in general include P. MCDONALD *et al., Animal Nutrition,* 5th ed. (1995); LEE RUSSELL MCDOWELL, *Vitamins in Animal and Human Nutrition,* 2nd ed. (2000); I.H. BURGER and J.P.W. RIVERS (eds.), *Nutrition of the Dog and Cat* (1989); HORST MARSCHNER, *Mineral Nutrition of Higher Plants,* 2nd ed. (1995); and WALTER MERTZ (ed.), *Trace Elements in Human and Animal Nutrition,* 5th ed., 2 vol. (1986–87).

Human metabolism and nutrition: General information is presented in JAMES L. GROFF and SAREEN S. GROPPER, *Advanced Nutrition and Human Metabolism,* 3rd ed. (2000); GORDON M. WARDLAW and MARGARET W. KESSEL, *Perspectives in Nutrition,* 5th ed. (2002); L. KATHLEEN MAHAN and SYLVIA ESCOTT-STUMP (eds.), *Krause's Food, Nutrition, & Diet Therapy,* 10th ed. (2000); ELEANOR NOSS WHITNEY and SHARON RADY ROLFES, *Understanding Nutrition,* 9th ed. (2001); BARBARA A. BOWMAN and ROBERT M. RUSSELL (eds.), *Present Knowledge in Nutrition,* 8th ed. (2001); CAROLYN D. BERDANIER, *Handbook of Nutrition and Food* (2002); FRANCES SIENKIEWICZ SIZER and ELEANOR NOSS WHITNEY, *Nutrition: Concepts and Controversies,* 9th ed. (2002); MELVIN H. WILLIAMS, *Nutrition for Health, Fitness, and Sport,* 6th ed. (2002); A. STEWART TRUSWELL, *ABC of Nutrition,* 3rd ed. (1999); MARTHA H. STIPANUK (ed.), *Biochemical and Physiological Aspects of Human Nutrition* (1999); and BONNIE S. WORTHINGTON-ROBERTS and SUE RODWELL WILLIAMS (eds.), *Nutrition Throughout the Life Cycle,* 4th ed. (2000).

Foods: The nutritional values of foods are presented in ANNA DE PLANTER BOWES and HELEN NICHOLS CHURCH, *Bowes and Church's Food Values of Portions Commonly Used,* 17th ed., rev. by JEAN A. THOMPSON PENNINGTON (1998); and HELEN CHARLEY and CONNIE WEAVER, *Foods: A Scientific Approach* (1998).

Nutrition recommendations: Recommended intakes for nutrients are detailed in a series of reports by the INSTITUTE OF MEDICINE (U.S.): *Dietary Reference Intakes for Calcium, Phosphorus, Magnesium, Vitamin D, and Fluoride* (1997); *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Pantothenic Acid, Biotin, and Choline* (1998); *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids* (2000); *Dietary Reference Intakes for Vitamin A, Vitamin K, Arsenic, Boron, Chromium, Copper, Iodine, Iron, Manganese, Molybdenum, Nickel, Silicon, Vanadium, and Zinc* (2001); and *Dietary Reference Intakes: Applications in Dietary Assessment* (2001); these reports update and augment the NATIONAL RESEARCH COUNCIL (U.S.), *Recommended Dietary Allowances,* 10th ed. (1989). See also WORLD HEALTH ORGANIZATION, *Energy and Protein Requirements* (1985). Dietary guidelines in various countries are presented in JAMES PAINTER, J.H. RAH, and Y.K. LEE, "Comparison of International Food Guide Pictorial Representations," *Journal of the American Dietetic Association,* 102:438–489 (April 2002); UNITED STATES DEPARTMENT OF AGRICULTURE and DEPARTMENT OF HEALTH AND HUMAN SERVICES, *Nutrition and Your Health: Dietary Guidelines for Americans,* 5th ed. (2000).

Nutrition and disease: Comprehensive treatments of nutrition in health and disease are found in MAURICE A. SHILS *et al.* (eds.), *Modern Nutrition in Health and Disease,* 9th ed. (1999); and RICHARD D. SEMBA and MARTIN W. BLOEM (eds.), *Nutrition and Health in Developing Countries* (2001). (J.W./K.J.C./A.S.T.)

Occultism

Occultism encompasses a wide range of theories and practices involving a belief in and knowledge or use of supernatural forces or beings. Such beliefs and practices—principally magical or divinatory—have occurred in all human societies throughout recorded history, with considerable variations both in their nature and in the attitude of societies toward them. In the West the term occultism has acquired intellectually and morally pejorative overtones that do not obtain in other societies where the practices and beliefs concerned do not run counter to the prevailing worldview.

Occult practices centre on the presumed ability of the practitioner to manipulate natural laws for his own or his client's benefit; such practices tend to be regarded as evil only when they also involve the breaking of moral laws. Some anthropologists have argued that it is not possible to make a clear-cut distinction between magic—a principal component of occultism—and religion.

Those aspects of occultism that appear to be common to all human societies—divination, magic, witchcraft, and alchemy—are treated in depth below. Features that are unique to Western cultures, and the history of their development, are treated only briefly.

The Western tradition of occultism, as popularly conceived, is of an ancient "secret philosophy" underlying all occult practices. This secret philosophy derives ultimately from Hellenistic magic and alchemy on the one hand and from Jewish mysticism on the other. The principal Hellenistic source is the *Corpus Hermeticum,* the texts associated with Hermes Trismegistos, which are concerned with astrology and other occult sciences and with spiritual regeneration.

The Jewish element is supplied by the Kabbala (the doctrine of a secret, mystical interpretation of the Torah), which had been familiar to scholars in Europe since the Middle Ages, and which was linked with the Hermetic texts during the Renaissance. The resulting Hermetic-Kabbalistic tradition, known as Hermetism, incorporated both theory and magical practice, with the latter presented as natural, and thus good, magic, in contrast to the evil magic of sorcery or witchcraft.

Alchemy was also absorbed into the body of Hermetism, and this link was strengthened in the early 17th century with the appearance of Rosicrucianism, an alleged secret brotherhood that utilized alchemical symbolism and taught secret wisdom to its followers, creating a spiritual alchemy that survived the rise of empirical science and enabled Hermetism to pass unscathed into the period of the Enlightenment.

During the 18th century the tradition was taken up by esoterically inclined Freemasons who could not find an occult philosophy within Freemasonry. These enthusiasts persisted, both as individual students of Hermetism and, in continental Europe, as groups of occult practitioners, into the 19th century, when the growth of religious skepticism led to an increased rejection of orthodox religion by the educated and a consequent search for salvation by other means—including occultism.

But those interested turned to new forms of occultism rather than to the Hermetic tradition: on the one hand to Spiritualism—the practice of alleged regular communication between the living and the spirits of the dead through a living "medium"—and on the other to Theosophy—a blend of Western occultism and Eastern mysticism that proved to be a most effective propagator of occultism but whose influence has declined markedly over the last 50 years.

Indeed, despite the 19th-century revival, occult ideas have failed to gain acceptance in academic circles, although they have occasionally influenced the work of major artists, such as the poet William Butler Yeats and the painter Wassily Kandinsky, and occultism in Europe and North America seems destined to remain the province of popular culture.

(R.A.Gi.)

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Eight, Division I, Section 811; Division II, Section 829.

This article is divided into the following sections:

Alchemy 77
  Nature and significance 77
    The chemistry of alchemy
    Goals
  Regional variations 77
    Chinese alchemy
    Indian alchemy
    Hellenistic alchemy
    Arabic alchemy
    Latin alchemy
  Modern alchemy 80
  Assessments of alchemy 81
    Accomplishments
    Interpretations
Divination: astrology 81
  Nature and significance 81
  Historical development 82
    Astral omens in the ancient Middle East
    Astral omens in Egypt, Greece, India, China, and Islām
    Astrology in the Hellenistic period (3rd century BC to 3rd century AD)
    Astrology after the Hellenistic period
    Astrology in modern times
Divination: other forms 85
  Nature and significance 85
  The structure of divination 85
    Types of divination
    The function of divination
  Varieties of divination 87
    Inductive divination
    Interpretive divination
    Intuitive divination
    Divination today
Magic 89
  Nature and scope 89
  Elements of magic 89
    Spells and material
    Performance and functions
  Definitional issues: magic, religion, and science 89
    Dichotomies in context
    Subcategories of magic
  Conceptual history 90
    Ancient Mediterranean world
    Medieval and early modern Europe
    European traditions and the modern world
    Globalization of the magic concept and the postcolonial world
  History of magic theories 92
    Early theories
    Recent theories
  Conclusion 92
Witchcraft 92
  Definitions: witchcraft and sorcery 93
  Witchcraft in Africa and the world 93
    General perceptions
    Witchcraft and reason
    Moral issues
  The witch-hunts 95
    Creation of a stereotype
    Witchcraft, heresy, and the Devil
    Institutions of the witch-hunts
    Decline
    Modern theories
  Contemporary witchcraft 97
Bibliography 97

Alchemy

Relationship of alchemy and astrology

Alchemy was the name given, in Latin Europe, from the 12th century, to an aspect of thought that corresponds to astrology, which is apparently an older tradition. Both represent attempts to discover the relationship of man to the cosmos and to exploit that relationship to his benefit. The first of these objectives may be called scientific, the second technological. Astrology (see below) is concerned with man's relationship to "the stars" (including the members of the solar system); alchemy, with terrestrial nature. But the distinction is far from absolute, since both are interested in the influence of the stars on terrestrial events. Moreover, both have always been pursued in the belief that the processes human beings witness in heaven and on earth manifest the will of the Creator and, if correctly understood, will yield the key to the Creator's intentions.

NATURE AND SIGNIFICANCE

That both astrology and alchemy may be regarded as fundamental aspects of thought is indicated by their apparent universality. It is notable, however, that the evidence is not equally substantial in all times and places. Evidence from ancient Middle America (Aztecs, Mayans) is still almost nonexistent; evidence from India is tenuous and from ancient China, Greece, and Islāmic lands is only relatively more plentiful. A single manuscript of some 80,000 words is the principal source for the history of Greek alchemy. Chinese alchemy is largely recorded in about 100 "books" that are part of the Taoist canon. Neither Indian nor Islāmic alchemy has ever been collected, and scholars are thus dependent for their knowledge of the subject on occasional allusions in works of natural philosophy and medicine, plus a few specifically alchemical works.

Nor is it really clear what alchemy was (or is). The word is a European one, derived from Arabic, but the origin of the root word, *chem,* is uncertain. Words similar to it have been found in most ancient languages, with different meanings, but conceivably somehow related to alchemy. In fact, the Greeks, Chinese, and Indians usually referred to what Westerners call alchemy as "The Art," or by terms denoting change or transmutation.

The chemistry of alchemy. Superficially, the chemistry involved in alchemy appears a hopelessly complicated succession of heatings of multiple mixtures of obscurely named materials, but it seems likely that a relative simplicity underlies this complexity. The metals gold, silver, copper, lead, iron, and tin were all known before the rise of alchemy. Mercury, the liquid metal, certainly known before 300 BC, when it appears in both Eastern and Western sources, was crucial to alchemy. Sulfur, "the stone that burns," was also crucial. It was known from prehistoric times in native deposits and was also given off in metallurgic processes (the "roasting" of sulfide ores). Mercury united with most of the other metals, and the amalgam formed coloured powders (the sulfides) when treated with sulfur. Mercury itself occurs in nature in a red sulfide, cinnabar, which can also be made artificially. All of these, except possibly the last, were operations known to the metallurgist and were adopted by the alchemist.

The alchemist added the action on metals of a number of corrosive salts, mainly the vitriols (copper and iron sulfates), alums (the aluminum sulfates of potassium and ammonium), and the chlorides of sodium and ammonium. And he made much of arsenic's property of colouring metals. All of these materials, except the chloride of ammonia, were known in ancient times. Known as sal ammoniac in the West, *nao sha* in China, *nao sadar* in India, and *nushādir* in Persia and Arabic lands, the chloride of ammonia first became known to the West in the *Chou-i ts'an t'ung ch'i,* a Chinese treatise of the 2nd century AD. It was to be crucial to alchemy, for on sublimation it dissociates into antagonistic corrosive materials, ammonia and hydrochloric acid, which readily attack the metals. Until the 9th century it seems to have come from a single source, the Flame Mountain (Huo-yen Shan) near T'u-lu-p'an (Turfan), in Central Asia.

Finally, the manipulation of these materials was to lead to the discovery of the mineral acids, the history of which began in Europe in the 13th century. The first was probably nitric acid, made by distilling together saltpetre (potassium nitrate) and vitriol or alum. More difficult to discover was sulfuric acid, which was distilled from vitriol or alum alone but required apparatus resistant to corrosion and heat. And most difficult was hydrochloric acid, distilled from common salt or sal ammoniac and vitriol or alum, for the vapours of this acid cannot be simply condensed but must be dissolved in water.

Goals. "Transmutation" is the key word characterizing alchemy, and it may be understood in several ways: in the changes that are called chemical, in physiological changes such as passing from sickness to health, in a hoped-for transformation from old age to youth, or even in passing from an earthly to a supernatural existence. Alchemical changes seem always to have been positive, never involving degradation except as an intermediate stage in a process having a "happy ending." Alchemy aimed at the great human "goods": wealth, longevity, and immortality.

Wealth, longevity, and immortality

Alchemy was not original in seeking these goals, for it had been preceded by religion, medicine, and metallurgy. The first chemists were metallurgists, who were perhaps the most successful practitioners of the arts in antiquity. Their theories seem to have come not from science but from folklore and religion. The miner and metallurgist, like the agriculturalist, in this view, accelerate the normal maturation of the fruits of the earth, in a magico-religious relationship with nature. In primitive societies the metallurgist is often a member of an occult religious society.

But the first ventures into natural philosophy, the beginnings of what is called the scientific view, also preceded alchemy. Systems of five almost identical basic elements were postulated in China, India, and Greece, according to a view in which nature comprised antagonistic, opposite forces—hot and cold, positive and negative, and male and female; *i.e.,* primitive versions of the modern conception of energy. Drawing on a similar astrological heritage, philosophers found correspondences among the elements, planets, and metals. In short, both the chemical arts and the theories of the philosophers of nature had become complex before alchemy appeared.

REGIONAL VARIATIONS

Chinese alchemy. Neither in China nor in the West can scholars approach with certitude the origins of alchemy, but the evidences in China appear to be slightly older. Indeed, Chinese alchemy was connected with an enterprise older than metallurgy—*i.e.,* medicine. Belief in physical immortality among the Chinese seems to go back to the 8th century BC, and belief in the possibility of attaining it through drugs to the 4th century BC. The magical drug, namely the "elixir of life" (elixir is the European word), is mentioned about that time, and that most potent elixir, "drinkable gold," which was a solution (usually imaginary) of this corrosion-resistant metal, as early as the 1st century BC—many centuries before it is heard of in the West.

Although non-Chinese influences (especially Indian) are possible, the genesis of alchemy in China may have been a purely domestic affair. It emerged during a period of political turmoil, the Warring States Period (from the 5th to the 3rd century BC), and it came to be associated with Taoism—a mystical religion founded by the 6th-century-BC sage Lao-tzu—and its sacred book, the *Tao-te Ching* ("Classic of the Way of Power"). The Taoists were a miscellaneous collection of "outsiders"—in relation to the prevailing Confucians—and such mystical doctrines as alchemy were soon grafted onto the Taoist canon. What is known of Chinese alchemy is mainly owing to that graft, and especially to a collection known as *Yün chi ch'i ch'ien* ("Seven Tablets in a Cloudy Satchel"), which is dated 1023. Thus, sources on alchemy in China (as elsewhere) are compilations of much earlier writings.

The oldest known Chinese alchemical treatise is the *Chou-i ts'an t'ung ch'i* ("Commentary on the *I Ching*"). In the main it is an apocryphal interpretation of the *I Ching* ("Classic of Changes"), an ancient classic especially esteemed by the Confucians, relating alchemy to the mystical mathematics of the 64 hexagrams (six-line figures used for divination). Its relationship to chemical practice

is tenuous, but it mentions materials (including sal ammoniac) and implies chemical operations. The first Chinese alchemist who is reasonably well known was Ko Hung (AD 283–343), whose book *Pao-p'u-tzu* (pseudonym of Ko Hung) contains two chapters with obscure recipes for elixirs, mostly based on mercury or arsenic compounds. The most famous Chinese alchemical book is the *Tan chin yao chüeh* ("Great Secrets of Alchemy"), probably by Sun Ssu-miao (AD 581–after 673). It is a practical treatise on creating elixirs (mercury, sulfur, and the salts of mercury and arsenic are prominent) for the attainment of immortality, plus a few for specific cures for disease and such other purposes as the fabrication of precious stones.

Objective of Chinese alchemy

Altogether, the similarities between the materials used and the elixirs made in China, India, and the West are more remarkable than are their differences. Nonetheless, Chinese alchemy differed from that of the West in its objective. Whereas in the West the objective seems to have evolved from gold to elixirs of immortality to simply superior medicines, neither the first nor the last of these objectives seems ever to have been very important in China.

Chinese alchemy was consistent from first to last, and there was relatively little controversy among its practitioners, who seem to have varied only in their prescriptions for the elixir of immortality or perhaps only over their names for it, of which one Sinologist has counted about 1,000. In the West there were conflicts between advocates of herbal and "chemical" (*i.e.*, mineral) pharmacy, but in China mineral remedies were always accepted. There were, in Europe, conflicts between alchemists who favoured gold making and those who thought medicine the proper goal, but the Chinese always favoured the latter. Since alchemy rarely achieved any of these goals, it was an advantage to the Western alchemist to have the situation obscured, and the art survived in Europe long after Chinese alchemy had simply faded away.

Chinese alchemy followed its own path. Whereas the Western world, with its numerous religious promises of immortality, never seriously expected alchemy to fulfill that goal, the deficiencies of Chinese religions in respect to promises of immortality left that goal open to the alchemist. A serious reliance on medical elixirs that were in varying degrees poisonous led the alchemist into permanent exertions to moderate those poisons, either through variation of the ingredients or through chemical manipulations. The fact that immortality was so desirable and the alchemist correspondingly valued enabled the British historian of science Joseph Needham to tabulate a series of Chinese emperors who probably died of elixir poisoning. Ultimately a succession of royal deaths made alchemists and emperors alike more cautious, and Chinese alchemy vanished (probably as the Chinese adopted Buddhism, which offered other, less dangerous avenues to immortality), leaving its literary manifestations embedded in the Taoist canons.

Indian alchemy. The oldest Indian writings, the Vedas (Hindu sacred scriptures), contain the same hints of alchemy that are found in evidence from ancient China, namely vague references to a connection between gold and long life. Mercury, which was so vital to alchemy everywhere, is first mentioned in the 4th- to 3rd-century-BC *Artha-śāstra*, about the same time it is encountered in China and in the West. Evidence of the idea of transmuting base metals to gold appears in 2nd- to 5th-century-AD Buddhist texts, about the same time as in the West. Since Alexander the Great had invaded India in 325 BC, leaving a Greek state (Gandhāra) that long endured, the possibility exists that the Indians acquired the idea from the Greeks, but it could have been the other way around.

Medical and religious concerns of Indian alchemy

It is also possible that the alchemy of medicine and immortality came to India from China, or vice versa; in any case, gold making appears to have been a minor concern, and medicine the major concern, of both cultures. But the elixir of immortality was of little importance in India (which had other avenues to immortality). The Indian elixirs were mineral remedies for specific diseases or, at the most, to promote long life.

As in China and the West, alchemy in India came to be associated with religious mysticism, but much later—not until the rise of Tantrism (an esoteric, occultic, meditative system), AD 1100–1300. To Tantrism are owed writings that are clearly alchemical (such as the 12th-century *Rasārṇava*, or "Treatise on Metallic Preparations").

From the earliest records of Indian natural philosophy, which date from the 5th–3rd centuries BC, theories of nature were based on conceptions of material elements (fire, wind, water, earth, and space), vitalism ("animated atoms"), and dualisms of love and hate or action and reaction. The alchemist coloured metals and on occasion "made" gold, but he gave little importance to that. His six metals (gold, silver, tin, iron, lead, and copper), each further subdivided (five kinds of gold, etc.), were "killed" (*i.e.*, corroded) but not "resurrected," as was the custom of Western alchemy. Rather, they were killed to make medicines. Although "the secrets of mercurial lore" became part of the Tantric rite, mercury seems to have been much less important than in China. The Indians exploited metal reactions more widely, but, although they possessed from an early date not only vitriol and sal ammoniac but also saltpetre, they nevertheless failed to discover the mineral acids. This is the more remarkable because India was long the principal source of saltpetre, which occurs as an efflorescence on the soil, especially in populous tropical countries. But it lacks the high degree of corrosivity of metals possessed by the vitriols and chlorides and played a small part in early alchemy. Saltpetre appears particularly in 9th- to 11th-century-AD Indian and Chinese recipes for fireworks, one of which—a mixture of saltpetre, sulfur, and charcoal—is gunpowder. Saltpetre first appears in Europe in the 13th century, along with the modern formula for gunpowder and the recipe for nitric acid.

Hellenistic alchemy. Western alchemy may go back to the beginnings of the Hellenistic period (*c.* 300 BC–*c.* AD 300), although the earliest alchemist whom authorities have regarded as authentic is Zosimos of Panopolis (Egypt), who lived near the end of the period. He is one of about 40 authors represented in a compendium of alchemical writings that was probably put together in Byzantium (Constantinople) in the 7th or 8th century AD and that exists in manuscripts in Venice and Paris. Synesius, the latest author represented, lived in Byzantium in the 4th century. The earliest is the author designated Democritus but identified by scholars with Bolos of Mende, a Hellenized Egyptian who lived in the Nile Delta about 200 BC. He is represented by a treatise called *Physica et mystica* ("Natural and Mystical Things"), a kind of recipe book for dyeing and colouring but principally for the making of gold and silver. The recipes are stated obscurely and are justified with references to the Greek theory of elements and to astrological theory. Most end with the phrase "One nature rejoices in another nature; one nature triumphs over another nature; one nature masters another nature," which authorities variously trace to the Magi (Zoroastrian priests), Stoic pantheism (a Greek philosophy concerned with nature), or to the 4th-century-BC Greek philosopher Aristotle. It was the first of a number of such aphorisms over which alchemists were to speculate for many centuries.

The importance of the *Physica et mystica*

In 1828 a group of ancient papyrus manuscripts written in Greek was purchased in Thebes (Egypt), and about a half-century later it was noticed that among them, divided between libraries in Leyden (The Netherlands) and Stockholm, was a tract very like the *Physica et mystica*. It differed, however, in that it lacked the former's theoretical embellishments and stated in some recipes that only fraudulent imitation of gold and silver was intended. Scholars believe that this kind of work was the ancestor both of the *Physica et mystica* and of the ordinary artist's recipe book. The techniques were ancient. Archaeology has revealed metal objects inlaid with colours obtained by grinding metals with sulfur, and Homer's description (8th century BC) of the shield of Achilles gives the impression that the artist in his time was virtually able to paint in metal.

Democritus is praised by most of the other authors in the Venice–Paris manuscript, and he is much commented upon. But only Zosimos shows what had become of alchemy after Bolos of Mende. His theory is luxuriant in imagery, beginning with a discussion of "the composition

of waters, movement, growth, embodying and disembodying, drawing the spirits from bodies and binding the spirits within bodies" and continuing in the same vein. The "base" metals are to be "ennobled" (to gold) by killing and resurrecting them, but his practice is full of distillation and sublimation, and he is obsessed with "spirits." Theory and practice are joined in the concept that success depends upon the production of a series of colours, usually black, white, yellow, and purple, and that the colours are to be obtained through *Theion hydôr* (divine or sulfur water—it could mean either).

Zosimos credits these innovations mainly to Maria (sometimes called "the Jewess"), who invented the apparatus, and to Agathodaimon, probably a pseudonym. Neither is represented (beyond Zosimos' references) in the Venice–Paris manuscript, but a tract attributed to Agathodaimon, published in 1953, shows him to be preoccupied with the colour sequence and complicating it by using arsenic instead of sulfur. Thus, the colour-producing potentialities of chemistry were considerable by the time of Zosimos.

Zosimos also shows that alchemical theory came to focus on the idea that there exists a substance that can bring about the desired transformation instantly, magically, or, as a modern chemist might say, catalytically. He called it "the tincture," and had several. It was also sometimes called "the powder" (*xêrion*), which was to pass through Arabic into Latin as elixir and finally (signifying its inorganic nature) as the "philosopher's stone," "a stone which is not a stone," as the alchemists were wont to say. It was sometimes called a medicine for the rectification of "base" or "sick" metals, and from this it was a short step to view it as a drug for the rectification of human maladies. Zosimos notes the possibility, in passing. When the objective of alchemy became human salvation, the material constitution of the elixir became less important than the incantations that accompanied its production. Synesius, the last author in the Venice–Paris manuscript, already defined alchemy as a mental operation, independent of the science of matter.

Thus, Greek alchemy came to resemble, in both theory and practice, that of China and India. But its objectives included gold making; thus it remained fundamentally different.

The significance of Hermetic literature

Arabic alchemy. Arabic alchemy is as mysterious as Greek in its origins, and the two seem to have been significantly different. The respect in which *Physica et mystica* was held by the Greek alchemists was bestowed by the Arabs on a different work, the *Emerald Tablet* of Hermes Trismegistos, the reputed Hellenistic author of various alchemical, occultic, and theological works. Beginning "That which is above is like to that which is below, and that which is below is like to that which is above," it is brief, theoretical, and astrological. Hermes "the thrice great" (Trismegistos) was a Greek version of the Egyptian god Thoth and the supposed founder of an astrological philosophy that is first noted in 150 BC. The *Emerald Tablet,* however, comes from a larger work called *Book of the Secret of Creation,* which exists in Latin and Arabic manuscripts and was thought by the Muslim alchemist ar-Rāzī to have been written during the reign of Caliph al-Ma'mūn (AD 813–833), though it has been attributed to the 1st-century-AD pagan mystic Apollonius of Tyana.

Some scholars have suggested that Arabic alchemy descended from a western Asiatic school and that Greek alchemy was derived from an Egyptian school. As far as is known, the Asiatic school was not Chinese or Indian. What is known is that Arabic alchemy was associated with a specific city in Syria, Harran, which seems to have been a fountainhead of alchemical notions. And it is possible that the distillation ideology and its spokeswoman, Maria—as well as Agathodaimon—represented the alchemy of Harran, which presumably migrated to Alexandria and was incorporated into the alchemy of Zosimos.

The existing versions of the *Book of the Secret of Creation* have been carried back only to the 7th or 6th century but are believed by some to represent much earlier writings, although not necessarily those of Apollonius himself. He is the subject of an ancient biography that says nothing about alchemy, but neither does the *Emerald Tablet* nor the rest of the *Book of the Secret of Creation.* On the other hand, their theories of nature have an alchemical ring, and the *Book* mentions the characteristic materials of alchemy, including, for the first time in the West, sal ammoniac. It was clearly an important book to the Arabs, most of whose eminent philosophers mentioned alchemy, although sometimes disapprovingly. Those who practiced it were even more interested in literal gold making than had been the Greeks. The most well-attested and probably the greatest Arabic alchemist was ar-Rāzī (*c.* 850–923/924), a Persian physician who lived in Baghdad. The most famous was Jābir ibn Ḥayyān, now believed to be a name applied to a collection of "underground writings" produced in Baghdad after the theological reaction against science. In any case, the Jābirian writings are very similar to those of ar-Rāzī.

Ar-Rāzī classified the materials used by the alchemist into "bodies" (the metals), stones, vitriols, boraxes, salts, and "spirits," putting into the latter those vital (and sublimable) materials, mercury, sulfur, orpiment and realgar (the arsenic sulfides), and sal ammoniac. Much is made of sal ammoniac, the reactive powers of which seem to have given Western alchemy a new lease on life. Ar-Rāzī and the Jābirian writers were really trying to make gold, through the catalytic action of the elixir. Both wrote much on the compounding of "strong waters," an enterprise that was ultimately to lead to the discovery of the mineral acids, but students have been no more able to find evidence of this discovery in the writings of the Arabic alchemists than in those of China and India. The Arabic strong waters were merely corrosive salt solutions.

Ar-Rāzī's writing represents the apogee of Arabic alchemy, so much so that students of alchemy have little evidence of its later reorientation toward mystical or quasi-religious objectives. Nor does it seem to have turned to medicine, which remained independent. But there was a tendency in Arabic medicine to give greater emphasis to mineral remedies and less to the herbs that had been the chief medicines of the earlier Greek and Arabic physicians. The result was a pharmacopoeia not of elixirs but of specific remedies that are inorganic in origin and not very different from the elixirs of ar-Rāzī. This new pharmacopoeia was taken to Europe by Constantine of Africa, a Baghdad-educated Muslim who died in 1087 as a Christian monk at Monte Cassino (Italy). The pharmacopoeia also appeared in Spain in the 11th century and passed from there to Latin Europe, along with the Arabic alchemical writings, which were translated into Latin in the 12th century.

The dissemination of the pharmacopoeia

Latin alchemy. In the 12th century the Christian West began to shed its habit of indifference or hostility to the secular literature of ancient and alien civilizations. Christian scholars were particularly attracted to Muslim Spain and Sicily and there made translations from both Arabic and Greek works, many of which were in some degree familiar, but some of which, including the literature of alchemy, were new.

The Greek alchemy of the Venice–Paris manuscript had much less impact than the work of ar-Rāzī and other Arabs, which emerged among the voluminous translations made in Spain about 1150 by Gerard of Cremona. By 1250 alchemy was familiar enough to enable such encyclopaedists as Vincent of Beauvais to discuss it fairly intelligibly, and before 1300 the subject was under discussion by the English philosopher and scientist Roger Bacon and the German philosopher, scientist, and theologian Albertus Magnus. To learn about alchemy was to learn about chemistry, for Europe had no independent word to describe the science of matter. It had been touched upon in works concerned with other forms of change—*e.g.,* the motion of projectiles, the aging of man, and similar Aristotelian concepts. On the practical side there were also artists' recipe books; but for the first time in the works of Bacon and Albertus Magnus change was discussed in a truly chemical sense, with Bacon treating the newly translated alchemy as a general science of matter for which he had great hopes.

The objective of European alchemy

But the more familiar alchemy became, the more clearly it was understood that gold making was the almost exclusive objective of alchemy, and Europeans proved no more

resistant to the lure of this objective than their Arabic predecessors. By 1350, alchemical tracts were pouring out of the scriptoria (monastic copying rooms), and the Europeans had even taken over the tradition of anonymity and false attribution. One authority wrote at length about supposed disagreements between two Arabs, Iahiae Abindinon and Geber Abinhaen, who were probably two versions of the name of Jābir ibn Ḥayyān. The most famous Jābirian work in Europe, *The Sum of Perfection,* is now thought to have been an original European composition. At about this time personal reminiscences of alchemists began to appear. Most famous was the Paris notary Nicolas Flamel (1330–1418), who claimed that he dreamed of an occult book, subsequently found it, and succeeded in deciphering it with the aid of a Jewish scholar learned in the mystic Hebrew writings known as the Kabbala. In 1382 Flamel claimed to have succeeded in the "Great Work" (gold making); certainly he became rich and made donations to churches.

Alchemy and Kabbalistic speculation

By 1300 alchemists had begun the discovery of the mineral acids, a discovery that occupied about three centuries between the first evidence of the new strong water (aqua fortis—*i.e.,* nitric acid) and the clear differentiation of the acids into three kinds: nitric, hydrochloric, and sulfuric. These three centuries saw prodigious efforts in European alchemy, for these spontaneously reactive and highly corrosive substances opened a whole new world of research. And yet, it was of little profit to chemistry, for the experiments were inhibited by the old objectives of separating the base metals into their "elements," concocting elixirs, and other traditional procedures.

The "water of life" (aqua vitae; *i.e.,* alcohol) was probably discovered a little earlier than nitric acid, and some physicians and a few alchemists turned to the elixir of life as an objective. John of Rupescissa, a Catalonian monk who wrote *c.* 1350, prescribed virtually the same elixirs for metal ennoblement and for the preservation of health. His successors multiplied elixirs, which lost their uniqueness and finally simply became new medicines, often for specific ailments. Medical chemistry may have been conceived under Islām, but it was born in Europe. It only awaited christening by its great publicist, Paracelsus (1493–1541), who was the sworn enemy of the malpractices of 16th-century medicine and a vigorous advocate of "folk" and "chemical" remedies. By the end of the 16th century, medicine was divided into warring camps of Paracelsians and anti-Paracelsians, and the alchemists began to move en masse into pharmacy.

Paracelsian pharmacy was to lead, by a devious path, to modern chemistry, but gold making still persisted, though methods sometimes differed. Salomon Trismosin, purported author of the *Splendor solis,* or "Splendour of the Sun" (published 1598), engaged in extensive visits to alchemical adepts (a common practice) and claimed success through "kabbalistic and magical books in the Egyptian language." The impression given is that many had the secret of gold making but that most of them had acquired it from someone else and not from personal experimentation. Illustrations, often heavily symbolic, became particularly important, those of *Splendor solis* being far more complex than the text but clearly exercising a greater appeal, even to modern students.

MODERN ALCHEMY

The possibility of chemical gold making was not conclusively disproved by scientific evidence until the 19th century. As rational a scientist as Sir Isaac Newton (1643–1727) had thought it worthwhile to experiment with it. The official attitude toward alchemy in the 16th to 18th century was ambivalent. On the one hand, The Art posed a threat to the control of precious metal and was often outlawed; on the other hand, there were obvious advantages to any sovereign who could control gold making. In "the metropolis of alchemy," Prague, the Holy Roman emperors Maximilian II (reigned 1564–76) and Rudolf II (reigned 1576–1612) proved ever-hopeful sponsors and entertained most of the leading alchemists of Europe.

This was not altogether to the alchemist's advantage. In 1595 Edward Kelley, an English alchemist and companion of the famous astrologer, alchemist, and mathematician John Dee, lost his life in an attempt to escape after imprisonment by Rudolf II, and in 1603 the elector of Saxony, Christian II, imprisoned and tortured the Scotsman Alexander Seton, who had been traveling about Europe performing well-publicized transmutations. The situation was complicated by the fact that some alchemists were turning from gold making not to medicine but to a quasi-religious alchemy reminiscent of the Greek Synesius. Rudolf II made the German alchemist Michael Maier a count and his private secretary, although Maier's mystical and allegorical writings were, in the words of a modern authority, "distinguished for the extraordinary obscurity of his style" and made no claim to gold making. Neither did the German alchemist Heinrich Khunrath (*c.* 1560–1601), whose works have long been esteemed for their illustrations, make such a claim.

Conventional attempts at gold making were not dead, but by the 18th century alchemy had turned conclusively to religious aims. The rise of modern chemistry engendered not only general skepticism as to the possibility of making gold but also widespread dissatisfaction with the objectives of modern science, which were viewed as too limited. Unlike the scientists of the Middle Ages and Renaissance, the successors of Newton and the great 18th-century French chemist Antoine-Laurent Lavoisier limited their objectives in a way that amounted to a renunciation of what many had considered the most important question of science, the relation of man to the cosmos. Those who persisted in asking these questions came to feel an affinity with the alchemists and sought their answers in the texts of "esoteric," or spiritual, alchemy (as distinct from the "exoteric" alchemy of the gold makers), with its roots in Synesius and other late Greek alchemists of the Venice–Paris manuscript.

This spiritual alchemy, or Hermetism, as its practitioners often prefer to call it, was popularly associated with the supposititious Rosicrucian brotherhood, whose so-called *Manifestoes* (author unknown; popularly ascribed to the German theologian Johann Valentin Andreä) had appeared in Germany in the early 17th century and had attracted the favourable attention not only of such reform-



Saturn, or *Mercurius senex,* being cooked in the bath until the spirit or white dove (*pneuma*) ascends. From Salomon Trismosin's *Splendor solis,* 1598. In the British Museum (MS, Harley 3469).

By courtesy of the trustees of the British Museum

ing alchemists as Michael Maier but also of many prominent philosophers who were disquieted by the mechanistic character of the new science. In modern times alchemy has become a focal point for various kinds of mysticism. The old alchemical literature continues to be scrutinized for evidence, because alchemical doctrine is claimed to have on more than one occasion come into the possession of man but always again been lost. Nor is its association with chemistry considered accidental. In the words of the famous 19th-century English spiritual alchemist Mary Anne Atwood,

Association of alchemy with chemistry

> Alchemy is an universal art of vital chemistry which by fermenting the human spirit purifies and finally dissolves it . . . Alchemy *is* philosophy; it is *the* philosophy, the finding of the Sophia in the mind.

ASSESSMENTS OF ALCHEMY

Accomplishments. The most persistent goals of alchemy have been the prolongation of life and the transmutation of base metals into gold. It appears that neither was accomplished, unless one credits alchemy with the consequences of modern chemotherapy and the cyclotron.

It has been said that alchemy can be credited with the development of the science of chemistry, a keystone of modern science. During the alchemical period the repertoire of known substances was enlarged (*e.g.*, by the addition of sal ammoniac and saltpetre), alcohol and the mineral acids were discovered, and the basis was laid from which modern chemistry was to rise. Historians of chemistry have been tempted to credit alchemy with laying this base while at the same time regarding alchemy as mostly "wrong." It is far from clear, however, that the basis of chemistry was in fact laid by alchemy rather than medicine. During the crucial period of Arabic and early Latin alchemy, it appears that innovation owed more to nascent medical chemistry than to alchemy.

But those who explore the history of the science of matter, where matter is considered on a wider basis than the modern chemist understands the term, may find alchemy more rewarding. Numerous Hermetic writers of previous centuries claimed that the aims of their art could yet be achieved—indeed, that the true knowledge had been repeatedly found and repeatedly lost. This is a matter of judgment, but it can certainly be said that the modern chemist has not attained the goal sought by the alchemist. For those who are wedded to scientific chemistry, alchemy can have no further interest. For those who seek the wider goal, which was also that of the natural philosopher before the advent of "mechanical," "Newtonian," or "modern" science, the search is still on.

Interpretations. Charlatanism was a prominent feature of European alchemy during the 16th century, and such monarchs as Rudolf II—even if they had mainly themselves to blame—were not entirely without reason in incarcerating some of their resident adepts. The picturesqueness of this era, which also saw the birth of the modern science of chemistry, has led many historians of chemistry to view alchemy in general as a fraud.

Other historians of chemistry have attempted to differentiate the good from the bad in alchemy, citing as good the discovery of new substances and processes and the invention of new apparatus. Some of this was certainly accomplished by alchemists (*e.g.*, Maria), but most of it is more justifiably ascribed to early pharmacists.

Modern Hermetic and psychological views

Scholars generally agree that alchemy had something to do with chemistry, but the modern Hermetic holds that chemistry was the handmaiden of alchemy, not the reverse. From this point of view the development of modern chemistry involved the abandonment of the true goal of the art.

Finally, a new interpretation was offered in the 1920s by the Swiss psychoanalyst Carl Jung, who, following the earlier work of the Austrian psychologist Herbert Silberer, judged alchemical literature to be explicable in psychological terms. Noticing the similarities between alchemical literature, particularly in its reliance on bizarre symbolic illustrations, and the dreams and fantasies of his patients, Jung viewed them as manifestations of a "collective unconscious" (inherited disposition). Jung's theory, still largely undeveloped, remains a challenge rather than an explanation. (R.P.M./R.A.Gi.)

Divination: astrology

Divination, the alleged art or science of foretelling the future by various natural, psychological, and other techniques, is a phenomenon found in all civilizations, both ancient and modern. In the context of ancient Latin language and belief, divination was concerned with discovering the will of the gods. Today, however, scholars no longer restrict the word to that earlier root meaning. Divinatory practices and the beliefs undergirding them are greater in scope than the ancient methods of discerning the will of the gods and the fatalistic view of the human condition that inspired so much of early Mediterranean religious thought. In some societies, in fact, divination is a practice to which many persons frequently resort, but never in terms of discovering the will of the gods. The idea of a godly providence controlling human affairs, in such societies, is unusual, although humbler spirits are often thought to intervene in troublesome ways. (G.K.P.)

The most significant form of divination is astrology, defined as the practice of forecasting earthly and human events through the observation and interpretation of the fixed stars and of the movement of the Sun, the Moon, and the planets. Devotees believe that an understanding of the influence of the planets and stars on earthly affairs allows them to both predict and affect the destinies of individuals, groups, and nations. Although often regarded as a science throughout its history, astrology is thought to be opposed to the findings and theories of modern Western science.

NATURE AND SIGNIFIGANCE

Astrology is a method of predicting mundane events based upon the assumption that the celestial bodies—particularly the planets and the stars considered in their arbitrary combinations or configurations (called constellations)—in some way either determine or indicate changes in the sublunar world. The theoretical basis for this assumption lies historically in Hellenistic philosophy and radically distinguishes astrology from the celestial *omina* ("omens") that were first categorized and cataloged in ancient Mesopotamia. Originally, astrologers presupposed a geocentric universe in which the "planets" (including the Sun and Moon) revolve in orbits whose centres are at or near the centre of the Earth, and in which the stars are fixed upon a sphere with a finite radius whose centre is also the centre of the Earth. Later, the principles of Aristotelian physics were adopted, according to which there is an absolute division between the eternal, circular motions of the heavenly element and the limited, linear motions of the four sublunar elements: fire, air, water, earth.

Special relations were believed to exist between particular celestial bodies and their varied motions, configurations with each other, and the processes of generation and decay apparent in the world of fire, air, water, and earth. These relations were sometimes regarded as so complex that no human mind could completely grasp them; thus, the astrologer might be readily excused for any errors. A similar set of special relations was also assumed by those whose physics was more akin to that of the Greek philosopher Plato. For the Platonic astrologers, the element of fire was believed to extend throughout the celestial spheres, and they were more likely than the Aristotelians to believe in the possibility of divine intervention in the natural processes through celestial influences upon the Earth, since they believed in the deity's creation of the celestial bodies themselves.

The role of the divine

The role of the divine in astrological theory varies considerably. In its most rigorous aspect, astrology postulates a totally mechanistic universe, denying to the deity the possibility of intervention and to man that of free will; as such it was vigorously attacked by orthodox Christianity and Islām. For some, however, astrology is not an exact science like astronomy but merely indicates trends and directions that can be altered either by divine or by human will. In the interpretation of Bardesanes, a Syrian Christian scholar (154–*c.* 222)—who has often been identified as a

Gnostic (a believer in esoteric salvatory knowledge and the view that matter is evil and spirit good)—the motions of the stars govern only the elemental world, leaving the soul free to choose between the good and the evil. Man's ultimate goal is to attain emancipation from an astrologically dominated material world. Some astrologers, such as the Harranians (from the ancient Mesopotamian city of Harran) and the Hindus, regard the planets themselves as potent deities whose decrees can be changed through supplication and liturgy or through theurgy, the science of persuading the gods or other supernatural powers. In still other interpretations—*e.g.,* that of the Christian Priscillianists (followers of Priscillian, a Spanish ascetic of the 4th century who apparently held dualistic views)—the stars merely make manifest the will of God to those trained in astrological symbolism.

Significance of astral omens. The view that the stars make manifest the divine will is closest to the concept that lies behind the ancient Mesopotamian collections of celestial omens. Their primary purpose was to inform the royal court of impending disaster or success. These might take the forms of meteorological or epidemic phenomena affecting entire human, animal, or plant populations. Frequently, however, they involved the military affairs of the state or the personal lives of the ruler and his family. Since the celestial *omina* were regarded not as deterministic but rather as indicative—as a kind of symbolic language in which the gods communicated with men about the future and as only a part of a vast array of ominous events—it was believed that their unpleasant forebodings might be mitigated or nullified by ritual means or by contrary omens. The *bāru* (the official prognosticator), who observed and interpreted the celestial *omina,* was thus in a position to advise his royal employer on the means of avoiding misfortunes; the omens provided a basis for intelligent action rather than an indication of an inexorable fate.

Purposes of astrology. The original purpose of astrology, on the other hand, was to inform the individual of the course of his life on the basis of the positions of the planets and of the zodiacal signs (the 12 astrological constellations) at the moment of his birth or conception. From this science, called genethlialogy (casting nativities), were developed the fundamental techniques of astrology.

Principal subdivisions of astrology

The main subdivisions of astrology that developed after genethlialogy are general, catarchic, and interrogatory.

General astrology studies the relationship of the significant celestial moments (*e.g.,* the times of vernal equinoxes, eclipses, or planetary conjunctions) to social groups, nations, or all humanity. It answers, by astrological means, questions formerly posed in Mesopotamia to the *bāru.*

Catarchic (pertaining to beginnings or sources) astrology determines whether or not a chosen moment is astrologically conducive to the success of a course of action begun in it. Basically in conflict with a rigorous interpretation of genethlialogy, it allows the individual (or corporate body) to act at astrologically favourable times and, thereby, to escape any failures predictable from his (or its) nativity.

Interrogatory astrology provides answers to a client's queries based on the situation of the heavens at the moment of his posing the questions. This astrological consulting service is even more remote from determinism than is catarchic astrology; it is thereby closer to divination by omens and insists upon the ritual purification and preparation of the astrologer.

Other forms of astrology, such as iatromathematics (application of astrology to medicine) and military astrology are variants on one or another of the above.

HISTORICAL DEVELOPMENT

Astral omens in the ancient Middle East. The astral omens employed in Mesopotamian divination were later commingled with what came to be known as astrology in the strict sense of the term and constituted within astrology a branch described as natural astrology. Though lunar eclipses apparently were regarded as ominous at a somewhat earlier period, the period of the 1st dynasty of Babylon (18th to 16th centuries BC) was the time when the cuneiform text *Enūma Anu Enlil,* devoted to celestial *omina,* was initiated. The final collection and codification of this series, however, was not accomplished before the beginning of the 1st millennium BC. But the tablets that have survived—mainly from the Assyrian library of King Ashurbanipal (7th century BC)—indicate that a standard version never existed. Each copy had its own characteristic contents and organization designed to facilitate its owner's consultation of the omens.

The four omen categories

The common categories into which the omens of *Enūma Anu Enlil* were considered to fall were four, named after the chief gods involved in the ominous communication: Sin, Shamash, Adad, and Ishtar. Sin (the Moon) contains omens involving such lunar phenomena as first crescents, eclipses, halos, and conjunctions with various fixed stars; Shamash (the Sun) deals with omens involving such solar phenomena as eclipses, simultaneous observations of two suns, and perihelia (additional suns); Adad (the weather god) is concerned with omens involving meteorological phenomena, such as thunder, lightning, and cloud formations, as well as earthquakes; and Ishtar (Venus) contains omens involving planetary phenomena such as first and last visibilities, stations (the points at which the planets appear to stand still), acronychal risings (rising of the planet in the east when the Sun sets in the west), and conjunctions with the fixed stars.

Though these omens are often cited in the reports of a network of observers established throughout the Assyrian Empire in the 7th century BC, they seem to have lost their popularity late in the period of the Persian domination of Mesopotamia (ending in the 4th century BC). During the later period new efforts were made, in a large number of works called *Diaries,* to find the correct correlations between celestial phenomena and terrestrial events. Before this development, however, portions of the older omen series were transmitted to Egypt, Greece, and India as a direct result of the Achaemenid domination (a dynasty ruling in Persia 559–330 BC) of these cultural areas or of their border regions.

Astral omens in Egypt, Greece, India, China, and Islām. The evidence for a transmission of lunar omens to Egypt in the Achaemenid period lies primarily in a demotic papyrus based on an original of about 500 BC. A more extensive use of Mesopotamian celestial omens is attested by the fragments of a book written in Greek in the 2nd century BC and claimed as a work addressed to a King Nechepso by the priest Petosiris. From this source, among others, the contents of *Enūma Anu Enlil* were included in the second book of the *Apotelesmatika,* or "Work on Astrology" (commonly called the *Tetrabiblos,* or "Four Books"), by Ptolemy, a Greek astronomer of the 2nd century AD; the first book of an astrological compendium, by Hephaestion of Thebes, a Greco-Egyptian astrologer of the 5th century AD; and the *On Signs* of John Lydus, a Byzantine bureaucrat of the 6th century. Yet another channel of transmission to the Greeks was through the Magusaeans of Asia Minor, a group of Iranian settlers influenced by Babylonian ideas. Their teachings are preserved in several classical works on natural history, primarily that of Pliny the Elder (*c.* AD 23–79), and the *Geoponica* (a late collection of agricultural lore).

In various Middle Eastern languages there also exist many texts dealing with celestial omens, though their sources and the question as to whether they are directly descended from a Mesopotamian tradition or are derived from Greek or Indian intermediaries is yet to be investigated. Of these texts the most important are those ascribed to Hermes Trismegistos by the Harranians and now preserved in Arabic, the *Book of the Zodiac* of the Mandaeans (a Gnostic sect still existing in Iraq and Khuzistan), the *Apocalypse,* attributed to the Old Testament prophet Daniel (extant in Greek, Syriac, and Arabic versions), and *The Book of the Bee* in Syriac.

Introduction of Mesopotamian literature to India

The transmission of Mesopotamian omen literature to India, including the material in *Enūma Anu Enlil,* apparently took place in the 5th century BC during the Achaemenid occupation of the Indus Valley. The first traces are found in Buddhist texts of this period, and Buddhist missionaries were instrumental in carrying this material to Central Asia, China, Tibet, Japan, and Southeast Asia. But the most important of the works of this In-

dian tradition and the oldest extant one in Sanskrit is the earliest version of the as yet unpublished *Gargasaṃhitā* ("Compositions of Garga") of about the 1st century AD. The original Mesopotamian material was modified so as to fit into the Indian conception of society, including the system of the four castes and the duty of the upper castes to perform the *saṃskāras* (sanctifying ceremonies).

There are numerous later compilations of omens in Sanskrit—of which the most notable are the *Bṛhatsaṃhitā*, or "Great Composition," of Varāhamihira (*c.* 550), the Jaina *Bhadrabāhu-saṃhitā*, or "Composition of Bhadrabāhu" (*c.* 10th century), and the *Pariśiṣṭas* ("Supplements") of the Atharvaveda (perhaps 10th or 11th century)—though these add little to the tradition. But in the works of the 13th century and later, entitled *Tājika*, there is a massive infusion of the Arabic adaptations of the originally Mesopotamian celestial omens as transmitted through Persian (*Tājika*) translations. In *Tājika* the omens are closely connected with general astrology; in the earlier Sanskrit texts their connections with astrology had been primarily in the fields of military and catarchic astrology.

Astrology in the Hellenistic period (3rd century BC to 3rd century AD). In the 3rd century BC and perhaps somewhat earlier, Babylonian diviners began—for the purpose of predicting the course of an individual's life—to utilize some planetary omens: positions relative to the horizon, latitudes, retrogressions, and other positions at the moment of birth or of computed conception. This method was still far from astrology, but its evolution was more or less contemporary and parallel with the development of the science of genethlialogy in Hellenistic Egypt.

Equally obscure are those individuals who, living in Egypt under the Ptolemies (a Greek dynasty ruling 305–30 BC), mathematicized the concept of a correspondence between the macrocosm (larger order, or universe) and the microcosm (smaller order, or man) as interpreted in terms of Platonic or Aristotelian theories concerning the Earth as the centre of the planetary system. They conceived of the ecliptic (the apparent orbital circle of the Sun) as being divided into 12 equal parts, or zodiacal signs, each of which consists of 30°; in this they followed the Babylonians. They further regarded each of these 12 signs as the domicile (or house) of a planet and subdivided each into various parts—decans of 10° each, *fines* ("bounds") of varying lengths, and *dōdecatēmoria* of 2°30′ each—each of which is also dominated by a planet. Scattered at various points throughout the ecliptic are the planets' degrees of exaltation (high influence), opposite to which are their degrees of dejection (low influence). Various arcs of the zodiac, then, are either primarily or secondarily subject to each planet, whose strength and influence in a geniture (nativity) depend partially on its position relative to these arcs and to those of its friends and enemies.

The zodiac

Furthermore, each zodiacal sign has a special relation with a part of the human body. The 12 signs are further divided into four triplicities, each of which governs one of the four elements. Numerous pairs of opposites (male–female, diurnal–nocturnal, hot–cold, and others), based on the speculations of the followers of Pythagoras, a Greek mystical philosopher of the 6th century BC, are connected with consecutive pairs of signs. Finally, a wide variety of substances in the elemental world and attributes of human character are more or less arbitrarily associated with the different signs. These lists of interrelationships provide the rationale for many of the astrologer's predictions.

The individual planet's influences are related both to its general indications when regarded as ominous in Mesopotamian texts and to the traits of its presiding deity in Greek mythology. But on them are also superimposed the system of the four elements and their four qualities, the Pythagorean opposites, and lists of sublunar substances. Furthermore, as in the omens, the modes of the planetary motions are carefully considered, since their strengths are partially determined by their phases with respect to the Sun. Also, they exert a mutual influence both by occupying each other's houses and by means of conjunction and aspects—opposition (to the seventh) and quartile (to the fourth or 10th) being generally considered bad, trine (to the fifth or ninth) and sextile (to the third or 11th) good.

Moreover, as the planetary orbits revolve from west to east, the zodiac rotates daily about the Earth in the opposite sense. From a given spot on the Earth's surface this latter motion—if the ecliptic were a visible circle—would appear as a succession of signs rising one after the other above the eastern horizon. Astrologers regard the one that is momentarily in the ascendant as the first place, the one to follow it as the second, and so on, with the one that rose immediately prior to the ascendant being the 12th. In genethlialogy each place in this *dōdecatropos* determines an aspect of the life of the native (one born under a particular sign); in other forms of astrology the place determines some appropriate aspect of the sublunar world.

Horoscopes

Astrologers, then, cast a horoscope by first determining for the given moment and locality the boundaries of the 12 places and the longitudes and latitudes of the seven planets. They read this horoscope by examining the intricate geometric interrelationships of the signs and their parts and of the planets of varying computed strengths with the places and each other and by associating with each element in the horoscope its list of sublunary correspondences. Any horoscopic diagram, of course, will yield a vast number of predictions, including many that are contradictory or extravagant. Astrologers thus must rely on their knowledge of the client's social, ethnic, and economic background and on their own experience to guide them in avoiding error and attaining credibility.

Since about 100 BC the above method has been the essential procedure of astrology, though various refinements and additional devices occasionally have been introduced, including those associated with the Hermetic tradition of Hermes Trismegistos and with Dorotheus of Sidon, an influential astrological poet of the third quarter of the 1st century AD. One is the system of lots, which are influential points as distant from some specified points in the horoscopic diagram as two planets are from each other. A second is the prorogator, a point on the ecliptic that, traveling at the rate of one degree of oblique ascension a year toward either the descendant or ascendant, determines a person's length of life. Another is the method of continuous horoscopy, under which anniversary diagrams are compared with the base nativity to provide annual readings. And, finally, certain periods of life are apportioned to their governing planets in a fixed sequence; these period governors in turn share their authority with the

Table 1: The Signs and Relationships of the Zodiac

sign	sex, nature	triplicity	house	decan		exaltation
				Greek	Indian	
Aries	masculine, moving	fire	Mars	Mars Sun Venus	Mars Sun Jupiter	Sun (19°)
Taurus	feminine, fixed	earth	Venus	Mercury Moon Saturn	Venus Mercury Saturn	Moon (3°)
Gemini	masculine, common	air	Mercury	Jupiter Mars Sun	Mercury Venus Saturn	
Cancer	feminine, moving	water	Moon	Venus Mercury Moon	Moon Mars Jupiter	Jupiter (15°)
Leo	masculine, fixed	fire	Sun	Saturn Jupiter Mars	Sun Jupiter Mars	
Virgo	feminine, common	earth	Mercury	Sun Venus Mercury	Mercury Saturn Venus	Mercury (15°)
Libra	masculine, moving	air	Venus	Moon Saturn Jupiter	Venus Saturn Mercury	Saturn (21°)
Scorpio	feminine, fixed	water	Mars	Mars Sun Venus	Mars Jupiter Moon	
Sagittarius	masculine, common	fire	Jupiter	Mercury Moon Saturn	Jupiter Mars Sun	
Capricorn	feminine, moving	earth	Saturn	Jupiter Mars Sun	Saturn Venus Mercury	Mars (28°)
Aquarius	masculine, fixed	air	Saturn	Venus Mercury Moon	Saturn Mercury Venus	
Pisces	feminine, common	water	Jupiter	Saturn Jupiter Mars	Jupiter Moon Mars	Venus (27°)

Table 2: Relationship of Positions in the Zodiac to Aspects of Life

place	*dōdecatropos* Greek	Indian
I	life	body
II	wealth	wealth
III	siblings	siblings
IV	parents	relatives
V	children	children
VI	health	enemies
VII	marriage	marriage
VIII	death	death
IX	travel, religion	religion
X	occupation, honors	occupation
XI	benefits, friends	gains
XII	losses, enemies	losses

other planets by granting them subperiods. All of these complications serve, critics argue, to provide astrologers with excuses for their errors.

Astrology after the Hellenistic period. *In India.* Greek astrology was transmitted to India in the 2nd and 3rd centuries AD by means of several Sanskrit translations, of which the one best known is that made in AD 149/150 by Yavaneśvara and versified as the *Yavanajātaka* by Sphujidhvaja in AD 269/270. The techniques of Indian astrology are thus not surprisingly similar to those of its Hellenistic counterpart. But the techniques were transmitted without their philosophical underpinnings (for which the Indians substituted divine revelation), and the Indians modified the predictions, originally intended to be applied to Greek and Roman society, so that they would be meaningful to them. In particular, they took into account the caste system, the doctrine of metempsychosis (transmigration of souls), the Indian theory of five elements (earth, water, air, fire, and space), and the Indian systems of values.

The Indians also found it useful to make more elaborate the already complex methodology of Hellenistic astrology. They added as significant elements: the *nakṣatras* (or lunar mansions); an elaborate system of three categories of *yogas* (or planetary combinations); dozens of different varieties of *daśās* (periods of the planets) and *antardaśās* (subperiods); and a complex theory of *aṣṭakavarga* based on continuous horoscopy. The number of subdivisions of the zodiacal signs was increased by the addition of the *horās* (15° each), the *saptāṃśas* (4¾° each), and the *navāṃśas* (3°20′ each); the number of planets was increased by the addition of the nodes of the Moon (the points of intersection of the lunar orbit with the ecliptic), and of a series of *upagrahas*, or imaginary planets. Several elements of Hellenistic astrology and its Sāsānian offshoot (see below *In Sāsānian Iran*), however—including the lots, the prorogator, the Lord of the Year, the triplicities, and astrological history—were introduced into India only in the 13th century through the *Tājika* texts. Besides genethlialogy, the Indians particularly cultivated military astrology and a form of catarchic astrology termed *muhūrta-śāstra* and, to a lesser extent, iatromathematics and interrogatory astrology.

In Sāsānian Iran. Shortly after Ardashīr I founded the Sāsānian empire in AD 226, a substantial transmission of both Greek and Indian astrology to Iran took place. There were Pahlavi (Iranian language) translations of Dorotheus of Sidon, Vettius Valens, Hermes, and an Indian called (in the Arabic sources) Farmasp. Since the Pahlavi originals are all lost, these translations provided the only knowledge of the Sāsānian science. Genethlialogy in Iran was essentially an imitation of the Hellenistic (though without any philosophy), onto which were grafted some Indian features, such as the *navāṃśas* and a Śaivite interpretation of illustrations of the Greco-Egyptian deities of the decans. The most influential and characteristic innovation of the Sāsānian astrologers was the development of the theory of astrological history—that is, the writing of history, both past and future, on the basis of extensions of the techniques of the prorogator, the Lord of the Year, the planetary periods, and the continuous horoscopy employed in Hellenistic genethlialogy. This was done in conjunction with Zoroastrian millenarianism (the division of the finite duration of the material creation into 12 millennia).

In Islām. Astrology entered Islāmic civilization in the 8th and 9th centuries in three simultaneous streams—Hellenistic, Indian, and Sāsānian. Arabic translations from the Greek and Syriac represented the Hellenistic science, from Sanskrit the Indian version, and from Pahlavi the Sāsānian combination of the two. But to these influences Islāmic astrology, through the work of Abū Ma'shar, an astrologer of the 9th century, added the Harranian adaptation of the Neoplatonic definition of the mode of astral influences in terms of Aristotelian physics. Abū Ma'shar further elaborated Sāsānian astrological history and greatly expanded the number of lots that an astrologer had to take into consideration. Much attention was paid by the Muslims to catarchic and interrogatory astrology, but, under attack by the theologians for denying divine intervention in the world and man's free will, astrology rapidly declined in its appeal to Muslim intellectuals after the Mongol invasions of the 13th century, though not before its influence had spread in India, the Latin West, and Byzantium.

Contributions of Abū Ma'shar

In Byzantium. During the last upsurge of paganism in the 5th and 6th centuries AD, Byzantium (the Eastern Roman Empire) boasted a host of astrologers: Hephaestion, Julian of Laodicea, "Proclus," Rhetorius, and John Lydus. Though their works are singularly unoriginal compilations, they remain the major sources for an understanding of earlier Hellenistic astrology. By the end of the 6th century, however, the general decline of the Byzantine Empire's intellectual life and the strong opposition of the church had combined to virtually obliterate astrology, though some practice of reading celestial omens survived in Byzantium as it did in western Europe. The science was revived only in the late 8th century and the 9th century under the impact of translations from Syriac and Arabic. The period from about 800 to 1200 was the most propitious for Byzantine astrology, though nothing was essentially added to astrological theories or techniques. This period was rivaled only by a last flowering of astrology in the late 14th century, when John Abramius and his students revised the older astrological treatises in Greek to provide the Renaissance with vulgate texts.

In western Europe. The astrological texts of the Roman Empire were written almost universally in Greek rather than in Latin; the only surviving exceptions are the poem *Astronomica* of Manilius (*c.* AD 15–20), the *Matheseos libri* ("Books on Astrology") of Firmicus Maternus (*c.* 335), and the anonymous *Liber Hermetis* ("Book of Hermes") from the 6th century. In the absence of astronomical tables in Latin, however, none of these was of any use, and astrology for all practical purposes disappeared with the knowledge of Greek in western Europe. It was revived only with the numerous translations of Arabic astrological and astronomical treatises executed in Spain and Sicily in the 12th and 13th centuries, supplemented by a few translations directly from the Greek. But the new astrology in the Latin-reading world remained essentially an offshoot of Islāmic astrology, gaining an adequate representation of its Hellenistic originals only in the 15th and 16th centuries. These two centuries also witnessed the fullest flowering of astrology in western Europe, frequently in conjunction with Neoplatonism and Hermetism. By the 17th century, however—with the displacement of the Earth from the centre of the universe in the new astronomy of Copernicus (1473–1543), Galileo (1564–1642), and Johannes Kepler (1571–1630) and with the rise of the new mechanistic physics of Descartes (1596–1650) and Newton (1643–1727)—astrology lost its intellectual viability and became increasingly recognized as scientifically untenable. Though Kepler attempted to devise a new method of computing astrological influences in the heliocentric (Sun-centred) universe, he did not succeed.

Effect of modern science on astrology

Astrology in modern times. In the West, Newtonian physics and Enlightenment rationalism largely eradicated the widespread belief in astrology. In countries such as

India, however, where only a small intellectual elite has been trained in Western physics and other sciences, it manages to retain here and there its respected position among the sciences. Its continued legitimacy is demonstrated by the fact that some Indian universities offer advanced degrees in astrology.

Recently in the West, however, astrology has gained a large popular following, and there have been attempts to reestablish a firm theoretical basis for it, notably by the French psychologist Michel Gauquelin in his *The Scientific Basis of Astrology* (1964), although with results that are at best inconclusive. The divisions of the year governed by the 12 zodiacal signs (which are derived from Hellenistic astrology) as depicted in newspapers, manuals, and almanacs are as follows:

Aries, the Ram, March 21–April 19
Taurus, the Bull, April 20–May 20
Gemini, the Twins, May 21–June 21
Cancer, the Crab, June 22–July 22
Leo, the Lion, July 23–August 22
Virgo, the Virgin, August 23–September 22
Libra, the Balance, September 23–October 23
Scorpio, the Scorpion, October 24–November 21
Sagittarius, the Archer, November 22–December 21
Capricorn, the Goat, December 22–January 19
Aquarius, the Water Carrier, January 20–February 18
Pisces, the Fish, February 19–March 20

Attempts have been made by practitioners of astrology to incorporate into the general astrological scheme the planets discovered since the Renaissance and to find some sort of statistical relation between planetary positions and human lives. None of these attempts appears to be at all convincing to the various skeptics and other critics of astrology, however, and no serious explanation seems to exist regarding the alleged spheres of influence of the planets, the alleged nature of their influences, or the manner in which they operate. Despite these criticisms and others like them, astrology continues to attact people from all walks of life—from the casual followers who read their horoscopes in the daily newspapers to those who have their star charts drafted by professional astrologers. In short, even though it is regarded as devoid of intellectual value by some, astrology in its modern and historic forms remains of great interest sociologically and popularly.

(D.E.P./R.A.Gi.)

Divination: other forms

While divination is most commonly practiced in the modern Western world in the form of horoscopic astrology, other forms of divination were and are of equal importance for other cultures, both ancient and contemporary.

NATURE AND SIGNIFICANCE

Concerns of divination

Divination is universally concerned with practical problems, private or public, and seeks information upon which practical decisions can be made; but the source of such information is not conceived as mundane, and the technique of getting it is necessarily fanciful. The mantic (divinatory) arts are many, and a broad understanding can only emerge from a survey of actual practices in various cultural settings. A short definition, however, may be offered as a preliminary guide: divination is the effort to gain information of a mundane sort by means conceived of as transcending the mundane.

Though the act of divination is attended by respect and the attitude of the participants in the divinatory act may be religious, the subject matter of divination (like that of magic) is ephemeral—*e.g.*, an illness, a worrisome portent, a lost object. Divination is a consultative institution, and the matter posed to a diviner may range from a query about a few lost coins to high questions of state. The casual or solemn nature of the matter is normally matched by that of the diviner in terms of attitude, technique, and style. Where the diviner is a private practitioner, the elaborateness of the procedure may be reflected in the fee. In contrast to the worldly motives of some diviners, the calling of diviner-priest was seen by the ancient Etruscans in Italy and the Maya in Mexico as sacred; his concern was for the very destiny of his people. Divination has many rationales, and it is difficult to describe the diviner as a distinctive social type. He may be a shaman (private curer employing psychic techniques), a priest, a peddler of sorcery medicines, or a holy man who speaks almost with the voice of prophecy. To appreciate the significance of the diviner's art in any culture or era, one must be familiar with prevailing beliefs about man and the world. In Christian times Europe has moved from a horror of necromancy (conceived not as consultation with a ghost but as a literal "raising of the dead") to an amused tolerance (among the educated) of spiritualism as a sort of parlour game. To assert that European beliefs about God and man have remained the same throughout the Christian Era would be to ignore the impact of modern science and secularization. On the other hand, to suppose that divination has been doomed by science and secularism would be to ignore the abiding popularity of horoscopic astrology and recurrent fashions for other mantic disciplines—and perhaps to misjudge the security of "modern" beliefs.



Barotse basket diviner. The diviner shakes various objects in the winnowing basket and by interpreting their final juxtaposition seeks to predict the outcome of an illness and to name the sorcerer responsible.
By courtesy of the Livingstone Museum, Republic of Zambia

THE STRUCTURE OF DIVINATION

Worldviews and philosophical undergirdings for divination

The extent to which a practice such as divination should be called a corollary of the beliefs entailed and the extent to which the opposite might be true (*i.e.*, the beliefs deriving from the practice as an after-the-fact explanation) is difficult to ascertain. Among the great cultures, the Chinese tradition has given the broadest scope to divination; yet there is no single Chinese religious cosmology, or theory on the ordering of the world, comparable to those of the Maya, Sanskritic (Hindu), or Judeo-Christian traditions, from which the variety of popular practice can be seen to derive. Sometimes, as with the flourishing business of astrology in Christian countries since the Renaissance, the metaphysical (transcendent) presuppositions of mantic practice may have been muted in order to minimize conflict with official religious and scientific doctrines. Generally, however, the philosophical underpinnings of divination need not be deep or well worked out, but, where they are, they will afford clues to fundamental beliefs about man and about visible or invisible nature. Some traditions of divination—such as astrology, geomancy (divination by means of figures or lines), or the Chinese divinatory disciplines—are so old and established that it is virtually impossible to discover their original contexts. Over the centuries such practices have survived many changes and have become perennial attempts to answer recurring questions about the human condition.

Established long ago in the hieratic (priestly) discipline of primitive theocracies, such a tradition still bears the marks of the specialists who worked out its systematic techniques. Since the practice is now observed only as

a folk or popular tradition, however, it would be rash to suppose that any legitimate philosophical tradition undergirding divination survives. Only in the case of the *I Ching*, the Chinese "Classic of Changes," have scholarly commentaries of any great intellectual substance accumulated over the millennia. Systematic studies of geomancy are recent, and the literature of astrology is as perishable as it is massive. Babylonian astrology, from which later forms are derived, arose in an agrarian Mesopotamian civilization concerned the vicissitudes of nature and the affairs of state. The mercantile, seafaring, and individualistic Greeks absorbed the mantic system of the collectivistic floodplain civilization of Mesopotamia, elaborated on it by adding the horoscopic discipline, and transmitted it, through Hellenistic, Egyptian, and Islāmic science, to Europe. In the course of this transformation, a two-way relationship between a society's view of the world and its system may be seen. Various priests and scholars have made their contributions to the system; yet there also is a clear correspondence between the general character of a culture and the uses it finds for divination. That is, the worldview implicit in the divination system itself may reflect the historical rather than the current context of use. It requires only practical understanding to consult a ouija board or use a forked stick to decide where to drill for water. Hence, men of very different beliefs may adopt the same practices, and a full correspondence between practice and belief can only be expected where both have developed in the same cultural context. Where much of the popularity of the mantic art derives from its "exotic" flavour, its symbolism may be little understood. By its very nature, however, divination tends to develop as a discipline, becoming the tradition of an organized body of specialists. This is because the means to which the diviner must resort are conceived as setting him apart. That is the case even among such peoples as the Azande (adjective and singular noun form Zande) of the Nile-Congo divide in Africa, where the resort to divination is frequent, and the most common techniques utilized are recognized to be within the competence of ordinary men. There, on a sensitive or contentious issue, an extraordinary credibility is desired, and the ultimate reliability of an oracle reflects the political standing of its owner—the king's oracle, for example, is viewed as the final authority, and the royal court is scrupulously organized to guard this vessel of power (divinatory and other) from contamination. Few societies are as enthusiastically given to divination as the Azande, who routinely employ it to explore their thoughts and who will not consider any important undertakings without oracular confirmation in advance. Among the Azande, the ordinary man could be considered a divinatory specialist. Elsewhere, men are content to reserve divination for special crises, and consultation must be with a recognized expert in order to distinguish an authentic answer from a spurious one.

Types of divination. As schools of dramatic art range from those relying on explicit technique to those teaching intuitive identification with a role, mantic skills range from the mechanical to the inspirational but most often combine both skills in a unique, dramatically coherent format. The comparative study of divinatory practices is at least as old as the 1st-century-BC Roman orator and politician Cicero's treatise *De divinatione* (*Concerning Divination*), and the convenient distinction there drawn between inductive and intuitive forms designates the range. An intermediate class, interpretive divination, allows a less rigid classification, since many divinatory disciplines do not rely strongly either upon inductive rigour or upon trance and possession.

Inductive, intuitive, and interpretive forms of divination

Inductive divination presupposes a determinative procedure, apparently free from mundane control, yielding unambiguous decisions or predictions. The reading of the "eight characters" of a Chinese boy and girl before proceeding to arrange a marriage—the year, month, day, and hour of birth of the two persons to be betrothed—illustrates this class of procedures. The "characters" are all predetermined by the accidents of birth date and hour, and it is supposed that all proper diviners would come to the same conclusions about them.

Interpretive divination requires the combination of correct procedure with the special gift of insight that sets a diviner apart from his fellows. The contemporary Maya diviner of Guatemala, seeking to diagnose an illness, will carefully pass a number of eggs over the patient's body in order to draw into them an essence of the affliction. The intact contents are then collected in water, and the diviner withdraws into a darkened corner to bend over the receptacle and read the signs of the eggs. His recitation then interprets the origin and nature of the disease.

Intuitive divination presupposes extraordinary gifts of insight or ability to communicate with beings in an extramundane sphere. The "Shaking Tent" rite of the Algonkian Indians of Canada illustrates the use of uncanny phenomena to lend credence to a mediumistic performance. The diviner, bound and cloaked, is no sooner placed in his barrel-like tent than the tent begins to shake with astonishing vigour and to fill the air with monstrous noises; and this continues with great effect until, all of a sudden, the communicating spirit makes his presence known from within the tent and undertakes to answer questions. It is difficult to explain away the phenomena of spirit possession as products of deliberate instruction.

Cosmological and psychological conditioning

The cosmological and psychological conditioning that affects divinatory practices within a cultural tradition will influence in a similar fashion all of its religious practices. The Greeks tended to the intuitive or "oracular" style, and the Etruscans, in contrast, elaborated upon the more systematic but less versatile inductive practice of Mesopotamia—developing an authoritative state religion in which the positions were monopolized by the ruling class. Greek divination was eccentric in that sanctuaries were located apart from the centres of political power; the Etruscan system, on the other hand, was concentric, focused at the summit itself. Rome eclectically incorporated both Greek and Etruscan elements, the ecstatic cult and the expert "reading" of livers—*i.e.,* haruspimancy. Rome, however, never allowed divination to become the central preoccupation of society as it had been for Etruria, nor did it become an autonomous force in society as it had been for the Greeks. In this, Rome represented a balance that is more congenial to modern Western thought. Throughout the ancient Mediterranean world, the notable exception being Egypt, divination was tied to expiation and sacrifice: fate was perceived as dire but not quite implacable, and the function of divination was to foresee calamity in order to forestall it. In trans-Saharan Africa, religion centres on expiation and sacrifice, and divination is a pivotal institution, but the Mediterranean notion of fate is not developed. Instead, the trouble of a person is attributed to witchcraft, sorcery, or ancestral vexation—all of which are believed to be arbitrary and morally undeserved. Divination is employed to discover the source of trouble in order to remove it, whether by sacrifice, countersorcery, or accusation and ordeal. The mind is turned to past events or hidden motives of the present time, however, and not to the future—that would be to borrow trouble.

The function of divination. The function of divination needs to be understood in its motivational context. It is not enough to say that information won from the diviner serves to allay uncertainty, locate blame, or overcome misfortune. Divination is motivated by the fact that information, whether spurious or true, will please a client. Unless one assumes that the information is usually accurate, one would expect clients to be displeased and subsequently skeptical. A careful assessment of the kinds of information that divinatory systems are required to yield is thus in order. The two main kinds are general information about the future and specific information about the past as it bears upon the future.

The first kind of information is yielded by horoscopic divination. It is usually so general that it cannot be properly tested. If such information were really specific, the prediction could interfere with its own fulfillment, acting as a warning or breeding overconfidence. The other kind of information demanded from diviners is specific enough to be tested and often is; but testing a particular diviner's competence is seldom seen as putting the institution to the test. Indeed, it is common in trans-Saharan societies

for a troubled client to consult a series of diviners until he finds one who is convincing. Again, many systems of divination have a double check built into them: the question is posed first in the positive and then in the negative, and the oracle must (obviously without manipulation) answer consistently. The chances are actually even that any oracle will fail to do so, yet the credibility of such oracles seems not to be lost. Technically, this means that false information can be obtained by a client without weakening his belief in the source. Early students of divinatory practice concluded that clients must be gullible, superstitious, illogical, or even "prelogical"; *i.e.*, culturally immature. Ethnographic studies do not confirm this, suggesting rather that what a client seeks from the diviner is information upon which he can confidently act. He is seeking, in so doing, public credibility for his own course of action. Consistent with this motive, he should set aside any finding that he thinks would lead him into doubtful action and continue his consultations until they suggest a course that he can take with confidence. The diviner's findings are judged pragmatically.

Occasions of divinatory consultation

Clients seek out a diviner when they are unsure how to behave—when there is illness, drought, death, or the fear of death; when there is suspicion of malevolence, theft, or breach of faith; when dreams or other symptoms are disturbing or the signs of the time seem bad. Divination serves the purpose of circumscription, of marking out and delimiting the area of concern: the nature of the crisis is defined, the source of anxiety is named. Concern becomes allegation, bafflement decision. The diviner may function as a stage manager, speeding up the action, rejecting false moves in advance, or indicating the secret fear or the hidden motive. Where divinatory practice is a recognized resource, a man who ignores it is considered arbitrary, and one who heeds it needs no further justification for his actions. In this sense, the ultimate function of divination is the legitimation of problematic decisions.

VARIETIES OF DIVINATION

Because dramatic effect is important, divination takes many forms and employs a wealth of devices. In a general way, it may be said that inductive divination employs nonhuman phenomena, either artificial or natural, as signs that can be unambiguously read. The prime condition is that the signs appear to be genuine, not manipulated. Interpretive divination commonly combines the use of nonhuman phenomena with human action, employing devices so complex, subtle, or fluid that the special gifts of the diviner seem required if the meaning is to be known. It is here that divination takes its most characteristically dramatic forms. Intuitive divination usually places little reliance upon artificial trappings, except for dramatic effect. The impressive performer may exhibit gifts like those that in a different context would have made him an effective actor, writer, or political leader. Where the diviner can produce voices other than his own, the impression is that the gods or spirits are speaking.

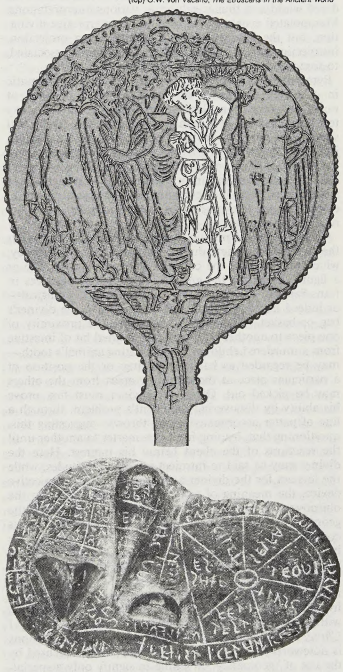
Inductive divination. To speculate that inductive divination from natural phenomena must be very old—*i.e.*, that it arose from early man's intimate acquaintance with nature—is tempting. In fact, however, evidence of an awareness of nature as a system among preliterate peoples is spotty, and this is particularly true in respect to astral observation. Divination from the skies is concerned preeminently with the future but presupposes a concern with cycles of time and history. Quite distinctive attitudes were taken toward the celestial clock by the ancient Maya astronomers and those of Mesopotamia; and distinct but related forms of astrology were developed in the Western, Indian, and Chinese civilizations. But the relation between astrology and scientific astronomy is quite apparent, and the two "sciences" were inseparable in the West until early modern times.

Forms of inductive divination

Associated with the observation of the heavens is the reading of signs in the weather and the movement of birds. The interpretation of lightning as a decipherable message from the gods—not simply as an outburst of divine anger—was brought to the level of a pseudoscience by the Etruscans. Winds and clouds, being suited to less exact observation, invited interpretive rather than inductive divination. Weather phenomena were also conceived of as in a special status relative to man, in that rain, drought, and natural disasters are forces that man seeks not simply to read but to control. Nonetheless, Hindu scripture discusses the art of interpreting "castles in the air"—celestial cities seen in towering clouds.

Augury, the art of interpreting omens, is the attempt to discover divine will in phenomena of animate nature. In Mesopotamia, augury was associated with sacrifice and perhaps developed from it. As the priests watched the rising smoke to divine the answer to a ritual query, they observed the movement of birds as auspicious or inauspicious. As a further augury the viscera of the sacrificial victim were examined, particularly the liver, which (rather than the heart) was conceived as the vital centre. The discipline of augury mapped cosmic space with the sacrificial altar at the centre, and each sector was assigned a definite meaning. Every event in the heavens could thus be charted and pondered. Similarly, haruspicy, the study

By courtesy of (bottom) the Museo Civico, Piacenza, Italy; from (top) O.W. von Vacano, *The Etruscans in the Ancient World*



(Top) Etruscan mirror showing a ritual scrutiny of a sacrificed liver. The various sections of the liver were believed to be a reflection of the divisions of heaven, each of which had a favourable or unfavourable meaning. The man (second from right) holds the liver so that the lobes can be examined and the details for foretelling the future are distinct. From Tuscania, Italy, 3rd century BC. In the Museo Archeologico, Florence. (Bottom) Bronze model of a sheep's liver used for divining. It is divided on its upper surface into about 40 sections, each with the name of an Etruscan divinity inscribed on it. From Piacenza, Italy, 4th–3rd century BC. In the Museo Civico, Piacenza, Italy.

of the liver, was developed by mapping it as a microcosm and reading it as one may read the palm.

Inductive divination from nature is associated with the reading of artificially contrived events, such as the movement of sacrificial smoke, the fall of an arrow shot upward, or the cast of dice or lots. A much-used natural–artificial technique consists in the braising of bone or shell to produce a system of signs. Scapulimancy—divination from a fire-cracked shoulder blade—was widespread in North America and Eurasia. The related but more elaborate Chinese technique of tortoiseshell divination was inspired by the idea of equating the carapace (back) and ventral (lower) shell with their view of a rounded sky over flat earth. Only the "earth" was inscribed and heated to produce signs. In general, however, artificial systems of signs are likely to be manipulatory, as they will be used in an artful way by the professional diviner—and in such cases interpretive techniques have to be taken into account.

Interpretive divination. Interpretive divination involves, in the main, the reading of portents, omens, or prodigies. To the scientifically minded, no event is without a cause. Yet, the apparently arbitrary event does occur in an ordered world and thus is subject to various interpretations. Manipulated events are an element of interpretive divination, but the less active forms depend upon projection, introjection, and free association—thus being associated, to some degree, with intuitive techniques.

Forms of interpretive divination

Pyromancy, divination by fire, may be highly dramatic in a society dependent on fire for light and safety at night. In some trans-Saharan societies the diviner may test an accusation at a seance around the fire, which will suddenly explode upon the "guilty" one. Elsewhere, objects may be overtly cast into the fire and signs read in the reaction. Hydromancy, divination by water, is usually less dramatic, ranging from the reading of reflections in a shallow surface, in the manner of the crystal gazer, to construing the movements of floating objects, as in the reading of tea leaves.

A range of related mantic practices may be grouped under the terms cleromancy, divination by lots, and geomancy, which may involve the casting of objects upon a map or a figure drawn on the ground. Cleromantic practices in trans-Saharan Africa may rely on the supposedly magical—or indeed horrifying—qualities of objects in the diviner's bag or basket. When they are thrown, the proximity of one piece to another—for example, a dried bit of intestine from a murdered child and a man-eating animal's tooth—may be regarded as having meaning; or the position of a particular piece at the centre or apart from the others may be picked out. Often, the diviner must first prove his ability by discovering the client's problem, through a line of patter accompanying the throws—suggesting this, questioning that, leaping from one matter to another until the reactions of the client betray his interest. Here the diviner may be said to introject ideas and attitudes, while the lots act for the diviner and client alike as a projective device, the meaning of which is only half-formed in the objective pattern cast. A far more elaborate practice is the geomancy of West Africa, in which elegant equipment is combined with impressive erudition to produce a seance in which lots are used to select verses, wherein the client is expected to find his answers. The nature of the lots employed, the number lore on which the selection of verses is based, and the verses themselves are entirely distinct from their counterparts in the Chinese yarrow (an herb with finely dissected leaves) tradition embodied in the *I Ching*, but the general equivalence of the two elaborations is noteworthy. The parallel has perhaps been obscured by the use of "geomancy" in China to signify only a specialized art by which propitious locations are selected.

Sometimes a diviner can be said to interpret signs so characteristic of his client that the practice falls between interpretive and intuitive arts. Somatomancy, body divination, is clearly interpretive in most forms, whether in China or the West, though the system of signs employed comprises private attributes of the client's physique. Examples are phrenology, employing features of the head that are normally unnoticed; and the reading of moles, where the body is treated as a microcosm bearing astrological signs. But oneiromancy, dream interpretation, employs explicitly psychic phenomena; and here the diviner may be said to assist the intuition of meaning by his client as often as he can be said to introject. The Ojibwa and Bella Coola Indians of North America were characteristically preoccupied with the meanings of their dreams.

Intuitive divination. The prototype of the intuitive diviner is the occasional shaman or curer who uses trance states. These are achieved idiopathically (*i.e.*, arising from the self spontaneously) or induced by drugs or by autokinetic (self-energized) techniques, such as hand trembling among the Navajo, a large North American Indian tribe. As a mantic art, trance is associated with oracular utterance and spirit possession. An impressive performance will be taken to represent the actual voice of a god or spirit addressing the client directly; and divination in this mode is known from diverse religious traditions, including Christianity. The idea that the gods may be importuned to speak on a matter of temporal human concern seems to be very ancient. In early Egypt incubation was practiced—*i.e.*, sleeping in the temple in the hope of being inspired by the resident god. The idea behind Maya maiden sacrifice was the same: a number of maidens were cast into a sacred cenote or deep well, and those who survived after some hours were brought back to recite the messages received during their ordeal—a virtual enactment of the journey into the underworld. As oracular utterance became regular, special techniques or contraptions were developed for making the god's image show assent or denial or for amplifying the sound of an unseen priest's voice. In nomadic societies today, however, the diviner may still achieve personal authority by passing into a trance before his fellows, trembling and speaking "as if possessed"—that is, as if his own spirit had ceased to inhabit his body and had been replaced by another.

Forms of intuitive divination

Related to possession is the conviction that malevolent persons are essentially unlike innocent ones, though not in outward appearance. When a test is devised for discovering malevolence, commonly conceived of as witchcraft or as a nonhuman force disguising itself in human form, the test takes the form of an ordeal. This may be a demonstration of invulnerability to harm, the presence of blessed qualities being viewed as inconsistent with malevolence; among the many types of ordeal are walking on coals and retrieving an object from boiling liquid. The ordeal may even involve death: in the ordeal by water, a witch was expected to float and so be spared for burning, but an innocent person would be accepted by the water and drown. In trans-Saharan poison ordeals the innocent person is expected to survive.

Intuitive divination may also be a wholly private affair. A Roman might hear a warning from the gods in a piece of conversation; the Aztec might discern a portent in an animal's howl. The North American Indian who sought a private vision through isolation, self-mutilation, and fasting would preserve the memory of that vision throughout life, turning to it as his unique guardian spirit.

Divination today. The immense popularity of horoscopes in the urban West today illustrates the almost exclusive concern with individual fortune-telling that characterizes divination in a mobile and competitive mass society. Chiromancy, Tarot (fortune-telling) cards, and crystal gazing represent respectively body divination, sortilege (divination by lots), and trancelike performance in styles suitable for what might be called a half-serious attempt to learn one's fate. Necromancy, in its modern spiritualist form, represents a slightly more serious and sustained effort to establish contact with extramundane beings. But astrology, in its various popular forms, is the form of divination best suited to mass consumption, since it is based on a well-articulated body of lore, touches matters of high destiny as well as individual fortune, and "personalizes" its introjective advice without the client's having to be interviewed. On the other hand, the more esoteric mantic arts have the appeal of discipline—an individual may enter into the lore deeply until it becomes a part of his own worldview. Study of the *I Ching* for divinatory purposes can involve this sort of commitment.

Popularity of horoscopic astrology

(G.K.P./R.A.Gi.)

Magic

The term *magic* is best defined as a concept used to describe a way of thinking that looks to invisible forces to influence events, effect change in material conditions, or present the illusion of change. Within the Western tradition, this way of thinking is distinct from religious or scientific modes; however, such distinctions and even the definition of magic are subject to wide debate.

NATURE AND SCOPE

Practices classified as magic include divination, astrology, incantations, alchemy, sorcery, spirit mediation, and necromancy. The term *magic* is also used colloquially in Western popular culture to refer to acts of conjuring and sleight of hand for entertainment. The purpose of magic is to acquire knowledge, power, love, or wealth; to heal or ward off illness or danger; to guarantee productivity or success in an endeavour; to cause harm to an enemy; to reveal information; to induce spiritual transformation; to trick; or to entertain. The effectiveness of magic is often determined by the condition and performance of the magician.

Magic is sometimes divided into the "high" magic of the intellectual elite, bordering on science, and the "low" magic of common folk practices. A distinction is also made between "black" magic, used for nefarious purposes, and "white" magic, ostensibly used for beneficial purposes. Although these boundaries are often unclear, magical practices have a sense of "otherness" because of the supernatural power that is believed to be channeled through the practitioner, who is a marginalized or stigmatized figure in some societies and a central one in others.

ELEMENTS OF MAGIC

Spells and material. The performance of magic involves words (*e.g.* spells, incantations, or charms) and symbolic numbers that are thought to have innate power, natural or man-made material objects, and ritual actions performed by the magician or other participants. A spell or incantation is believed to draw power from spiritual agencies to accomplish magic. Knowledge of spells or symbolic numbers is often secret (occult), and the possessor of such knowledge can be either greatly revered or feared. In some cases, the spell is the most highly regarded component of the magical rite or ceremony. The Trobriand Islanders of Melanesia, for example, regarded using the right words in the right way as essential to the efficacy of the rite being performed. Among the Maori of New Zealand the power of words is thought to be so important that mistakes in public recitations are believed to cause disasters for individuals or the community. Moreover, spells often employ an esoteric vocabulary that adds to the respect accorded rites. Belief in the transformative power of words is also common in many religions. Shamans, spirit mediums, and mystics, for example, repeat specific sounds or syllables to achieve an ecstatic state of contact with spiritual forces or an enlightened state of consciousness. Even modern magic for entertainment retains a residual of the spell with its use of the term *abracadabra*.

Medicines

Much anthropological literature refers to the objects used in magic as "medicines," hence the popular use of the term *medicine man* for magician. These medicines include herbs, animal parts, gemstones, sacred objects, or props used in performance and are thought to be potent in themselves or empowered by incantations or rituals. In some cases, medicines that are intended to heal are physiologically effective; for example, willow bark is used by some Chinese as an analgesic, and garlic and onions were used as antibiotics in medieval Europe. Other medicines that are meant to cause harm, such as toad extracts and bufadienolides, are, in fact, known poisons. Other materials have a symbolic relationship to the intended outcome, as with divination from animal parts. In sorcery a magician may employ something belonging to the intended victim (*e.g.*, hair, nail parings, or a piece of clothing) as part of the ritual. The rite itself may be symbolic, as with the drawing of protective circles in which to call up spirits or the destruction of a wax image to harm a victim. Plants or other objects can also symbolize desired outcomes: in rites to ensure a canoe's speed, the Trobriand use light vegetable leaves to represent the ease with which the craft will glide over the water.

Performance and functions. Because magic is based on performance, ritual and the magician's knowledge and ability play a significant role in its efficacy. The performance of magic also presumes an audience, either the spiritual forces addressed, the patient-client, or the community. Both the magician and the rite itself are concerned with the observance of taboos and the purification of the participants. Magicians, like priests presiding over religious rituals, observe restrictions of diet or sexual activity to demarcate the rite from ordinary and profane activities and to invest it with sanctity.

Instrumental and expressive functions

Foremost among the many roles magic plays are its "instrumental" and "expressive" functions. Based in the attempt to influence nature or human behaviour, magic's instrumental function is measured by its efficacy in achieving the desired result. Anthropologists identify three main types of instrumental magic: the productive, the protective, and the destructive. Productive magic is employed to solicit a successful outcome from human labour or nature, such as a bountiful hunt or harvest or good weather. Protective magic aims to defend an individual or community from the vagaries of nature and the evil of others. The use of amulets to ward off contagious diseases or the recitation of charms before a journey are examples. Lastly, destructive magic, or sorcery, is intended to harm others, often is motivated by envy, and is socially disruptive. Consequently, the use of countermagic against sorcery may relieve some social tension within a community.

Magic's expressive function results from the symbolic and social meanings attached to its practices, though its performers may not necessarily be aware of this function. Magic can provide a sense of group identity through shared rituals that give power or strength to members. At the same time, it can isolate the magician as a special person within or on the margins of society. Magic can also serve as a creative outlet or form of entertainment. It is, therefore, inseparable from the total system of thought, belief, and practice in a given society.

DEFINITIONAL ISSUES: MAGIC, RELIGION, AND SCIENCE

Dichotomies in context. The term *magic* cannot be defined in isolation because of its broad parameters, important role in many societies, and interactions with related phenomena. *Magic* is a generic label used by outsiders (theoretically, objective observers) to describe specific practices in societies in which this word or its conceptual equivalent may not even exist. As a result, diverse phenomena are lumped together on the assumption that they operate in the same way. This artificial construct of magic also exists only in relation to what it is not—primarily, religion and science as alternate modes of rationality. Such definitions of magic privilege cultures with a strong scientific orientation and stigmatize those that practice magic instead of religion. Consequently, defining magic and identifying magicians requires an understanding of the cultural contexts in which these labels are used.

Magic and Western thought

Although magic has an ambiguous relationship with Western religion and science, it is rooted in the main institutional, social, and intellectual traditions in Western history. Moreover, modern attempts to arrive at a universal definition of magic reflect a Western bias. In a model developed by the British anthropologist Sir James Frazer (1854–1941), magic is characterized as an early stage in human development, superseded first by religion and then by science. The debate over the relationship between magic, religion, and science that dominated much of the discussion about magic throughout the 20th century is evident in the fieldwork of anthropologists, the theories of sociologists of religion, and critiques by postmodernists. Consequently, research in comparative religions, history, and anthropology in the second half of the 20th century moved away from the evolutionary model toward more context-sensitive interpretations, while other studies developed new models for cross-cultural comparison. Nonetheless, the magic-religion-science model retains considerable interpretive power, and the dichotomies used to distin-

guish magic from religion or science are pervasive in popular discourse.

Magic and religion. Magic continues to be widely perceived as an archaic worldview, a form of superstition lacking the intrinsic spiritual value of religion or the rational logic of science. Religion, according to seminal anthropologist Sir Edward Burnett Tylor (1832–1917), involves a direct, personal relationship between humans and spiritual forces; in religion's highest form, that relationship is with a personal, conscious omnipotent spiritual being. Magic, on the other hand, is characterized as external, impersonal, and mechanical, involving technical acts of power. Magic seeks to manipulate spiritual powers, while religious prayer supplicates spiritual forces, a distinction explored by Bronislaw Malinowski (1884–1942) in his work on the Trobriand Islanders. Moreover, according to Émile Durkheim (1858–1917), religion is communal because its adherents, bound together by shared belief, form a church. Magic, on the other hand, involves no permanent ties between believers and only temporary ties between individuals and the magicians who perform services for them; however, magic, too, may have a communal dimension.

Magic and science. Magic, like religion, is concerned with invisible, nonempirical forces; yet, like science, it also makes claims to efficacy (the ability to produce a desired outcome). Unlike science, which measures outcomes through empirical and experimental means, magic invokes a symbolic cause-effect relationship. Moreover, like religion and unlike science, magic has an expressive function in addition to its instrumental function. Magical rainmaking strategies, for example, may or may not be efficacious, but they serve the expressive purpose of reinforcing the social importance of rain and farming to a community.

Subcategories of magic. The view of magic as pre-religious or nonscientific has contributed both to subtle distinctions between magic and other practices and to the recognition of subcategories of magic. Notably, anthropologists distinguish magic from witchcraft, defining the former as the manipulation of an external power by mechanical or behavioral means to affect others and the latter as an inherent personal quality that allows the witch to achieve the same ends. However, the line between the two is not always clear, and in some parts of the world an individual may operate in both ways. Similarly, the distinction between "black" magic and "white" magic is obscure since both practices often use the same means and are performed by the same person. Scholars also distinguish between magic and divination, whose purpose is not to influence events but to predict or understand them. Nevertheless, the mystical power of diviners may be thought to be the same as that behind magic. Ultimately, despite these distinctions and the variety of unique roles that practitioners play in their own societies, most end up classified under the universal term *magician.* Often even religious figures such as priests, shamans, and prophets are identified as magicians because many of their activities include acts defined as "magical" by modern scholars.

"Black" and "white" magic

In the end, distinctions between magic and religion or science are harder to make in practice than in theory; scholars therefore use labels such as *magico-religious* to describe activities or persons who cross this artificial dividing line. Similarly, the boundary between magic and science is permeable, since the modern scientific method (observation and experimentation) evolved from forms of scientific magic such as alchemy and astrology. Thus, the evolutionary model, which draws sharp distinctions between magic, religion, and science, cannot account for the essential similarity between various phenomena.

CONCEPTUAL HISTORY

The claim that magic is found in all human societies rests on a definition that is rooted in Western cultural assumptions, and both these assumptions and the use of the term *magic* have undergone change over time and place. Consequently, to understand beliefs and practices in other societies that appear similar to European magic, it is necessary to apply the context-sensitive and comparative methods that become increasingly important in the study of anthropology, history, and religion.

Ancient Mediterranean world. The root word for magic (Greek: *mageia;* Latin: *magia*) derives from the Greek term *magoi,* which refers to a Median tribe in Persia and their religion, Zoroastrianism. The Greco-Roman tradition held that magicians possessed arcane or secret knowledge and the ability to channel power from or through any of the polytheistic deities, spirits, or ancestors of the ancient pantheons. Indeed, many of the traditions associated with magic in the classical world derive from a fascination with ancient Middle Eastern beliefs and are concerned with a need for countermagic against sorcery. Ancient Roman society was particularly concerned with sorcery and countersorcery, contests associated with the development of competitive new urban classes whose members had to rely on their own efforts in both material and magical terms to defeat their rivals and attain success.

Ambivalence toward magic carried into the early Christian era of the Roman Empire and its subsequent heirs in Europe and Byzantium. In the Gospel According to Matthew, the Magi who appeared at the birth of Jesus Christ were both Persian foreigners of Greco-Roman conception and wise astrologers. As practitioners of a foreign religion they seemed to validate the significance of Jesus' birth. However, *magus,* the singular form of *magi,* has a negative connotation in the New Testament in the account of Simon Magus (Acts 8:9–25), the magician who attempted to buy the miraculous power of the disciples of Christ. In medieval European Christian legends, his story developed into a dramatic contest between true religion, with its divine miracles, and false demonic magic, with its illusions. Nonetheless, belief in the reality of occult powers and the need for Christian counterrituals persisted.

Medieval and early modern Europe. During the period of Europe's conversion to Christianity (*c.* 300–1050), magic was strongly identified with paganism, the label Christian missionaries used to demonize the religious beliefs of Celtic, Germanic, and Scandinavian peoples. Church leaders appropriated and Christianized native practices and beliefs. For example, medicinal remedies found in monastic manuscripts combined Christian formulas and rites with Germanic folk rituals to empower natural ingredients to cure ailments caused by poisons, elf-attack, demonic possession, or other invisible forces. Although coopted and condemned by Christian leaders of this period, magic survived in a complex relationship with the dominant religion. Similar acculturation processes occurred later in Latin America and Africa, where indigenous beliefs in spiritual forces and magical practices coexist, sometimes uneasily, with Christian theology.

In high medieval Europe (*c.* 1050–1350), the battle between religion and magic occurred as the struggle against heresy, the church's label for perverted Christian belief. Magicians, like heretics, were believed to distort or abuse Christian rites to do the Devil's work. By the 15th century, belief in the reality of human pacts with the Devil and the magical powers acquired through them contributed to the persecution of those accused of harming others with their magic. Also in the high Middle Ages the demonization of Muslims and Jews contributed to the suspicion of the "other." Marginal groups were routinely accused of ritual baby-killing. In lurid accounts of the "blood libel," Jews were charged with stealing Christian children for sacrifice. Similar accusations were made against witches by Christians and against Christians by the ancient Romans.

Magic and heresy

Although magic was widely condemned during the Middle Ages, often for political or social reasons, the proliferation of magic formulas and books from the period indicates its widespread practice in various forms. Richard Kieckhefer has identified two major categories of magic: "low" magic includes charms (prayers, blessings, adjurations), protective amulets and talismans, sorcery (the misuse of medical and protective magic), divination and popular astrology, trickery, and medical magic through herbs and animals; and "high," or intellectual, magic, includes more learned forms of astrology, astral magic, alchemy, books of secrets, and necromancy. Moreover, magic served as a literary device of the time, notably the presence of Merlin in the Arthurian romances. Although medieval European magic retained its sense of otherness by borrow-

ing from Jewish practices and Arabic scientific sources such as the astral magic manual *Picatrix,* it also drew from the mainstream Christian tradition. Necromancy, for example, used Latin Christian rites and formulas to compel the spirits of the dead to obey.

By the late Middle Ages (*c.* 1350–1450) and into the early modern period (*c.* 1450–1750), magic was regarded as part of a widespread and dangerously antisocial demonic cult that included the condemned practices of sorcery, necromancy, and witchcraft. Accused heretics, witches, and magicians were subject to inquisitions designed to uncover these connections and to destroy the means of transmission (*e.g.,* the burning of condemned books and/or the "guilty" parties). The influential manual *Malleus maleficarum* ("The Hammer of Witches," 1486) by Jacob Sprenger and Henry Krämer describes witchcraft in great detail; moreover, this oft-reprinted volume is responsible for the association of witchcraft with women that becomes dominant in the early modern period. This conspiracy theory of demonic magic contributed to the early modern "witch craze" that occurred at a time of growing tension between magic, religion, and nascent science.

Witchcraft and magic

Nonetheless, despite the persecution of "black" magic and its alleged practitioners, forms of "white" magic persisted in Europe on the boundaries between magic, mysticism, and emerging empiricism. During the Renaissance there was renewed interest in ancient Middle Eastern practices, neoplatonic mysticism, and Arabic texts on alchemy and astrology. Pico della Mirandola sought hidden knowledge in Jewish Kabbala, a mystical practice for unlocking the divine secrets contained in written and unwritten Hebrew Scriptures, while Giordano Bruno explored the mystical traditions of Hermeticism, based on works of the legendary Alexandrian prophet of the 1st–3rd century Hermes Trismegistus. Although generally tolerated because their practices were perceived to be within the main Judaic and Christian Hermetic tradition, practitioners of alchemy were sometimes considered to be evil magicians who acquired their knowledge through a pact with the Devil (as in the Faust legends).

European traditions and the modern world. The European fascination with the magical traditions of the ancient Middle East was extended to those of East and South Asia in the early modern period. Orientalism, as literary and cultural critic Edward Said labeled this phenomenon, has its roots in the sense of the "other" found in the earliest definitions of magic (notably the Magi as Persian foreigners) and in the Renaissance penchant for Egyptian, Hebrew, and Arabic materials.

A popular "scientific" worldview prevails in modern Western societies that suggests the triumph of human reason. Enlightenment rationalism and the Scientific Revolution—ironically rooted in Renaissance experiments in magic and motivated in part by Reformation pragmatism—led to the modern triumph of scientific reasoning over magic, evident, for example, in 19th-century exposés of magic tricksters as charlatans. Notably, spirit rappers, mediums who "conversed" with spirits who replied by knocking on a table, were easily exposed as the ones doing the knocking. Modern popular magic has appeared in the realm of entertainment, generally as a plot device in stories and movies, as tricks aimed at children, and as mysterious sleight-of-hand illusions in magic shows that delight the audience's sense perceptions and challenge their reasoning ability. The fascination with occult knowledge and mystical powers derived from nonmainstream or foreign sources persists in the West in astrological charts in newspapers, theories of interplanetary aliens and government conspiracies to hide them, occult rituals in some New Age religions, and interest in traditional practices that have an esoteric flavour, such as *feng shui* (geomancy, the traditional Asian practice of aligning graves, homes, and temples with cosmic forces). This persistence suggests, in part, the impact of globalization on postmodern worldviews challenging the dominance of a strictly scientific mode of rationality.

Globalization of the magic concept and the postcolonial world. Western conceptions of magic, religion, and science were exported to other parts of the globe in the modern period by traders, conquerors, missionaries, anthropologists, and historians. European travelers in the 16th–19th centuries functioned as primitive ethnographers whose written observations are invaluable historical resources. However, their accounts, often coloured by their Judeo-Christian assumptions about religion versus magic, illuminate how indigenous peoples were treated as "children" to be educated or, in the case of some conquerors, as subhuman races to be enslaved. During the latter part of the 19th century, anthropologists began to analyze magic and its part in the evolution of the world's religions. Their work was characterized by a fundamental distinction rooted in the magic-religion-science evolutionary model: the world is divided between historical, literate urbanized cultures, or "civilizations" (for example, the ancient traditions of East and South Asia) and nonliterate, tribal archaic, or "primitive," societies (such as those found in parts of Africa, the Americas, and Oceania). Historians viewed complex societies characterized by urbanization, centralization, and written traditions as more advanced and measured their progress as civilizations according to the evolutionary model. Nomadic, tribal, agricultural, or nonurbanized societies with strong oral traditions were often perceived by early European observers as developmentally stagnant people without history. While these views are no longer accepted, their residual effect is still felt in the way magic, religion, and science are conceptualized. Anthropologists of religion traditionally distinguished between the "religion" practiced by the world's main faiths, which often marginalize magic as superstition, and the beliefs of small nonliterate societies in which "magic" may in fact be central to religious belief. Here the distinction between religion and magic seems unfounded. Indeed, as some postcolonial societies try to distance themselves from Western logic, ancient religious traditions are pivotal to the reassertion of cultural identity and autonomy. West African vodun (voodoo), which spread to the Caribbean, the Americas, and elsewhere, is one example of an indigenous religious practice that is tied to cultural identity and used subversively as a rallying point for postcolonial resistance to Western values.

Western magic and world cultures

The Western concept of magic as a set of beliefs, values, and practices that are not fully religious or scientific does not find its equivalent in non-Western languages and cultures; conversely, concepts found in other cultures may be untranslatable into a Western framework. For example, Hawaiian historian David Malo (*c.* 1793–1853), discussing Christianity and traditional Hawaiian religion, found *ho'omana* (to make, to do, or to imbue with supernatural, divine, or miraculous power) the closest translation for English *religion,* contrary to its characterization by Westerners as a magical component in Polynesian beliefs. A modern Japanese dictionary uses a transliteration, *majikku,* for the English word *magic.* It also uses the English word *magic* to translate several Japanese words beginning with *ma-,* the kanji character representing a vengeful spirit of the dead (in East Asian folk belief, an ancestor not cared for properly; in Buddhist cosmology, an evil demonic figure). While superficially similar to the Christian notion of magic as demonic, the cosmologies regarding these demons differ significantly. Moreover, *ma-* does not have the range of meanings that *magic* has in Western thought.

On the other hand, specific practices identified as magic—*e.g.,* divination, spells, spirit mediation—are found worldwide, even if the word *magic* is not. For example, in China various practices such as divination through oracle bones, offerings to dead ancestors, and *feng shui* can be classified as magic, religion, or science, but it is questionable whether these categories have any validity in Chinese thought; rather these practices are an intrinsic part of the worldviews expressed in China's main religious and philosophical systems (ancestor worship, Confucianism, Taoism, and Buddhism). In modern China, some communities deal with crisis by combining seemingly contradictory practices—including supplication and coercion of gods, appeals to ancestral spirits, folk cures, and modern inoculations. Such syncretism has been common in East Asia; notably, in 6th-century Japan, the native nature worship of Shinto blended with imported forms of Buddhism without the kind of conflict that occurred during the conversion of Europe to Christianity. In modern East Asia,

conflict between magic, religion, and science introduced by Western concepts of magic occurs alongside a strong tradition of syncretism that blends empirical science with practices that Westerners often perceive as unscientific magic or religious superstition.

Asian religious traditions such as Hinduism, Buddhism, and Taoism teach that material life is illusory and focus on understanding the principles and spiritual forces that lie behind physical experience. Adepts in these traditions who have achieved a level of understanding of these cosmic forces often appear to have the ability to manipulate physical reality in ways that seem magical. The point of demonstrations by street magicians and snake charmers in India is to show the illusory quality of material reality in order to draw attention to the universal, timeless, and cosmic. Purposeful deception in magic is thus used to illustrate the deceptiveness of human apprehensions of reality. The mystical component of magic is also clear in Tantra and other nonconformist sects of Hinduism or Buddhism, which use mystical words, symbols, and diagrams in their rituals. Whether these practices are magic or religion depends upon one's point of view.

Anthropological and sociological studies of modern nonliterate societies in the Americas, Oceania, and Africa have given rise to new global terminology. Beginning in the second half of the 20th century, some sociologists and anthropologists turned the tables on earlier scholarship by applying the methods used for examining extant nonliterate ("primitive") societies to literate, urban societies of the past, which previously had been evaluated by the criteria reserved for the study of "civilizations." For example, the phenomenon of shamanism and the word *shaman*, as defined by Mircea Eliade (1907–86) in his exploration of ecstatic states, has been applied not only to "primitive" cultures but to premodern Christian Europe.

HISTORY OF MAGIC THEORIES

Early theories. *Foundations.* Because of the impact of anthropological theory on the study of magic, its development and history bear reviewing. The first important figure in this line was Sir Edward Burnett Tylor, whose *Primitive Culture* (1871) saw magic as a "pseudo-science" in which the "savage" postulated a direct cause-effect relationship between the magical act and the desired outcome. Tylor regarded magic as "one of the most pernicious delusions that ever vexed mankind," but he did not approach it as superstition or heresy. Instead, he studied it as a phenomenon based on the "symbolic principle of magic," a scheme of thought founded on a rational process of analogy. He also realized that magic and religion are parts of a total system of thought. Although he believed that magic and animistic beliefs became less prevalent in the later stages of history, he did not view magic and religion as alternative stages in the evolutionary development of mankind.

The Golden Bough

That conclusion was left for Sir James Frazer in *The Golden Bough* (1890), in which he ordered magic, religion, and science in a grandiose evolutionary scheme. Magic preceded religion because the former was logically simpler. This notion, however, was based on his erroneous assumption that the Australian Aborigines, examples of a "primitive" people, believed in magic but not in religion.

Sociological theories. Another line of theorists, including sociologists Durkheim and Mauss, widened the discussion by defining magic in terms of its social function. In *The Elementary Forms of the Religious Life* (1912), Durkheim argued that magical rites involved the manipulation of sacred objects by the magician on behalf of individual clients; the socially cohesive significance of religious rites proper (by priests) was therefore largely lacking. Durkheim's views were furthered by A.R. Radcliffe-Brown in *The Andaman Islanders* (1922) and to a lesser extent by Malinowski in *Argonauts of the Western Pacific* (1922) and *Magic, Science and Religion* (1925).

Subsequent studies of the working of systems of magic, especially in Africa and Oceania, built upon the work of Malinowski and Radcliffe-Brown along with that of Sir Edward Evans-Pritchard in *Witchcraft, Oracles and Magic Among the Azande* (1937). In his seminal book, Evans-Pritchard demonstrated that magic is an integral part of religion and culture used to explain events that cannot otherwise be understood or controlled. The Zande of The Sudan accept magic, together with witchcraft and oracles, as a normal part of nature and society. These phenomena form a closed logical system, each part of which buttresses the other and provides a rational system of causation.

Psychological theories. These anthropological and sociological approaches focused on magic as a social phenomenon, but the role of individual psychology was implicit in the views of Tylor and Frazer and brought out more in the work of Malinowski, who frequently offered psychological explanations for belief in magic. Sigmund Freud's influential view of magic as the earliest phase in the development of religious thought (*Totem and Taboo,* 1918) followed Frazer's model and posited an essential similarity between the thought of children, neurotics, and "savages." According to Freud, all three assumed that wish or intention led automatically to the fulfillment of the desired end. This view, based on outmoded notions about "primitive" cultures, was revised as the result of new field research. Although Claude Lévi-Strauss also initially equated these three groups, he later modified this view in his analysis of the work of Mauss, which focuses on the structural linguistics of terms such as *mana* that are deployed in the study of magic. His work laid the foundation for later deconstructions of the concept of magic.

Freud's view

Recent theories. *Comparative religions.* The rise of the study of comparative religions led to new theories that accounted for both world religions and localized belief systems. The work of Eliade, including his study of shamanism, is an important example of this approach, as is that of Ninian Smart, who devised a six-dimensional (experiential, mythic, doctrinal, ethical, ritual, and social) worldview analysis for cross-cultural comparison that can be applied to different belief systems, whether called magic or religion. Likewise, Judaic scholar Jacob Neusner suggested the neutral rubric "modes of rationality" to avoid pejorative comparisons between systems of thought otherwise classified as magic, religion, science, or philosophy. The broader base established by the comparative religions approach avoids the difficulties of distinguishing urban literate from nonurban nonliterate societies and the perils of the magic-religion-science progression.

Postmodern dialogue. Postmodern scholarship continues to challenge older anthropological notions. The work of such anthropologists as Victor Turner (1920–83), Clifford Geertz, and Marshall Sahlins has had a wide impact on the social sciences and humanities. Central to the challenge to the traditional magic-religion-science paradigm was *Magic, Science, Religion, and the Scope of Rationality* (1990), in which Stanley Jeyaraja Tambiah deconstructs the European history of the progress model and the work of anthropologists from Tylor forward. Other anthropologists have questioned the model of the rise and decline of magic in European thought in Keith Thomas's *Religion and the Decline of Magic* (1971), a study of early modern England, and Valerie Flint's *The Rise of Magic in Early Medieval Europe* (1991). These cross-disciplinary debates, along with the rejection of the Western magic-religion-science paradigm, have contributed to more sensitive treatments of magical practices in diverse societies.

CONCLUSION

The study of magic as a distinct cultural phenomenon has a long history in anthropological, sociological, and historical studies. Although some distinctions between magic and other religious or scientific activities may be useful, magic cannot be studied in isolation as it once was. Practices classified as magic represent essentially an aspect or reflection of the worldview held by a particular people at a particular point in their own historical development. Magic, like religion and science, is thus a part of a culture's total worldview.

(J.F.M./R.A.Gi./K.L.J.)

Witchcraft

The term witchcraft refers to the exercise or invocation of alleged supernatural powers to control people or events, practices typically involving sorcery or magic. Although

defined differently in disparate historical and cultural contexts, witchcraft has often been seen, especially in the West, as the work of crones who meet secretly at night, indulge in cannibalism and orgiastic rites with the Devil, and perform black magic. Witchcraft thus defined exists more in the imagination of contemporaries than in any objective reality. Yet this stereotype has a long history and has constituted for many cultures a viable explanation of evil in the world. The intensity of these beliefs is best represented by the European witch-hunts of the 14th to 18th century, but witchcraft and its associated ideas are never far from the surface of popular consciousness and—sustained by folk tales—find explicit focus from time to time in popular television and films and in fiction.



"Linda Maestra," by Francisco de Goya, showing a young witch being instructed by an experienced hag.
By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd

DEFINITIONS: WITCHCRAFT AND SORCERY

The modern English word *witchcraft* has three principal connotations: the practice of magic or sorcery worldwide; the beliefs associated with the Western witch-hunts of the 14th–18th centuries; and varieties of the modern movement called Wicca, frequently mispronounced "wikka."

The terms *witchcraft* and *witch* derive from Old English *wiccecraeft:* from *wicca* (masculine) or *wicce* (feminine), pronounced "witchah" and "witchuh," respectively, denoting someone who practices sorcery. Roughly equivalent words in other European languages have different connotations, and none precisely translates another. The difficulty is even greater with the relevant words in African, Asian, and other languages. The problem of defining witchcraft is made more difficult because the concepts underlying these words also change according to time and place, sometimes radically. Moreover, different cultures do not share a coherent pattern of witchcraft beliefs, which often blend other concepts such as magic, sorcery, religion, folklore, theology, technology, and diabolism. Some societies regard a witch as a person with inherent supernatural powers, but in the West witchcraft has been more commonly believed to be an ordinary person's free choice to learn and practice magic with the help of the supernatural. (The terms *West* and *Western* in this article refer to European societies themselves and to post-Columbian societies influenced by European concepts.) The answer to the old question "Are there such things as witches?" therefore depends upon individual belief and upon definition, and no single definition exists. One thing is certain: the emphasis on the witch in art, literature, theatre, and film has little relation to external reality.

Misconceptions

False ideas about witchcraft and the witch-hunts persist today. First, the witch-hunts did *not* occur in the Middle Ages but in what historians call the "early modern" period (the late 14th to the early 18th century), the era of the Renaissance, the Reformation, and the Scientific Revolution. There was neither a witch-cult nor any cult, either organized or disorganized, of a "Horned God" or of any "Goddess"; Western "witches" were *not* members of an ancient pagan religion; and they were *not* healers or midwives. Moreover, not all persons accused of witchcraft were women, let alone old women; indeed, there were "witches" of all ages and sexes. Witches were *not* a persecuted minority, because witches did not exist: the people hurt or killed in the hunts were not witches but victims forced by their persecutors into a category that in reality included no one. The witch-hunts did *not* prosecute, let alone execute, millions; they were not a conspiracy by males, priests, judges, doctors, or inquisitors against members of an old religion or any other real group.

A sorcerer, magician, or "witch" attempts to influence the surrounding world through occult (*i.e.,* hidden) means. In Western society until the 14th century, "witchcraft" had more in common with sorcery in other cultures—such as those of India or Africa—than it did with the witchcraft of the witch-hunts. Before the 14th century, witchcraft was much alike in villages from Ireland to Russia and from Sweden to Sicily; the similarities derived not from cultural diffusion or from secret cults but from the age-old human desire to achieve one's purposes whether by open or occult means. In many ways, like their counterparts worldwide, early Western sorcerers and witches worked secretly for private ends, as contrasted with the public practice of religion. They were usually feared as well as respected, and they used a variety of means to attempt to achieve their goals, including incantations (formulas or chants invoking evil spirits), divination and oracles (to predict the future), amulets and charms (to ward off hostile spirits and harmful events), potions or salves, and dolls or other figures (to represent their enemies). Witches sought to gain or preserve health, to acquire or retain property, to protect against natural disasters or evil spirits, to help friends, and to seek revenge. Often the magic was instead an effort to construct symbolic reality. Sorcery was sometimes believed to rely on the power of gods or other spirits, leading to the belief that witches used demons in their work. (J.B.R.)

WITCHCRAFT IN AFRICA AND THE WORLD

General perceptions. The same dichotomy between sorcery and witchcraft exists (sometimes more ambiguously) in the beliefs of many African and indigenous peoples throughout the world. The Navajo believe that witches meet at night, wear nothing except a mask and jewelry, sit among baskets of corpses, and have intercourse with dead women. In some African cultures witches are believed to assemble in cannibal covens, often at graveyards or around a fire, to feast on blood, which, like vampires, they extract from their victims. If they take the soul from a victim's body and keep it in their possession, the victim will die. Like those in Western society suspected of child abuse and satanism, in the popular imagination, African witches are believed to practice incest and other perversions. Sometimes, as in the Christian tradition, their malevolent power is believed to derive from a special relationship with an evil spirit with whom they have a "pact," or they exercise it through "animal familiars" (assistants or agents) such as dogs, cats, hyenas, owls, or baboons.

The Zande

In other cases the witch's power is thought to be based in his or her own body. Among the Zande of the Congo area and some other central African peoples, the source of this evil-working capacity is believed to be located in the witch's stomach, and its power and range increase with age. It can be activated merely by wishing someone ill and is thus a kind of unspoken, or implicit, curse. At the same time the Zande believe that evil deeds can be wrought even more effectively by the manipulation of spells and potions and the use of powerful magic. In anthropological terminology this is technically "sorcery," and thus the Zande practitioners may more properly be termed "sorcerers."

In many African cultures witches are believed to act unconsciously; unaware of the ill they cause, they are driven by irrepressible urges to act malevolently. It is thus easy for those accused of witchcraft, but who are not conscious of wishing anyone ill, to assume that they unknowingly did what is attributed to them. This, along with the effects of suggestion and torture, in a world where people take the reality of witchcraft for granted, goes far to explain the striking confessions of guilt that are so widely reported in Africa and elsewhere and that are otherwise hard to comprehend. It is worth noting, however, that, if witches believe they are unconscious agents, this is generally not the view of those who feel victimized by them.

Witchcraft and reason. Whatever the basis of their power, witches (and sorcerers) are regularly credited with causing all manner of disease and disaster. Sickness, and even death, as well as a host of lesser misfortunes, are routinely laid at their door. In many parts of Africa and Asia, epidemics and natural disasters have been interpreted as acts of witchcraft. For some unhappy candidates in many less developed countries, the same malign influence is cited to explain (at least in part) failure in examinations, elections, or finding employment. Members of certain Afro-Brazilian cults, for example, believe that job loss is due to witchcraft, and they participate in a ritual, the "consultation," to counter the evil.

However, like their ancient and early modern European counterparts, modern Africans and Asians who believe firmly in the reality of witchcraft do not lack the power of rational reasoning. To suppose that these are incompatible alternatives is a common mistake. In reality, pragmatic and mystical explanations of events usually exist in parallel or combination but operate in different contexts and at different levels. For example, research has shown that African farmers who believe in witches do not expect witchcraft to account for obvious technical failures. If one's home collapses because it was poorly constructed, no witch is needed to explain this. Witchcraft enters the picture when rational knowledge fails. It explains the diseases whose causes are unknown, the mystery of death, and, more generally, strange and inexplicable misfortunes. Moreover, like ancient and modern Westerners, people in modern Africa and other parts of the world who take the reality of witchcraft for granted usually also believe in other sources of supernatural power—*e.g.*, divinities and spirits.

Explanation of misfortune

There is thus no inconsistency in the actions of the sick African who consults both a medical doctor and a witch doctor. The first treats the external symptoms, while the second uncovers the hidden causes. Just as the sick African takes preventative measures prescribed by the medical doctor, he or she might also take steps against the supernatural. To protect against witchcraft, the patient might wear amulets, take "medicine," or practice divination. Similarly, the Navajo protect themselves against witches with "gall medicine" or with sand or pollen paintings. If preventative measures prove ineffective for the Navajo, then the confession of a witch is thought to cure the evil magic, and torture is sometimes used to extract that confession.

Witchcraft explains the problem posed when one seeks to understand why misfortune befalls oneself rather than someone else. It makes sense of the inequalities of life: the fact that one person's crops or herds fail while others' prosper. Equally, witchcraft can be invoked to explain the success of others. In this "limited good" scenario—where there is implicitly a fixed stock of resources and where life is generally precarious, with little surplus to distribute in time of need—those who succeed too flagrantly are assumed to do so at the expense of others less fortunate. The "witch," therefore, is typically someone who selfishly wants more than he or she ostensibly deserves, whose aspirations and desires are judged excessive and illegitimate.

Moral issues. However, there is a narrow, ambiguous line between good and evil here. Among some African peoples "witchcraft" is intrinsically neither morally good nor bad, and among others the supernatural activities of "witches" are, according to their perceived effects, divided into good, or protective, and bad, or destructive, witchcraft. Traditional and modern African leaders sometimes surround themselves with protective "witch doctors" and are themselves thought to be endowed with supernatural power. This is the positive charisma of which witchcraft is the negative counterpart. In the colonial period these ideas were extended to Europeans, who, in the Belgian Congo and British Central Africa at the time of independence, were feared as cannibalistic witches. This was somewhat ironic since colonial regimes, unlike their missionary predecessors, did not believe in witchcraft and made accusations of witchcraft illegal in most of sub-Saharan Africa—which has been largely reversed by their successor regimes.

This ambiguity between good and evil can also be found among the Mapuche, an indigenous people of Chile. They believe that young women take up sorcery and as old women become powerful witches who use "bad medicine" to obtain their ends. They are aligned with evil forces and use them to harm or gain advantage over others. Their training and use of plants and animals in their medicine is similar to that of the shamans who use "good medicine" and other magic against forces of evil.

The distinctions between good and bad supernatural power are relative and depend on how moral legitimacy is judged. This becomes clear when the spiritual power invoked is studied more closely. In a number of revealing African cases, the word that denotes the essence of witchcraft (*e.g.*, *tsau* among the West African Tiv and *itonga* among the East African Safwa), the epitome of illegitimate antisocial activity, also describes the righteous wrath of established authority, employed to curse wrongdoers.

This essential ambivalence is particularly evident in Haitian voodoo, where there is a sharp distinction between man-made evil magic powers, connected with zombies (beings identified as familiars of witches in the beliefs of some African cultures), and benevolent invisible spirits identified with Roman Catholic saints. This antithesis between witchcraft and religion, however, is always problematic: after his death, the malevolent spirits or powers that an ancestor has used for his personal benefit become accrued by his descendants' protective spirits (*loas*).

So everything depends on the moral evaluation made by the community of the victims of misfortune: have they received their just deserts or is their plight unjustified? Witchcraft and sorcery are involved only in the latter case, where they provide a moral philosophy of unmerited misfortune. This is particularly important in religions that lack the concepts of heaven and hell. Where one cannot take refuge in the reassuring belief that life's injustices will be adjusted in the hereafter, witchcraft indeed provides a way of shrugging off responsibility and of coming to terms with an unjust fate. According to these "instant" religions, the just should prosper and the unjust should suffer the consequences of their evil deeds here on earth.

Psychodynamics of witchcraft

The psychodynamics here are equally revealing. Those who interpret their misfortunes in terms of witchcraft will often use similar means to discover the source of their woes, which is often traced to the malice and jealousy of their enemies. In Africa and elsewhere, the bewitched person seeks help from a diviner to establish the evil person responsible. The diviner, often in a trance, uses a number of different techniques to discover the witch, including throwing dice or opening a Bible or Qur'ān at random. Another form of divination involves administering poison to a chicken and mentioning the name of a suspected witch. If the chicken dies, then the suspect is a witch. Whatever the process, the result is always the same: the bewitched "victim" finds the source of his woes among his rivals, typically neighbours, coworkers, or other competitors. Accusations often follow the lines of community conflict. In Chile the tensions between the Mapuche and neighbouring Chilean peasants are revealed in accusations that the Chileans use witchcraft to cheat the Mapuche and conversely that the Mapuche use it to harm the crops or livestock of the Chileans. Among the Navajo, competition over grazing lands and water rights or between jealous lovers is the source of witchcraft accusations. In some polygynous societies in Africa, these accusations are particularly prevalent between competing co-wives, but they are by no means always targeted at women. Ultimately, the effect of successful accusations is to call into question or to rupture an untenable relationship. (I.M.L.)

THE WITCH-HUNTS

Creation of a stereotype. Although accusations of witchcraft in contemporary cultures provide a means to express or resolve social tensions, these accusations had different consequences in premodern Western society where the mixture of irrational fear and a persecuting mentality led to the emergence of the witch-hunts. In the 11th century attitudes toward witchcraft and sorcery began to change, a process that would radically transform the Western perception of witchcraft and associate it with heresy and the Devil. By the 14th century, fear of heresy and of Satan had added charges of diabolism to the usual indictment of witches, *maleficium* (malevolent sorcery). It was this combination of sorcery and its association with the Devil that made Western witchcraft unique. From the 14th through the 18th century, witches were believed to repudiate Jesus Christ, to worship the Devil and make pacts with him, to employ demons to accomplish magical deeds, and to desecrate the crucifix and the consecrated bread and wine of the Eucharist (Holy Communion). It was also believed that they rode through the air at night to "sabbats" (secret meetings), where they engaged in sexual orgies and even had sex with Satan; that they changed shapes (from human to animal or from one human form to another); that they often had "familiar spirits" in the form of animals; and that they kidnapped and murdered children for the purpose of eating them or rendering their fat for magical ointments. This fabric of ideas was a fantasy. Although some people undoubtedly practiced sorcery with the intent to harm, and some may actually have worshiped the Devil, in reality no one ever fit the concept of the "witch."

Since 1970 careful research has elucidated law codes and theological treatises from the era of the witch-hunts and uncovered much information about how fear, accusations, and prosecutions actually occurred in villages, local law courts, and courts of appeal in Roman Catholic and Protestant cultures in western Europe. Charges of *maleficium* were prompted by a wide array of suspicions. It might have been as simple as one person blaming his misfortune on another. The most common suspicions concerned livestock, crops, storms, disease, property and inheritance, sexual dysfunction or rivalry, family feuds, marital discord, stepparents, sibling rivalries, and local politics. *Maleficium* was a threat not only to individuals but also to public order, for a community wracked by suspicions about witches could split asunder. No wonder the term *witch-hunt* has entered common political parlance to describe such campaigns as that of Senator Joseph McCarthy in his attempt to root out "communists" in the United States in the 1950s.

Witchcraft, heresy, and the Devil. Another accusation that often accompanied *maleficium* was trafficking with evil spirits. In the Middle East—in ancient Mesopotamia, Syria, Canaan, and Palestine—belief in the existence of evil spirits was universal, so that both religion and magic were thought to be needed to appease, offer protection from, or manipulate these spirits. In Greco-Roman civilization, Dionysiac worship included meeting underground at night, sacrificing animals, practicing orgies, feasting, and drinking. Classical authors such as Aeschylus, Horace, and Virgil described sorceresses, ghosts, furies, and harpies with hideous pale faces and crazed hair; clothed in rotting garments, they met at night and sacrificed both animals and humans. A bizarre set of accusations, including the sacrifice of children, was made by the Syrians against the Jews in Hellenistic Syria in the 2nd century BC. These accusations would also be made by the Romans against the Christians, by early Christians against heretics (dissenters from the core Christianity of the period) and Jews, by later Christians against witches, and, as late as the 20th century, by Protestants against Catholics.

Ancient accusations

Along with this older tradition, attitudes toward witches and the witch-hunts of the 14th–18th centuries stemmed from a long history of the church's theological and legal attacks on heretics. Accusations similar to those expressed by the ancient Syrians and early Christians appeared again in the Middle Ages. In France in 1022 a group of heretics in Orléans was accused of orgy, infanticide, invocations of demons, and use of the dead children's ashes in a blasphemous parody of the Eucharist. These allegations would have important implications for the future because they were part of a broader pattern of hostility toward and persecution of marginalized groups. This pattern took shape in 1050–1300, which was also an era of enormous reform, reorganization, and centralization in both the ecclesiastical and secular aspects of society, an important aspect of which was suppressing dissent. The visible role played by women in some heresies during this period may have contributed to the stereotype of the witch as female.

Witches in the Middle Ages

The Devil, whose central role in witchcraft beliefs made the Western tradition unique, was an absolute reality in both elite and popular culture, and failure to understand the prevailing terror of Satan has misled some modern researchers to regard witchcraft as a "cover" for political or gender conspiracies. The Devil was deeply and widely feared as the greatest enemy of Christ, keenly intent on destroying soul, life, family, community, church, and state. Witches were considered Satan's followers, members of an antichurch and an antistate, the sworn enemies of Christian society in the Middle Ages, and a "counter-state" in the early modern period. If witchcraft existed, as people believed it did, then it was an absolute necessity to extirpate it before it destroyed the world.

Because of the continuity of witch trials with those for heresy, it is impossible to say when the first witch trial occurred. Even though the clergy and judges in the Middle Ages were skeptical of accusations of witchcraft, the period 1300–30 can be seen as the beginning of witch trials. In 1374 Pope Gregory XI declared that all magic was done with the aid of demons and thus was open to prosecution for heresy. Witch trials continued through the 14th and early 15th centuries, but with great inconsistency according to time and place. By 1435–50, the number of prosecutions had begun to rise sharply, and toward the end of the 15th century, two events stimulated the hunts: Pope Innocent VIII's publication in 1484 of the bull *Summis desiderantes affectibus* ("Desiring with the Greatest Ardour"), condemning witchcraft as Satanism, the worst of all possible heresies, and the publication in 1486 of Heinrich Krämer and Jacob Sprenger's *Malleus maleficarum* ("The Hammer of Witches"), a learned but cruelly misogynist book blaming witchcraft chiefly on women. Widely influential, it was reprinted numerous times. The hunts were most severe from 1580 to 1630, and the last known execution for witchcraft was in Switzerland in 1782. The number of trials and executions varied widely according to time and place, but in fact no more than about 110,000 persons in all were tried for witchcraft, and no more than 40,000 to 60,000 executed. Although these figures are alarming, they do not remotely approach the feverishly exaggerated claims of some 20th-century writers.

Institutions of the witch-hunts. The "hunts" were not pursuits of individuals already identified as witches but ef-

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co., Ltd.



Devil and witches trampling a cross, from *Compendium maleficarum*, 1608.



Frontispiece to *The Discovery of Witches*, a pamphlet prepared in 1647 by Matthew Hopkins, English witchfinder, in which he explained his methods. Hopkins is shown with two of the women he accused and the familiars he claimed were theirs. The woman on the left names Sir John Holt, chief justice of England, who did not believe Hopkins' evidence of witchcraft was genuine. In the British Museum.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.

forts to identify those who were witches. The process began with suspicions and, occasionally, continued through rumours and accusations to convictions. The overwhelming majority of processes, however, went no further than the rumour stage, for actually accusing someone of witchcraft was a dangerous and expensive business. Accusations originated with the ill-will of the accuser, or, more often, the accuser's fear of someone having ill-will toward him. The accusations were usually made by the alleged victims themselves, rather than by priests, lords, judges, or other "elites." Successful prosecution of one witch sometimes led to a local hunt for others, but larger hunts and regional panics were confined (with some exceptions) to the years from the 1590s to the 1640s.

Geographic distribution

Three-fourths of European witch-hunts occurred in western Germany, the Low Countries, France, northern Italy, and Switzerland, areas where prosecutions for heresy had been plentiful and charges of diabolism were prominent. In Spain, Portugal, and southern Italy, witch prosecutions seldom occurred, and executions were very rare. There were additional hunts in Spanish America, where the European pattern of accusations continued even though the differences between the folklore of the Europeans and Native Americans introduced some minor variations into the accusations. In Mexico the Franciscan friars linked indigenous religion and magic with the Devil; prosecutions for witchcraft in Mexico began in the 1530s, and by the 1600s indigenous peasants were reporting stereotypical pacts with the Devil. Like the Spanish colonies, the English colonies repeated the European stereotype with a few minor differences. The first hanging for witchcraft in New England was in 1647, after the witch-hunts had already abated in Europe, though a peculiar outbreak in Sweden in 1668–76 bore some similarity to that in New England. Although the lurid trials at Salem (now in Massachusetts) continue to draw much attention from American authors, they were only a swirl in the backwater of the witch-hunts. The outbreak at Salem, where 19 people were executed, was the result of a combination of church politics, family feuds, and hysterical children, all in a vacuum of political authority. Prosecutions of witches in Austria, Poland, and Hungary took place as late as the 18th century.

Responsibility for witch-hunts

The responsibility for the witch-hunts can be distributed among theologians, legal theorists, and the practices of secular and ecclesiastical courts. The theological worldview—derived from the early Christian fear of Satan and reinforced by the great effort to reform and conform that began in 1050—was intensified again by the fears and animosities engendered by the Reformation of the 16th century. The Protestant Reformation and Catholic Counter-Reformation heightened the fear of witchcraft by promoting the idea of personal piety (the individual alone with his or her Bible and God), which enhanced individualism while downplaying community. The emphasis on personal piety exacerbated the rigid characterization of people as either "good" or "bad." It also aggravated feelings of guilt and the psychological tendency to project negative intentions onto others. Moreover, just as the growth of literacy and of reading the Bible helped spread dissent, so did they provoke resistance and fear. Sermons and didactic treatises, including "devil books" warning of Satan's power, spread both the terror of Satan and the corresponding frantic need to purge society of him. Both Protestants and Catholics were involved in the prosecutions, as the theology of the Protestant Reformers on the Devil and witchcraft was virtually indistinguishable from that of the Catholics. More differences existed among Protestants and among Catholics than between the two religious groups, and regions in which Protestant-Catholic tensions were high did not produce significantly more trials than other regions.

Because accusations and trials of witches took place in both ecclesiastical and secular courts, the law played at least as important a role as religion in the witch-hunts. Local courts were more credulous and therefore more likely to be strict and even violent in their treatment of supposed witches than were regional or superior courts. Crude practices such as pricking witches to see whether the Devil had desensitized them to pain; searching for the "devil's mark," an oddly shaped mole or wart; or "swimming" (throwing the accused into a pond; if she sank, she was innocent because the water accepted her) occurred on the local level. Where central authority—*i.e.*, bishops, kings, or the Inquisition—was strong, convictions were fewer and sentences milder. Ecclesiastical and civil authorities usually tried to restrain witch trials and rarely manipulated witch-hunts to obtain money or power.

The witch executions occurred in the early modern period, the time in Western history when capital punishment and torture were most widespread. Judicial torture, happily in abeyance since the end of the Roman period, was revived in the 12th and 13th centuries; other brutal and sadistic tortures occurred but were usually against the law. Torture was not allowed in witch cases in Italy or Spain, but where used it often led to convictions and the identification of supposed accomplices. The latter was the greatest evil of the system, for a victim might be forced to name acquaintances, who were in turn coerced into naming others, creating a long chain of accusations. Witch trials were equally common in ecclesiastical and secular courts before 1550, and then, as the power of the state increased, they took place more often in secular ones.

Inquisition

Among the main effects of the papal judicial institution known as the Inquisition was in fact the restraint and reduction of witch trials that resulted from the strictness of its rules. It investigated whether the charges resulted from personal animosity toward the accused; it obtained physicians' statements; it did not allow the naming of accomplices either with or without torture; it required the review of every sentence; and it provided for whipping, banishment, or even house arrest instead of death for first offenders. Like the Inquisition, the Parlement of Paris (the supreme court of northern France) severely restrained the witch-hunts. After an outbreak of hunts in France in 1587–88, increasingly skeptical judges began a series of restraining reforms marked by the requirement of "obligatory appeal" to the Parlement in cases of witchcraft, making accusations even more expensive and dangerous.

Decline. The decline of witch-hunts, like their origins, was gradual. By the late 16th century, many prosperous

and professional people in western Europe were accused, so that the leaders of society began to have a personal interest in checking the hunts. The legal use of torture declined in the 17th and 18th centuries, and there was a general retreat from religious intensity following the wars of religion (from the 1560s to 1640s). The gradual demise during the late 17th and early 18th century of the previous religious, philosophical, and legal worldview encouraged an existent but often suppressed skepticism; increasing literacy, mobility, and means of communication set the stage for acceptance of this changing outlook.

Nevertheless, the reasons for the decline in the witch-hunts are as difficult to discern as the reasons for their origins. The theory best supported by the evidence is that the increasing power of the centralized courts such as the Inquisition and the Parlement acted to begin a "decriminalization" of witchcraft. These courts reduced the number of witch trials significantly by 1600, half a century before legal theory, legislation, and theology began to dismiss the notion of witchcraft in France and other countries.

Modern theories. Explanations of the witch-hunts continue to vary, but recent research has shown some of these theories to be improbable or of negligible value. Most scholars agree that the prosecutions were not driven by political or gender concerns; they were not attacks on backward, or rural, societies; they did not function to express or relieve local tensions; they were not a result of the rise of capitalism or other macroeconomic changes; they were not the result of changes in family structure or in the role of women in society; and they were not an effort by cultural elites to impose their views on the populace. Moreover, the evidence does not indicate a close correlation between socioeconomic tension and witchcraft, though agrarian crises seem to have had some effect.

Women and the witch-hunts

One of the most important aspects of the hunts remains unexplained. No satisfactory explanation for the preponderance of women among the accused has appeared. Although the proportions varied according to region and time, on the whole about three-fourths of convicted witches were female. Women were certainly more likely than men to be economically and politically powerless, but that generalization is too broad to be helpful, for it holds true for societies in periods where witchcraft is absent. The malevolent sorcery more often associated with men, such as harming crops and livestock, was rarer than that ascribed to women. Young women were sometimes accused of infanticide, but midwives and nurses were not particularly at risk. Older women were more frequently accused of casting malicious spells than were younger women, because they had had more time to establish a bad reputation, and the process from suspicion to conviction often took so long that a woman might have aged considerably before charges were actually advanced. Although many witchcraft theorists were not deeply misogynist, many others were, notably the authors of the infamous *Malleus maleficarum*. Resentment and fear of the power of the "hag," a woman released from the constraints of virginity and then of maternal duties, has been frequently described in Mediterranean cultures. Folklore and accounts of trials indicate that a woman who was not protected by a male family member might have been the most likely candidate for an accusation, but the evidence is inconclusive. Children were often accusers (as they were at Salem), but they were sometimes also among the accused. Most accused children had parents who had been accused of witchcraft.

In the long run it may be better simply to describe the witch-hunts than to try to explain them, since the explanations are so diverse and complicated. Yet one general explanation is valid: the unique character of the witch-hunts was consistent with the prevailing worldview of intelligent, educated, experienced people for more than three centuries.

CONTEMPORARY WITCHCRAFT

Academics tend to dismiss contemporary witchcraft (known as "Wicca"), at the heart of the modern Neo-Pagan movement, as a silly fad or an incompetent technology, but some now understand it as an emotionally consistent but deliberately anti-intellectual set of practices. At the start of the 21st century, perhaps a few hundred thousand people (mostly in North America and Britain) practiced Wicca and Neo-Paganism, a modern Western reconstruction of pre-Christian religions that draws upon the diversity of worldwide polytheistic religions to create a new and diverse religious movement. The rise of Wicca and Neo-Paganism is due in part to increasing religious tolerance and syncretism, a growing awareness of the symbolism of the unconscious, the retreat of Christianity, the popularity of fantasy and science fiction, the growth of feminism, the ascendancy of deconstructionist and relativist theory, and the emphasis upon individuality and subjectivity. Most modern Neo-Pagans, distrustful of the demands of traditional religions, eschew doctrine and engage in the ritual expression of "symbolic and experiential" meanings. Although Neo-Paganism incorporates the emotional involvement and ritual practices associated with religion into its tradition, many Neo-Pagans prefer to think of themselves as practicing magic rather than religion, and, although their emphasis is on opening themselves up to hidden powers through rites, chants, or charms, most do not call themselves "witches," as Wiccans do. Both Wiccans and Neo-Pagans also have strong ecological and environmental concerns, worship the Goddess and other deities, and celebrate the change of seasons with elaborate rituals. Whether magic or religion, these groups reject intellectual coherence and objectivity in favour of personal experience and dismiss science as well as traditional religion.

Wicca and Neo-Paganism

Although some Wiccans claim to be part of the "old ways" and "ancient tradition," their religion is new. Wicca is creative, imaginative, and entirely a 20th-century invention, with no connection to ancient paganism or the alleged "witches" of the witch-hunts. No cult of the "Goddess" played a significant role in Western culture between late antiquity and the mid-20th century. Wicca, in fact, originated about 1939 with an Englishman, Gerald Gardner, who constructed it from the fanciful works of the self-styled magician Aleister Crowley; the fake "ancient" document *Aradia* (1899); the Hermetic Order of the Golden Dawn and other late-19th- and early-20th-century occult movements; and Margaret Murray's *The Witch-Cult in Western Europe* (1921) and article "Witchcraft" in the 14th edition of the *Encyclopædia Britannica* (1929), which put forth in its most popular form her theory that the witches of western Europe were the lingering adherents of a once general pagan religion that had been displaced, though not completely, by Christianity. Gardner, backed by Murray, who wrote a laudatory introduction to his book *Witchcraft Today* (1954), fixed this erroneous notion of an ancient witch-cult in the public consciousness, and it has been nurtured there by Robert Graves's *The White Goddess* (1948) and innumerable more recent quasi-fictional and fictional accounts. (J.B.R.)

BIBLIOGRAPHY

General works. RICHARD CAVENDISH (ed.), *Man, Myth, & Magic: The Illustrated Encyclopedia of Mythology, Religion, and the Unknown*, new ed. edited and comp. by YVONNE DEUTCH, 12 vol. (1983), generally reliable; WILLIAM A. LESSA and EVON Z. VOGT (eds.), *Reader in Comparative Religion: An Anthropological Approach*, 4th ed. (1979), with articles on divination, magic, and witchcraft; LESLIE SHEPARD (ed.), *Encyclopedia of Occultism & Parapsychology*, 2nd ed., 3 vol. (1984–85); and LYNN THORNDIKE, *A History of Magic and Experimental Science*, 8 vol. (1923–58, reissued from various printings, 1964), which is concerned with Europe from the period of the Roman Empire through the 17th century.

Hermetism. HERBERT LEVENTHAL, *In the Shadow of the Enlightenment: Occultism and Renaissance Science in Eighteenth Century America* (1976); A.D. NOCK (ed.) and A.-J. FESTUGIÈRE (trans.), *Corpus Hermeticum*, 4 vol. (1945–54, reprinted 1980); WAYNE SHUMAKER, *The Occult Sciences in the Renaissance: A Study in Intellectual Patterns* (1972, reprinted 1979); JAMES WEBB, *The Occult Establishment* (1976, reissued 1981), studies of occultist movements in Europe in the 19th and 20th centuries; EDGAR WIND, *Pagan Mysteries in the Renaissance*, rev. and enlarged ed. (1968, reissued 1980), the influence of Neoplatonic and esoteric thought on Renaissance arts; and FRANCES A. YATES, *The Rosicrucian Enlightenment* (1972, reissued 1978) and *The Occult Philosophy in the Elizabethan Age* (1979, reis-

sued 1983), on the Hermetic-Kabbalist tradition and its place in 16th- and 17th-century thought in Europe.

Alchemy. ALAN PRITCHARD, *Alchemy: A Bibliography of English-Language Writings* (1980), a comprehensive work, valuable for its listing of secondary material; TITUS BURCKHARDT, *Alchemy: Science of the Cosmos, Science of the Soul* (1967; originally published in German, 1960); C.A. BURLAND, *The Arts of the Alchemists* (1967), a valuable source for iconography; MIRCEA ELIADE, *The Forge and the Crucible,* 2nd ed. (1978; originally published in French, 1956), a comparison of alchemy with the practices of the primitive metallurgist; RICHARD RUSSELL (trans.), *The Works of Geber* (1678, reprinted 1928; originally published in Latin, 1545), the most famous and important medieval Latin work on alchemy; E.J. HOLMYARD, *Alchemy* (1957, reprinted 1968), a general history; C.G. JUNG, *Psychology and Alchemy,* 2nd rev. ed. (1968, reprinted 1980; originally published in German, 2nd rev. ed., 1952), a classic psychological interpretation of alchemy; ROBERT P. MULTHAUF, *The Origins of Chemistry* (1967), a discussion of Greek, Arabic, and European alchemy, with an extensive bibliography; PRAPHULLA CHANDRA RAY, *A History of Hindu Chemistry from the Earliest Times to the Middle of the Sixteenth Century, A.D.,* 2 vol. (1902–09), a valuable study; HERBERT SILBERER, *Problems of Mysticism and Its Symbolism* (1917, reprinted 1971 as *Hidden Symbolism of Alchemy and the Occult Arts;* originally published in German, 1914), an early Freudian interpretation; NATHAN SIVIN, *Chinese Alchemy: Preliminary Studies* (1968); and *Ambix* (3 times per year), the journal of the Society for the History of Alchemy and Early Chemistry and the principal source of new historical research.

Divination. *Astrology:* AUGUSTE BOUCHÉ-LECLERCQ, *L'Astrologie grecque* (1899, reprinted 1979), the fundamental work on Greek astrology; JEAN-LOUIS BRAU, HELEN WEAVER, and ALLAN EDWARDS, *Larousse Encyclopedia of Astrology* (1980, reissued 1982; originally published in French, 1977), useful for modern astrology; FRANZ CUMONT, *Astrology and Religion Among the Greeks and Romans* (1912, reprinted 1960), a review of astrology's position in pagan religions; ROBERT EISLER, *The Royal Art of Astrology* (1946), an excellent critical history; EUGENIO GARIN, *Astrology in the Renaissance: The Zodiac of Life* (1983; originally published in Italian, 1976), on astrology's relationship to other aspects of Hermetism; MICHEL GAUQUELIN, *The Scientific Basis of Astrology: Myth or Reality* (1969; originally published in French, 1964); ELLIC HOWE, *Urania's Children: The Strange World of the Astrologers* (1967), a history of astrology in England and Germany in the 19th and 20th centuries; E.S. KENNEDY and DAVID PINGREE, *The Astrological History of Māshā' allāh* (1971), an Arabic text and English translation on Sāsānian astrology; HANS LEWY, *Chaldaean Oracles and Theurgy: Mysticism, Magic and Platonism in the Later Roman Empire,* new ed. by MICHEL TARDIEU (1978), on astrology's role in contemporary practices; CARLO ALFONSO NALLINO, "Astrologia e astronomia presso i Musulmani. 1. Astrologia," in his *Raccolta di scritti editi e inediti,* ed. by MARIA NALLINO, vol. 5, *Astrologia, astronomia, geografia,* pp. 1–41 (1944), the best discussion of astrology in Islām; DAVID PINGREE, *Census of the Exact Sciences in Sanskrit,* 4 vol. (1970–81), an account of Indian astrologers; and G.J. TOOMER (trans.), *Ptolemy's Almagest* (1984), the source of all subsequent astrological theory.

Other forms: WILLIAM BARRETT and THEODORE BESTERMAN, *The Divining Rod* (1926, reissued 1968), with a bibliography of water divining; WILLIAM R. BASCOM, *Ifa Divination: Communication Between Gods and Men in West Africa* (1969); HENRY CALLAWAY, *The Religious System of the Amazulu* (1870, reissued 1970), with an analysis of their divinatory practices; GEORGES DUMÉZIL, *Archaic Roman Religion,* 2 vol. (1970; originally published in French, 1966), with an appendix on the religion of the Etruscans; ROBERT FLACELIÈRE, *Greek Oracles,* 2nd ed. (1976; originally published in French, 1961); WILLIAM A. LESSA, *Chinese Body Divination: Its Forms, Affinities and Functions* (1968); MICHAEL LOEWE and CARMEN BLACKER (eds.), *Divination and Oracles* (1981), nine studies covering the ancient and Oriental worlds; RENÉ DE NEBESKY-WOJKOWITZ, *Oracles and Demons of Tibet* (1956, reprinted 1976); VICTOR W. TURNER, *Ndembu Divination: Its Symbolism & Techniques* (1961, reprinted 1969); and EVON Z. VOGT and RAY HYMAN, *Water Witching, U.S.A.,* 2nd ed. (1979), an ethnographic study.

Magic. ARTHUR C. LEHMANN and JAMES E. MYERS (compilers), *Magic, Witchcraft, and Religion: An Anthropological Study of the Supernatural,* 5th ed. (2001), is an important collection of essays; and NINIAN SMART, *Worldviews: Crosscultural Explorations of Human Beliefs,* 3rd ed. (2000), is an introduction to various belief systems using the six-dimension approach. Among the most important early works on magic are EDWARD B. TYLOR, *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Language, Art, and Custom,* 3rd American ed., 2 vol. (1889, reprinted 1977); ÉMILE DURKHEIM, *The Elementary Forms of the Religious Life* (1915, reissued 1976; originally published in French, 1912); and J.G. FRAZER, *The Golden Bough,* 3rd ed., 12 vol. (1911–15). Other significant early works are BRONISLAW MALINOWSKI, *Magic, Science, and Religion, and Other Essays,* ed. by Robert Redfield (1948, reprinted 1992); MARCEL MAUSS, *A General Theory of Magic* (1972, reprinted 1975; originally published in French, 1902–03), an early anthropological study; and SIGMUND FREUD, *Totem and Taboo* (1950, reissued 1999; originally published in German, 1912–13), a psychoanalytic interpretation. Studies that employ a context-sensitive or cross-cultural approach include MIRCEA ELIADE, *Shamanism: Archaic Techniques of Ecstasy* (1964, reissued 1989; originally published in French, 1951); CLAUDE LÉVI-STRAUSS, *Introduction to the Work of Marcel Mauss* (1987; originally published in French, 1950); and STANLEY JEYARAJA TAMBIAH, *Magic, Science, Religion, and the Scope of Rationality* (1990). JACOB NEUSNER, ERNEST S. FRERICHS, and PAUL VIRGIL MCCRACKEN FLESHER (eds.), *Religion, Science, and Magic: In Concert and in Conflict* (1989, reissued 1992), is a collection of essays on definitional theories and case studies. Useful studies of magic in the ancient world are GEORG LUCK (trans. and ed.), *Arcana Mundi: Magic and the Occult in the Greek and Roman Worlds* (1985, reissued 1987), a collection of texts; and MARVIN MEYER and RICHARD SMITH (eds.), *Ancient Christian Magic: Coptic Texts of Ritual Power* (1994, reissued 1999). CHARLES BURNETT, *Magic and Divination in the Middle Ages: Texts and Techniques in the Islamic and Christian Worlds* (1996); VALERIE I.J. FLINT, *The Rise of Magic in Early Medieval Europe* (1991, reissued 1993); RICHARD KIECKHEFER, *Magic in the Middle Ages* (1989, reissued 2000); and HENRY MAGUIRE (ed.), *Byzantine Magic* (1995), are helpful surveys of medieval attitudes toward magic. KEITH THOMAS, *Religion and the Decline of Magic: Studies in Popular Beliefs in Sixteenth and Seventeenth Century England* (1971, reissued 1997), an influential study based on an anthropological theory; D.P. WALKER, *Spiritual and Demonic Magic: From Ficino to Campanella,* new ed. (1958, reissued 2000), which examines magical theory of the Renaissance; and FRANCES A. YATES, *Giordano Bruno and the Hermetic Tradition* (1964, reissued 1999), are good introductions to magic in the early modern period. Important studies of magic in modern world civilizations are FRANCIS L. HSU, *Exorcising the Trouble Makers: Magic, Science, and Culture* (1983), on China; E.E. EVANS-PRITCHARD, *Witchcraft, Oracles, and Magic Among the Azande* (1937, reissued 1968), on African magic and witchcraft; A.R. RADCLIFFE-BROWN, *The Andaman Islanders* (1922, reissued 1964); and LEE SIEGEL, *Net of Magic: Wonders and Deceptions in India* (1991).

Witchcraft. Good introductions are EDWARD PETERS, *The Magician, the Witch, and the Law* (1978, reissued 1992); L.M. LEWIS, *Religion in Context: Cults and Charisma,* 2nd ed. (1996); and MARY DOUGLAS (ed.) *Witchcraft Confessions & Accusations* (1970). JEFFREY BURTON RUSSELL, *Mephistopheles: The Devil in the Modern World* (1986, reissued 1992), is an especially valuable study of the relationship between witchcraft beliefs and the Devil. LUCY MAIR, *Witchcraft* (1969), is a useful general introduction with emphasis on African examples. Witchcraft and cannibalism are discussed in PEGGY REEVES SANDAY, *Divine Hunger: Cannibalism as a Cultural System* (1986). JEAN COMAROFF and JOHN COMAROFF (eds.), *Modernity and its Malcontents* (1993), examines witchcraft and modern politics in Africa. ROBIN BRIGGS, *Witches & Neighbors: The Social and Cultural Context of European Witchcraft* (1996, reissued 1998), is perhaps the best book on the subject; and ALAN C. KORS and EDWARD PETERS (eds.), *Witchcraft in Europe, 1100–1700: A Documentary History* (1972, reprinted 1995), is an important collection of primary sources on European witchcraft beliefs. Important studies of the origins of witchcraft in medieval and early modern Europe are STUART CLARK, *Thinking with Demons: The Idea of Witchcraft in Early Modern Europe* (1997, reissued 1999); RICHARD KIECKHEFER, *European Witch Trials: Their Foundations in Popular and Learned Culture, 1300–1500* (1976); and JEFFREY BURTON RUSSELL, *Witchcraft in the Middle Ages* (1972, reissued 1984). Among the best studies of witchcraft in the various European nations and New World are ELAINE G. BRESLAW, *Tituba, Reluctant Witch of Salem: Devilish Indians and Puritan Fantasies* (1996); FERNANDO CERVANTES, *The Devil in the New World: The Impact of Diabolism in New Spain* (1994); RUTH MARTIN, *Witchcraft and the Inquisition in Venice, 1550–1650* (1989); and E. WILLIAM MONTER, *Witchcraft in France and Switzerland: The Borderlands during the Reformation* (1976). CHRISTINA LARNER, *Enemies of God: The Witch-Hunt in Scotland* (1981, reissued 1983); and BRIAN P. LEVACK, *The Witch-Hunt in Early Modern Europe,* 2nd ed. (1995), examine the persecutions of the early modern period. AIDAN A. KELLY, *A History of Modern Witchcraft, 1939–1964* (1991); and T.M. LUHRMANN, *Persuasions of the Witch's Craft: Ritual Magic in Contemporary England* (1989, reissued 1994), provide thoughtful discussions of contemporary witchcraft.

(R.P.M./D.E.P./G.K.P./J.F.M./K.L.J./J.B.R./I.M.L./Ed.)

Occupational Diseases and Disorders

Occupational diseases and disorders are those associated with a particular occupation or industry. They result from a variety of biological, chemical, physical, and psychological factors that are present in the work environment or are otherwise encountered in the course of employment. Occupational medicine is concerned with the effect of all kinds of work on health and the effect of health on a worker's ability and efficiency.

Occupational diseases are essentially preventable and can be ascribed to faulty working conditions. The control of occupational health hazards decreases the incidence of work-related diseases and accidents and improves the health and morale of the work force, leading to decreased absenteeism and increased worker efficiency. In most cases the moral and economic benefits far outweigh the costs of eliminating occupational hazards.

This article discusses general occupational health hazards and the disorders they cause, as well as the role of occupational health services. More detailed information about specific disorders can be found in the articles dealing with human diseases and the structures of the human body, such as CANCER; INFECTIOUS DISEASES; and RESPIRATION AND RESPIRATORY SYSTEMS.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 424.

The article is divided into the following sections:

Historical overview 99
- The preindustrial era
- The industrial era
- The 20th century

Aims and functions of occupational health services 100
- Job placement
- Safety training
- Supervision of high-risk groups
- Control of recognized hazards
- Identification of unrecognized hazards
- Treatment
- General health education and surveillance

Disorders due to chemical agents 100
- Gases
- Metals
- Organic compounds
- Dusts

Disorders due to physical agents 102
- Temperature
- Atmospheric pressure
- Noise
- Vibration
- Other mechanical stresses
- Ionizing radiation
- Nonionizing radiation

Disorders due to infectious agents 103

Disorders due to psychological factors 103

Bibliography 103

HISTORICAL OVERVIEW

The preindustrial era. The first recorded observation of an occupational disease may be a case of severe lead colic suffered by a worker who extracted metals. It is described in the third book of *Epidemics,* attributed to Hippocrates, the Greek physician of the 4th century BC. Other early writers also recognized the association between certain disorders and occupations. The Roman scholar Pliny, in the 1st century AD, described mercury poisoning as a disease of slaves because mines contaminated by mercury vapour were considered too unhealthy for Roman citizens and thus were worked only by slaves. In general, however, physicians of antiquity were not concerned with the health of workers.

During the Middle Ages the rise of metalliferous mining in central Europe inspired the German mineralogist Georgius Agricola to make a detailed study of gold-and silver-mining operations. In his *De Re Metallica,* published posthumously in 1556, Agricola described the primitive methods of ventilation and personal protection in use, common mining accidents and disasters, and such miners' occupational diseases as the "difficulty in breathing and destruction of the lungs" caused by the harmful effects of dust inhalation.

A more comprehensive account of occupational disorders was written by Bernardino Ramazzini, a professor of medicine first at the University of Modena and later at the University of Padua. His *De Morbis Artificum Diatriba* (1700; *Diseases of Workers*) contains descriptions of the diseases associated with 54 different occupations, from the mercury poisoning of Venetian mirror makers to the diseases afflicting learned men. Ramazzini believed that a physician must determine the patient's occupation in order to discover the cause of the patient's disorder. He is generally regarded as the father of occupational medicine.

The father of occupational medicine

The industrial era. The Industrial Revolution of the 18th century had a profound impact on occupational diseases. Rapid technological progress and industrial growth had led to crowded, unsanitary working and living conditions, with a corresponding rise in the number of accidents and deaths caused by the new machinery and exposure to toxic materials. In 1775 Percivall Pott, a London surgeon, linked the frequent occurrence of scrotal cancer among chimney sweeps to the soot ingrained into their skin by prolonged exposure to flue dusts. Charles Turner Thackrah, a Leeds physician, further advanced the study of occupational medicine in Britain with his *The Effects of the Principal Arts, Trades and Professions . . . on Health and Longevity . . .* (1831), which described lung diseases caused by dust that commonly afflicted miners and metal grinders. In 1895 Britain introduced a statutory notification system that required medical personnel to report all occurrences of certain diseases to the chief inspector of factories. Other industrial nations followed Britain's lead, and legal provisions for the health of the worker continued to be instituted throughout the 19th and 20th centuries.

The 20th century. Although such classic occupational diseases as lead poisoning and anthrax have declined in incidence in industrialized countries, none have been eradicated. Furthermore, new diseases continue to develop as a result of advances in technology. X rays were discovered in 1895, and 20 years later nearly 100 radiologists were estimated to have died as a result of occupational exposures. Asbestos-related disease was first reported in the first half of the 20th century, and in 1974 hemangiosarcoma, a rare malignant tumour of the liver, was discovered among workers involved in the polymerization of vinyl chloride monomer. Other occupational diseases related to the introduction of industrial processes and materials may well be recognized in the future.

Paralleling the development of new technology and occupational hazards has been the development of occupational health services. No longer concerned primarily with the prevention of industrial accidents and diseases among manual workers, industrial medicine now aims to protect and improve the health of all classes and kinds of workers. In 1950 a joint committee of the International Labour Organisation and the World Health Organization (ILO/WHO) defined the concerns of occupational health as:

The ILO/WHO definition of occupational health

> the promotion and maintenance of the highest degree of physical, mental and social well-being of workers in all occupations; the prevention among workers of departures from

health caused by their working conditions; the protection of workers in their employment from risks resulting from factors adverse to health; the placing and maintenance of the worker in an occupational environment adapted to his physiological equipment and, to summarise, the adaptation of work to man and of each man to his job.

In most countries in the West the responsibility for health and safety at work is placed on the employer, although the government may establish safety standards. Occupational health services are provided as benefits by employers and generally are separate from other community health services. In the former Soviet Union occupational health and hygiene were given high priority and were fully integrated in the general medical care system. In the developing and Third World countries, many of which are undergoing rapid industrialization, the importance of occupational health is increasingly realized. The problems of exposure to occupational hazards, however, are frequently compounded by preexisting malnutrition and a high incidence of infectious disease. Occupational health services in these countries are often most practical and cost-effective, therefore, when combined with primary health care delivery.

AIMS AND FUNCTIONS OF OCCUPATIONAL HEALTH SERVICES

The primary concerns of occupational health services remain those specified by the ILO/WHO in 1950, although work-related diseases are now considered as well as purely occupational diseases. The actual services offered are essentially preventive in nature and are summarized below.

Job placement. People with certain preexisting medical conditions may be at a disadvantage in some jobs. A preemployment health questionnaire or medical examination can be of great value in such cases by determining job unsuitability before training time and expense have been incurred. Job suitability may also need to be regularly monitored in order to assure employee health and ability. Airline pilots, for example, undergo regular medical checkups because a pilot with failing vision or one who suffers from an undetected heart condition that can lead to a heart attack could endanger many lives. The health service can also give valuable advice with regard to alternative employment when a worker is found to be unfit for a particular job.

Safety training. An occupational health service has a responsibility to keep all employees informed about hazards in the workplace. The measures taken to protect employee health should be thoroughly explained so that workers understand the necessity of complying with such irksome or unpleasant restrictions as the wearing of protective clothing and face masks. First aid facilities should be organized and employees instructed about first aid procedures in case of accidental injuries or other emergencies.

Supervision of high-risk groups. Exposure levels considered safe for a young male worker may be hazardous for a pregnant woman (the fetus, especially during the first three months of development, is particularly sensitive to environmental toxic agents). Pregnant women, as well as such other vulnerable groups as the very young, the elderly, and the disabled, therefore require appropriate medical surveillance and advice about specific precautionary measures they can take.

Control of recognized hazards. A complex system of environmental and biological monitoring has been developed for the control of known hazards at work. Occupational health practice is concerned with monitoring the concentration of toxic substances in the environment, determining safe exposure levels, suggesting procedures to limit worker exposure, and monitoring workers for signs of overexposure. Occupational health specialists can also contribute to the prevention of health risks by assisting in the planning and design of new equipment and factories.

Identification of unrecognized hazards. Occupational health services can play a major role in the detection of new health hazards of all types. Clinical observation and study may reveal a causal relationship between patterns of sickness or mortality in groups of workers and their occupational exposure. Examples of hazards identified in this manner include lung and nasal cancer among nickel workers, lung cancer in asbestos workers, and coronary heart disease among workers exposed to carbon disulfide (used in the manufacture of rayon).

Treatment. Quick, on-site treatment of work injuries and poisonings can prevent complications and aid recovery. Such treatment can also be economically beneficial by saving traveling and waiting time. Furthermore, physicians and nurses who are unfamiliar with their patients' working conditions may keep workers with minor injuries away from work longer than necessary. An occupational treatment service offers opportunities for specialized counseling and health education.

General health education and surveillance. Occupational health services may have to provide general medical care for workers and their families in developing countries with inadequate community health services. Even when general health care is provided elsewhere, an occupational health service can offer an effective and often economically advantageous program of health education and counseling. By advising employees on such topics as smoking, alcohol or drug abuse, exercise, and diet, the occupational health service can improve worker health and efficiency and reduce illness and absenteeism. The health service is also in a position to organize employee health surveillance programs for the early diagnosis of disease.

DISORDERS DUE TO CHEMICAL AGENTS

Exposure to hazardous chemicals

Hazardous chemicals can act directly on the skin, resulting in local irritation or an allergic reaction, or they may be absorbed through the skin, ingested, or inhaled. In the workplace ingestion of toxic chemicals is usually accidental and most commonly results from handling contaminated food, drink, or cigarettes. Substances that occur as gases, vapours, aerosols, and dusts are the most difficult to control, and most hazardous chemicals are therefore absorbed through the respiratory tract. If inhaled, airborne contaminants act as irritants to the respiratory tract or as systemic poisons. Toxicity in such cases depends on the contaminant's concentration, particle size, and physicochemical properties, particularly its solubility in body fluids. An individual's reaction to any hazard depends primarily on the length, pattern, and concentration of exposure but is also affected by such factors as age, sex, ethnic group, genetic background, nutritional status, coexistent disease, concomitant exposure to other toxic agents, life-style, and history of previous exposure to the agent in question. The wide range of both naturally occurring and synthetic chemical compounds that can give rise to adverse health effects can be roughly organized into four major categories: gases, metals, organic compounds, and dusts.

Gases. Gases may act as local irritants to inflame mucous surfaces. Common examples include sulfur dioxide, chlorine, and fluorine, which have pungent odours and can severely irritate the eyes and the respiratory tract. Some gases, such as nitrogen oxides and phosgene, are much more insidious. Victims may be unaware of the danger of exposure because the immediate effects of these gases may be mild and overlooked. Several hours after exposure, however, breathlessness and fatal cardiorespiratory failure due to pulmonary edema (collection of fluid in the lungs) may develop.

Asphyxiants

Gases that interfere with oxygen supply to the tissues are known as asphyxiants and are of two principal types. Simple asphyxiants are physiologically inert gases that act by diluting atmospheric oxygen. If the concentration of such gases is high enough, hypoxia (deficiency of oxygen reaching the tissues of the body) results. Victims of mild hypoxia may appear to be intoxicated and may even resist rescue attempts. Common examples of simple asphyxiants are methane and carbon dioxide.

In contrast to simple asphyxiants, chemical asphyxiants, such as carbon monoxide and hydrogen sulfide, are highly reactive. They cause a chemical action that either prevents the blood from transporting oxygen to the tissues or interferes with oxygenation in the tissues. For example, carbon monoxide, a frequently encountered gas produced by incomplete combustion, combines with hemoglobin in the blood and reduces its oxygen-carrying capacity. In low concentration carbon monoxide poisoning can cause symptoms of fatigue, headache, nausea, and vomiting, but

heavy exposure leads to coma and death. It is especially dangerous because it is both colourless and odourless. Hydrogen sulfide, however, can be recognized by its characteristic smell, suggestive of rotten eggs. It is produced when sulfur compounds decompose and acts by inhibiting the respiratory enzyme cytochrome oxidase, thus giving rise to severe tissue hypoxia. In addition to its asphyxiant properties, hydrogen sulfide also acts as an irritant to the eyes and mucous membranes.

Preventing gas poisoning involves preventing exposure. Workers should never enter enclosed spaces that have suspect atmospheres alone; workplaces should provide adequate ventilation, and air should be regularly tested for contamination. If exposure does occur, treatment involves the removal of the victim from the contaminated atmosphere, artificial respiration, and administration of oxygen or recommended antidotes. Victims exposed to gases with insidious delayed effects should be kept under medical observation for an appropriate period.

Metals. Metals and their compounds are among the poisons most commonly encountered in the home and workplace. Even metals essential for life can be toxic if they are present in excessive amounts. Iron, for example, is an essential element and is sometimes given therapeutically; if taken in overdose, however, it can be lethal.

Mercury poisoning

Mercury poisoning, one of the classic occupational diseases, is a representative example of metal poisoning. Exposure to mercury can occur in many situations, including the manufacture of thermometers, explosives, fungicides, drugs, paints, batteries, and various electrical products. The disorders it can cause vary depending on the type of mercury compound and the method of exposure.

Ingestion of mercury salts such as mercuric chloride (corrosive sublimate) leads to nausea, vomiting, and bloody diarrhea. Kidney damage resulting in death may follow in extreme cases. Inhalation or absorption through the skin of mercury vapour causes salivation, loosening of the teeth, and tremor; it also affects the higher centres of the brain, resulting in irritability, loss of memory, depression, anxiety, and other personality changes. This mental deterioration, known as erethism, led to the well-known saying "mad as a hatter," because, in the past, hatters commonly became ill when they used mercury salts to make felt out of rabbit fur. Poisoning with organic mercury compounds (used in fungicides and pesticides) results in permanent neurological damage and can be fatal.

Other hazardous metals commonly encountered in industry include arsenic, beryllium, cadmium, chromium, lead, manganese, nickel, and thallium. Some have been shown to be carcinogenic, including certain compounds of nickel (linked to lung and nasal cancer), chromium (lung cancer), and arsenic (lung and skin cancer).

Organic compounds. The organic compounds that pose the greatest occupational hazards are various aromatic, aliphatic, and halogenated hydrocarbons and the organophosphates, carbamates, organochlorine compounds, and bipyridylium compounds used as pesticides.

Pesticides

Pesticides are used the world over; and, even though precautionary measures (such as using protective clothing and respirators, monitoring contamination of equipment and clothing, keeping workers out of recently sprayed areas, and requiring workers to wash thoroughly after exposure) can be instituted, poisoning not infrequently occurs in agricultural communities. The organophosphates and the generally less toxic carbamates exert their effects by inhibiting cholinesterase, an enzyme that prevents stimulation from becoming too intense or prolonged by destroying the acetylcholine involved in the transmission of impulses in the autonomic nervous system. Cholinesterase inhibitors allow the accumulation of acetylcholine, causing symptoms related to parasympathetic overactivity, such as chest tightness, wheezing, blurring of vision, vomiting, diarrhea, abdominal pain, and in severe cases respiratory paralysis. Atropine and certain oximes counteract their effects.

Paraquat and diquat, the bipyridylium compounds, are deadly if ingested. Skin contact or inhalation of a concentrate of paraquat can cause fatal lung damage. Because no specific antidote is known, treatment consists of minimizing the body's absorption of the poison.

The organochlorine compounds, such as DDT, are being progressively phased out of use. Because they are fat-soluble and very stable, they accumulate and remain in the fatty tissues of the body for prolonged periods. Symptoms of poisoning include nausea, irritability, weakness, muscle tremors, and convulsions. There is no specific antidote.

Other groups of pesticides that are used less frequently or are less hazardous include the organomercury compounds (see above *Metals*); the dinitro and arsenic compounds; and nicotine.

Hydrocarbons

Hydrocarbons are used industrially in the derivation of other compounds and in solvents, degreasing agents, refrigerants, fire extinguishers, dry cleaning agents, paint removers, and other products. Many are volatile and can be absorbed by inhalation; some are fat-soluble and can be readily absorbed following spills on the skin.

Gasoline, fuel oils, and other petroleum products are common examples of aliphatic hydrocarbons. If they are ingested or inhaled, dizziness, weakness, nausea, or irritation of the lungs may follow. In very severe cases victims may become unconscious or experience convulsions. Direct contact causes skin irritation and dryness. Prolonged exposure to certain petroleum oils may result in skin cancer.

The aromatic hydrocarbon benzene provides the basis for the synthesis of many other organic compounds. It is rapidly absorbed following inhalation or skin contact. Symptoms from mild exposure include dizziness, headache, euphoria, confusion, and nausea. Long-term exposure may be followed by bone marrow depression, anemia, spontaneous bleeding, and leukemia. Several aromatic hydrocarbons are known to be carcinogens. Particularly hazardous are naphthylamine, benzidine, and 4-amino diphenyl, which cause bladder cancer. Previously used in the synthetic dye, synthetic rubber, cable-making, and chemical industries, they have been banned in a number of countries.

When aliphatic and aromatic hydrocarbons have hydrogen atoms in their structure replaced by halogens (often chlorine), they are known as halogenated hydrocarbons. In general, increasing the chlorination of aliphatic hydrocarbons increases their toxicity, while the reverse is true of the aromatic series. Many chlorinated hydrocarbons, including chloroform and trichloroethylene, act as depressants on the central nervous system, producing anesthetic or narcotic effects that may be abused. Occupational exposure to many solvents may act synergistically with alcohol, resulting in more damage than either agent could produce on its own. Some halogenated hydrocarbons cause extensive disorders in addition to their common narcotic effect. Inhaling or ingesting the solvent carbon tetrachloride, for example, leads to liver damage; and exposure to vinyl chloride causes Raynaud's phenomenon (spasms in the small arteries that cause the extremities to become pale and cold, as well as painful), necrosis of the small bones of the hand, liver damage, and a rare, highly malignant tumour of the liver.

Workers exposed to hydrocarbons should wear protective clothing or masks when appropriate, moderate alcohol consumption, and verify that work areas are well ventilated and that recommended exposure levels are not exceeded.

Dusts. The inhalation of a variety of dusts is responsible for a number of lung and respiratory disorders, whose symptoms and severity depend on the composition and size of the dust particle, the amount of dust inhaled, and the length of exposure. The lung diseases known as the pneumoconioses result when certain inhaled mineral dusts are deposited in the lungs, where they cause a chronic fibrotic reaction that leads to decreasing capacity for exercise and increasing breathlessness, cough, and respiratory difficulty. No specific treatment is known, but as with all respiratory disorders patients are urged to quit smoking, which aggravates the condition. Suggested measures for limiting exposure include using water and exhaust ventilation to lower dust levels and requiring workers to wear respirators or protective clothing, but such procedures are not always feasible. Coal worker's pneumoconiosis, silicosis, and asbestosis are the most common pneumoconioses.

Pneumoconioses

As its name suggests, coal worker's pneumoconiosis (also

known as black lung) occurs most frequently among coal miners and workers involved in the transporting or processing of coal. It is generally benign in its early stages, but after a variable number of years of exposure to coal dust, progressive massive fibrosis may develop, ending in cardiorespiratory failure. Miners and quarry workers are the people most likely to suffer from silicosis. Because silica is found in many rocks and is used in a variety of industries, workers involved in stonecutting, grinding, drilling, foundry work, sandblasting, pottery making, and the manufacture of abrasives are also at risk. Silicosis is an aggressive form of pulmonary fibrosis that speeds the progress of tuberculosis. Routine chest X rays can aid early diagnosis by revealing abnormal shadowing. Asbestosis is more difficult to detect in the early stages because chest X rays usually reveal little until the disease is advanced. From onset asbestosis progresses more rapidly than the other pneumoconioses and can result from relatively low exposure. Asbestos is the general term for a number of fibrous silicates that are used primarily in various fireproofing, insulation, and cement products. In addition to pulmonary fibrosis, inhaling asbestos fibres has also been shown to cause lung and other cancers.

Asthma

Prolonged exposure to certain plant and animal dusts can cause asthma, even in people without a predisposition for allergies. Specific hazards include dusts from flour, grains, and wood and wood products. Cotton workers and others handling hemp or flax may develop a condition known as byssinosis, similar to asthma. The group of diseases known as farmer's lung, malt worker's lung, bird fancier's lung, and so forth are caused by an allergic inflammatory reaction to the fungal spores present in moldy hay or barley, bird droppings, feathers, and a variety of other organic materials. Symptoms initially resemble those of influenza or pneumonia, but repeated episodes eventually lead to pulmonary fibrosis with chronic respiratory impairment. The only treatment for these disorders is avoiding exposure to the dusts.

DISORDERS DUE TO PHYSICAL AGENTS

Temperature. When working in a hot environment, humans maintain normal body temperature by perspiring and by increasing the blood flow to the surface of the body. The large amounts of water and salt lost in perspiration then need to be replaced. In the past, miners who perspired profusely and drank water to relieve their thirst experienced intense muscular pain—a condition known as miner's cramps—as a result of restoring their water but not their salt balance. When salt in the requisite amount was added to their drinks, workers no longer developed miner's cramps. Heat exhaustion is characterized by thirst, fatigue, giddiness, and often muscle cramps; fainting can also occur. Heatstroke, a more serious and sometimes lethal condition, results when prolonged exposure to heat and high humidity prevents efficient perspiration (by preventing evaporation of sweat), causing the body temperature to rise above 106° F (41° C) and the skin to feel hot and dry. If victims are not quickly cooled down, coma, convulsions, and death can follow. To prevent heat exhaustion or heatstroke, workers unaccustomed to high temperatures should allow adequate time (ranging from days to weeks) for their bodies to become acclimatized before performing strenuous physical tasks.

Work in cold environments may also have serious adverse effects. Tissue damage that does not involve freezing can cause inflammatory swelling known as chilblains. Frostbite, or the freezing of tissue, can lead to gangrene and the loss of fingers or toes. If exposure is prolonged and conditions (such as wet or tight clothing) encourage heat loss, hypothermia, a critical fall in body temperature, may result. When body temperature falls below 95° F (35° C), physiological processes are slowed, consciousness is impaired, and coma, cardiorespiratory failure, and death may ensue. Workers exposed to extreme cold require carefully designed protective clothing to minimize heat loss, even though a degree of acclimatization occurs with time.

Decompression sickness

Atmospheric pressure. Decompression sickness (caisson disease) can result from exposure to high or low atmospheric pressure. Under increased atmospheric pressure (such as that experienced by deep-sea divers or tunnel workers), fat-soluble nitrogen gas dissolves in the body fluids and tissues. During decompression the gas comes out of solution and, if decompression is rapid, forms bubbles in the tissues. These bubbles cause pains in the limbs (known as the bends), breathlessness, angina, headache, dizziness, collapse, coma, and in some cases death. Similarly, the gases in solution in the body tissues under normal atmospheric pressure form bubbles when pressure rapidly decreases, as when aviators in unpressurized aircraft ascend to high altitudes too quickly. Emergency treatment of decompression sickness consists of rapid recompression in a compression chamber with gradual subsequent decompression. The condition can be prevented by allowing sufficient decompression time for the excess nitrogen gas to be expelled naturally.

Noise. Exposure to excessive noise can be unpleasant and can impair working efficiency. Temporary or permanent hearing loss may also occur, depending on the loudness or intensity of the noise, its pitch or frequency, the length and pattern of exposure, and the vulnerability of the individual. Prolonged exposure to sound energy of intensity above 80 to 90 decibels is likely to result in noise-induced hearing loss, developing first for high frequencies and progressing downward. The condition can be prevented by enclosing noisy machinery and by providing effective ear protection. Routine audiometry gives an indication of the effectiveness of preventive measures.

Vibration. Whole-body vibration is experienced in surface and air transport, with motion sickness its most familiar effect. A more serious disorder, known as Raynaud's syndrome or vibration white finger (VWF), can result from the extensive use of vibratory hand tools, especially in cold weather. The condition is seen most frequently among workers who handle chain saws, grinders, pneumatic drills, hammers, and chisels. Forestry workers in cold climates are particularly at risk. Initial signs of VWF are tingling and numbness of the fingers, followed by intermittent blanching; redness and pain occur in the recovery stage. In a minority of cases the tissues, bones, and joints affected by the vibration may develop abnormalities; even gangrene may develop. VWF can be prevented by using properly designed tools, avoiding prolonged use of vibrating tools, and keeping the hands warm in cold weather.

Other mechanical stresses. Muscle cramps often afflict workers engaged in heavy manual labour as well as typists, pianists, and others who frequently use rapid, repetitive movements of the hand or forearm. Tenosynovitis, a condition in which the sheath enclosing a tendon to the wrist or to one of the fingers becomes inflamed, causing pain and temporary disability, can also result from prolonged repetitive movement. When the movement involves the rotation of the forearm, the extensor tendon attached to the point of the elbow becomes inflamed, a condition commonly known as tennis elbow.

Ionizing radiation. Ionizing radiation damages or destroys body tissues by breaking down the molecules in the tissues into positively or negatively charged particles called ions. Radiation that is capable of causing ionization may be electromagnetic (X rays and gamma rays) or particulate (radiation of electrons, protons, neutrons, alpha particles, and other subatomic particles) and has many uses in industry, medicine, and scientific research.

Ionizing radiation injury is in general dose-dependent. Whole-body exposure to doses in excess of 1,000 rads results in acute radiation syndrome and is usually fatal. Doses in excess of 3,000 rads produce cerebral edema (brain swelling) within a matter of minutes, and death within days. Lesser doses cause acute gastrointestinal symptoms, such as severe vomiting and diarrhea, followed by a week or so of apparent well-being before the development of the third toxic phase, which is characterized by fever, further gastrointestinal symptoms, ulceration of the mouth and throat, hemorrhages, and hair loss. There is an immediate drop in the white-cell elements of the blood, affecting the lymphocytes first and then the granulocytes and platelets, with a slower decline in the red cells. If death does not occur, these symptoms may last for many months before slow recovery begins.

Carcinogenic effects of exposure to radiation

Delayed effects of exposure to radiation include the development of leukemia and other cancers. Examples include the skin cancers that killed many of the pioneering scientists who worked with X rays and radioactive elements; the lung cancer common among miners of radioactive ores; and the bone cancer and aplastic anemia that women who painted clock dials with a luminous mixture containing radium and mesothorium developed as a result of ingesting small amounts of paint when they licked their paintbrushes to form a point.

Nonionizing radiation. Nonionizing forms of radiation include electromagnetic radiation in the radio frequency, infrared, visible light, and ultraviolet ranges. Exposure to radiation in the radio frequency range occurs in the telecommunications industry and in the use of microwaves. Microwaves produce localized heating of tissues that may be intense and dangerous. Various other disorders, mainly of a subjective nature, have been reported in workers exposed to this frequency range. Infrared radiation can be felt as heat and is commonly used in industry in drying or baking processes. Prolonged exposure to the radiation can result in severe damage to the skin and especially to the lens of the eye, where cataracts may be produced. Working under poor lighting conditions can adversely affect worker efficiency and well-being and may even cause temporary physical disorders, such as headache or dizziness. Proper lighting should provide adequate, uniform illumination and appropriate contrast and colour, without any flickering or glare. Exposure to ultraviolet radiation from the Sun or such industrial operations as welding or glassblowing causes erythema of the skin (a condition familiarly known as sunburn), skin cancer, and inflammation of the conjunctiva and cornea. Pigmentation offers natural protection against sunburn, and clothing and glass can also be used as effective shields against ultraviolet radiation. Lasers emit intense infrared, visible, or ultraviolet radiation of a single frequency that is used in surgery, for scientific research, and for cutting, welding, and drilling in industry. Exposure to these beams can burn the skin and cause severe damage to the eye.

DISORDERS DUE TO INFECTIOUS AGENTS

A large number of infectious diseases are transmitted to humans by animals. Many such diseases have been largely eliminated, but some still pose hazards. Anthrax, for example, can be acquired by workers handling the unsterilized hair, hide, and bone of infected animals; and slaughterhouse workers, farmers, veterinarians, and others in contact with infected animals, milk, and milk products still frequently contract brucellosis.

Contact with contaminated water is another common method of acquiring infectious diseases. Many workers are infected by organisms that thrive in the puddles or stagnant water found in sewers, canals, paddies, slaughterhouses, irrigation projects, and mines.

Laboratory workers, nurses, surgeons, and other health care workers may contract infectious diseases such as tuberculosis in the course of their work. To help prevent infection, these workers should wear appropriate protective clothing and exercise care when handling contaminated needles or other equipment. Contaminated material should be appropriately bagged, labeled, and disposed.

DISORDERS DUE TO PSYCHOLOGICAL FACTORS

Psychological factors are important determinants of worker health, well-being, and productivity. Studies have shown the benefits to workers who feel satisfied and stimulated by their jobs, who maintain good relationships with their employers or supervisors and with other employees, and who do not feel overworked. Such workers have lower rates of absenteeism and job turnover and higher rates of output than average.

Boredom and mental stress

The two psychological hazards commonly encountered at work are boredom and mental stress. Workers who perform simple, repetitious tasks for prolonged periods are subject to boredom, as are people who work in bland, colourless environments. Boredom can cause frustration, unhappiness, inattentiveness, and other detriments to mental well-being. More practically, boredom decreases worker output and increases the chances of error and accident. Providing refreshment and relaxation breaks or other outside stimulus can help relieve boredom.

Mental stress often results from overwork, although nonoccupational factors, such as personal relationships, life-style, and state of physical health, can play a major role. Job dissatisfaction, increased responsibility, disinterest, competition, feelings of inadequacy, and bad working relationships can also contribute to mental stress. Stress affects both mental and physical health, causing anger, irritation, fatigue, aches, nausea, ulcers, migraine, asthma, colitis, or even breakdown and coronary heart disease. Moderate exercise, meditation, relaxation, and therapy can help workers to cope with stress.

BIBLIOGRAPHY. GEORG AGRICOLA, *Georgius Agricola De Re Metallica,* translated by HERBERT CLARK HOOVER and LOU HENRY HOOVER (1912, reprinted 1950; originally published in Latin, 1556), an illustrated 16th-century study of working conditions in the mines of central Europe; BERNARDINO RAMAZZINI, *Diseases of Workers,* translated and revised by WILMER CAVE WRIGHT (1940, reprinted 1983; originally published in Latin, rev. ed., 1713), a classic; H.E. SIGERIST, "Historical Background to Industrial and Occupational Diseases," *Bulletin of the New York Academy of Medicine,* 12:597–609 (November 1936), a brief account of the recognition of occupational disease; DONALD HUNTER, *The Diseases of Occupations,* 6th ed. (1978), a historical overview of occupational medicine; R.S.F. SCHILLING (ed.), *Occupational Health Practice,* 2nd ed. (1981), a detailed account of industrial hygiene and disease prevention practices; and LUIGI PARMEGGIANI (ed.), *Encyclopaedia of Occupational Health and Safety,* 3rd rev. ed. (1983), a comprehensive reference source prepared under the auspices of the International Labour Organisation. Other comprehensive works include MARCUS M. KEY *et al.* (eds.), *Occupational Diseases: A Guide to Their Recognition,* rev. ed. (1977); LINDA ROSENSTOCK and MARK R. CULLEN, *Clinical Occupational Medicine* (1986); and JOHN C. BARTONE, *Occupational Diseases: International Survey with Medical Subject Directory and Bibliography* (1983).

Specific problems are studied in G. KAZANTZIS and L.J. LILLY, "Mutagenic and Carcinogenic Effects of Metals," in vol. 1 of LARS FRIBERG, GUNNAR F. NORDBERG, and VELIMIR B. VOUK (eds.), *Handbook on the Toxicology of Metals,* 2nd ed., 2 vol. (1986); W. KEITH C. MORGAN and ANTHONY SEATON, *Occupational Lung Diseases,* 2nd ed. (1984); DAVID F. GOLDSMITH, DEBORAH M. WINN, and CARL M. SHY (ed.), *Silica, Silicosis, and Cancer: Controversy in Occupational Medicine* (1986); MICHAEL ALDERSON, *Occupational Cancer* (1986); ARTHUR F. DISALVO (ed.), *Occupational Mycoses* (1983); RICHARD R. WEEDEN, *Poison in the Pot: The Legacy of Lead* (1984); A.J. BRAMMER and W. TAYLOR (eds.), *Vibration Effects on the Hand and Arm in Industry* (1982); and DIANA CHAPMAN WALSH and RICHARD H. EGDAHL (eds.), *Women, Work, and Health: Challenges to Corporate Policy* (1980). *Preventing Illness and Injury in the Workplace* (1985), is a survey on the prevention and control of occupational disease in the United States prepared by the Office of Technology Assessment.

(G.Ka.)

Oceanic Arts

The term Oceania has been variously used to refer to the land areas of the central and southern Pacific. For the purposes of this article, it is understood to include the island continent of Australia and the island groups of Melanesia, Polynesia, and Micronesia. The largest island in Melanesia, and the second largest island in the world, is New Guinea. New Guinea lies north of Australia and is on the western edge of Oceania. It is divided into the independent country of Papua New Guinea on the east and the Indonesian province of Irian Jaya on the west. Northeast and east of New Guinea lies the arc of islands that constitutes the rest of Melanesia. Included in this group are the Admiralty Islands, New Britain, and New Ireland (all in the Bismarck Archipelago and administratively part of Papua New Guinea), the Solomon Islands, Vanuatu (formerly the New Hebrides), New Caledonia, and Fiji. The small islands of Micronesia lie north of Melanesia and form a parallel arc that extends from Palau (Belau) and the Mariana Islands in the northwest through the Caroline and the Marshall islands to Kiribati (formerly the Gilbert Islands) in the southeast. East of both Melanesia and Micronesia is Polynesia, a vast triangular area formed by New Zealand in the southwest, Easter Island in the southeast, and Hawaii in the north. Within this triangle lie, roughly from west to east, Fiji (which serves as a transitional territory between Melanesia and Polynesia), Tuvalu (formerly the Ellice Islands), Tonga, Samoa, the Cook Islands, and French Polynesia (including the Society, the Austral [Tubuai], and the Marquesas [Marquises] islands).

The variety of Oceania's cultures has always been a matter for wonder. A small number of people, no more than three to four million at the time of first contact with Westerners, practiced a wide range of economies based largely on the ecological niches in which they lived. In the broad expanses of Australia, the tribal communities lived largely as hunting-and-gathering nomads. In New Guinea they ranged from nomadism to large-scale horticulture, with some coastal groups dependent mainly on fishing. The other Melanesian islanders were predominantly horticulturalists. With certain exceptions, Melanesian societies were acephalous, with men of energy and ambition competing for status that lasted only for their lifetimes. Polynesian and Micronesian societies, on the other hand, were hierarchical; in the most formalized, dynastic rulers, of reputedly divine descent, dominated several lower classes, the most inferior being prisoners and slaves.

Religion was everywhere a paramount concern, and throughout Oceania the supernatural and the activities associated with it were the inspiration for the greater achievements of the arts. These were not limited to the dances, chants, figures, masks, and architectural sculptures connected with actual ritual; so pervasive was belief that it infused with symbolism the creation of what might seem mundane and secular objects. In Australia and Melanesia, male domination of women largely excluded the latter from religious cults and thus from many types of artistic endeavour. In fact, in much of Melanesia, religion and art centred on segregated male meetinghouses, dance grounds, secret societies, and initiation ceremonies. In Polynesia and eastern Micronesia, however, religious ceremonies were often held at open-air sacred enclosures. Thus, although architecture was well-developed, only a few areas had buildings as artistically significant as the men's houses in Melanesia. But, whatever the type, almost every example of art produced in Oceania, whether for ritual or functional use, reflects its creator's sense of the spiritual world.

This article discusses the visual art, music, dance, and literature of the traditional peoples of Oceania. Although it includes the effects of Western colonization and the adaptation of traditional forms to modern technology, it does not treat the postcolonial adoption of wholly Western styles and forms. The literatures of Australia and New Zealand, for example, are derived from and strongly influenced by the English literary tradition; they are therefore discussed separately in the article AUSTRALIA AND NEW ZEALAND, LITERATURES OF. Information on the geographic, economic, and historical background of Oceanic arts can be found in the article PACIFIC ISLANDS.

For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, sections 613, 621–622, and 624–629, and the *Index*.

The article is divided into the following sections:

Visual arts 104
  General characteristics 104
  Early styles 106
  Effects of European contact 109
  Australian visual art 109
  Melanesian visual art 110
  Polynesian visual art 121
  Micronesian visual art 126
Music and dance 126
  The role of music and dance 126
  Musical instruments 128
  Regional styles and traditions 128
Literature 130
  The role of the author 130
  Oral traditions 131
  Modern literature 132
Bibliography 133

Visual arts

GENERAL CHARACTERISTICS

Materials and techniques. Until the 16th and 17th centuries, when European cultures appeared upon the scene, Oceanian cultures maintained various forms of Neolithic technology. The only exception was in the northwest of New Guinea, where the people living around Geelvink Bay (Teluk Cenderawasih) imported very small quantities of metal from the Indonesians of the Moluccas (Maluku). The technique of forging was jealously guarded, virtually as a cult secret; some tools were traded but only in quantities far too small to have made much impact on normal working conditions.

Throughout the rest of Melanesia and in Polynesia and Micronesia, the basic tool remained the stone blade, which was hafted as an adz or an ax and sometimes interchangeably as both. Tridacna shell was sometimes used for blades in parts of Oceania where stone was in short supply, including Micronesia and the Solomon Islands. When obsidian was available, it was chipped into blades for use as both weapons and tools. Other working materials included bamboo and bivalve shells, which take extremely sharp edges. Some fine cutting and engraving was done with unhafted boar tusks or with hafted shark and rodent teeth. Animal bones served as gouges, awls, and needles. All these tools were employed in working wood, which with rare exceptions was the main medium used throughout Oceania.

Tools

Clay was also employed, mainly for sculptures, for some small musical instruments (whistles), and for pottery in Melanesia and New Guinea. The making of clay vessels

was almost exclusively women's work, except in a few small areas in New Guinea and the northern Solomons. The usual method involved spiral coiling of rolls of clay. The decorating of the pot was the work of men.

Some working of shell and turtle shell was done with simple drilling and abrading equipment. The carving of stone, although obviously presenting far more arduous and time-consuming problems than wood, was undertaken remarkably often and occurred throughout the Pacific Islands; hammering, pecking, and polishing were the main methods. Even so resistant a material as jade was mastered by grinding with abrasives.

Paint and painting animate sculpture—often literally, in religiosymbolic terms, as paint was considered to have magical, vivifying powers. Paints were generally ochres, with some vegetable-derived pigments. Water was the usual medium, occasionally supplemented with sap. Brushes were the fibrous ends of chewed or frayed sticks, small feather bundles, pieces of wood, and sometimes the most elementary applicator of all, the finger. Apart from sculpture, the surfaces used for painting were rock faces, bark, and tapa (cloth made from pounded bark). Rock painting was most common in Australia, where panels of bark were also used. In Melanesia, paintings were made mainly on sago-palm spathes and sheets of tapa cloth. In Polynesia the women manufactured great quantities of tapa, which they then decorated with abstract designs using vegetable dyes. The techniques they employed included painting, stenciling with leaf templates, rubbing over relief-design tables, stamping, and printing with carved bamboo rollers.

The only areas where weaving was practiced were the Caroline Islands, the Polynesian outliers east of the Solomon Islands, some of the Santa Cruz Islands, parts of Vanuatu, the Saint Matthias Group (northwest of New Ireland), and a few places on the northern coast of Irian Jaya. Spinning was unknown; instead of yarn or thread, strips of banana fibre were used on a simple backstrap loom. Weaving was a woman's craft in the Caroline and Saint Matthias islands but was practiced by men elsewhere. A form of "finger weaving," as in net making, was used by Maori women in creating textiles from flax fibres.

Museum fur Volkerkunde, Basel, Switz. (Vb 28418-71); photograph, P. Horner



Figure 1: Initiation scene in a cult house. From Abelam, Papua New Guinea. In the Museum of Ethnology, Basel, Switz. Height 390 cm.

Architecture

The architecture of the Pacific Islands was varied and sometimes large in scale. Buildings reflected the structure and preoccupations of the societies that constructed them, with considerable symbolic detail. Technically, most buildings in Oceania were no more than simple assemblages of poles held together with cane lashings; only in the Caroline Islands were complex methods of joining and pegging known.

Oceanic artists' quest for media was consummately opportunistic; they regarded almost anything from the lavish natural world that surrounded them as potentially usable. The marine world yielded shells of all kinds, especially conus, cowrie, and nassa shells. Birds gave down, beaks, and plumes (those of the birds of paradise were especially prized); animals provided teeth, tusks, and skins; insects supplied their brilliant wing cases. The vegetable realm was drawn upon for flowers, leaves, and fibres. The assembly of such materials into single objects was rare in Polynesia and Micronesia, but the practice was typical of Australian and Melanesian styles and contributed brilliantly to their more spectacular effects. The most basic medium of all was the human body, which received both removable and permanent decorations, including scarification, enhanced by treatment to raise keloid welts, in New Guinea and tattooing with needles and pigments elsewhere.

Artist and society. In societies whose members are largely self-reliant, some degree of craft skill is practically universal. Men make their own canoes, build their own houses, and carve simple household equipment such as hooks and stools; individuals are responsible for decorating their own belongings, including their bodies. In the case of body decoration, however, which can be culturally prescribed in form, highly skilled in execution, and dense in symbolism, the more lavish displays usually entail more than the wearer's sole efforts. Tattooing and scarification, usually tokens of ritual or hierarchical status, were the work of esteemed specialists.

To progress beyond simple skills, a craftsman not only required the will to excel but sometimes was subject, in theory at least, to socially defined restrictions. There seems to have been an inclination to regard artistic talent as passing from father to son, or from mother to daughter when appropriate; but, in cases in which this was true, society's concept of the role of the artist probably played a bigger role than heredity.

Artistic training

In many societies the artist was—and still is today—expected to begin his career as an apprentice to a known master, often working on preparatory tasks or the less-demanding details of a project. In some parts of Melanesia, among the Kilenge of New Britain, for example, or in the Solomons, artistic progress is recognized as covering several stages. The apprentice grows into an independent worker with limited skills and eventually, if he has talent and ambition, becomes a master in his turn. In the Solomons the aspirant is actually expected to produce test pieces for approval by his peers and mentors. Elsewhere the process is apparently less formal and, particularly for grandiose projects, less individualistic. Large-scale projects are often an affair of communal effort under specialized supervision. In Papua New Guinea several men at a time may work on a single large architectural carving among the Kwoma, and a whole team may paint one of the huge gables of the Abelam. Individuals, however, may carve major sacred objects when they are inspired by dreams or induced visions. These interventions by the supernatural world can be quite common: if work goes badly, the failure is attributed less to the workers' incompetence than to the displeasure of the spirits concerned.

In Polynesia, with its more sharply graded societies, the role of artist was more closely related to the religious expert (for instance, the Maori tohunga) than it was in Melanesia. Indeed, in Hawaii and elsewhere carvers formed a special priestly class, and their work was accompanied at every stage with rituals and prayers. The New Zealand Maori considered carving a sacred activity, surrounded by spiritual and physical dangers. Myths of the origins of carving connected it directly to the gods, and its subjects linked it intimately to the ancestors. Carving was one of eight proverbial attainments of a chief, and young Maori of

high rank were trained in the formal schools of learning. There were cases of chiefs being captured and enslaved for their talents and, conversely, of slaves celebrated as artists.

Remuneration of artists

The material rewards were not great. While the carver and painter was preoccupied with his work, it was the business of his employer to keep him well fed. On completion, the artist received agreed amounts of valuables, but he might well give away some of them (among the Kilenge at least) to those who praised him. Praise and esteem were in fact the main rewards and were steps toward the making of a "Big Man" of power and influence in Melanesian communities; in Polynesia mana—personal prestige and moral authority—was achieved in the same way. Equal or even greater credit often went to the man who commissioned the work, for he was regarded as its true author. His achievement in seeing that the work was first instigated and then carried through to a successful conclusion earned him fame and prestige.

Aesthetics. Pacific languages seem to be deficient in terms to express appreciation of or reactions to art, apart from a few that designate the mastery of individual specialists. Little is understood, moreover, about the islanders' aesthetic concepts. Reactions to works of art seem to range from the pragmatic and rational in the secular realm to the violently emotional in the religious. At a fairly simple level, aesthetic appreciation is expressed as approval of the manner in which a work has been accomplished, of its compliance with possibly unformulated but nevertheless well-understood standards. Craftsmanship and suitability to function are highly valued.

In general, innovation does not seem to have been highly prized. Nevertheless, changes have certainly taken place in the arts over the long period of Pacific history, even though, in the absence of more than a scattering of archaeological examples, such changes are difficult to document. One technique used by artists to attain success was to copy models of recognized excellence and symbolic soundness; old works were often retained precisely for this purpose. The inevitable introduction of variations in these situations, as a result of variations in individual talent, was largely ignored, and the intention of identity between old and new objects was accepted as always having been achieved. The ideal of the local tradition was thus maintained, even though actual stylistic fluctuations must have occurred over time.

In some areas the exotic was deeply admired and therefore copied: in parts of New Guinea, for instance, certain items captured in warfare are known to have been duplicated. Such cases were probably comparatively rare, however. More often works displaying special craft techniques (such as work in ivory imported by Fijians from Tonga) were treasured because it was accepted by the importers that the imports were beyond their skills to manufacture for themselves.

Role of religion in art

The Maori of New Zealand developed the most precise aesthetic terminology of Oceania, describing both the innate properties of a work and its effect on the viewer. A masterpiece possesses *ihi* (power), emanates *wana* (authority), and inspires *wehi* (awe and fear). The belief that art and religion overlap is widespread in the Pacific, and religious objects are often works of visual art (though not invariably). These objects are not considered sacred in themselves, however; they are humanly worked things into which supernatural beings can be induced for human purposes. These supernaturals are always powerful, unpredictable, and thus dangerous. In New Guinea their destructive power may turn against the object itself, causing a carving to rot, self-consumed; or an object may become so loaded with accumulated power that it has to be buried or otherwise eliminated. It is possible that the practice of abandoning elaborate and painstakingly made carvings after ritual use—as in New Ireland and among the Asmat of Irian Jaya—was inspired by such beliefs. In many societies an uninitiated person who glimpsed the sacred objects would be executed, but it is likely that the offended spirits were considered the killers, not the men who acted for them and performed the execution. Among the Maori, ancestral heirlooms were not to be touched without ritual purification, and mistakes in ritual, especially in the building of meetinghouses, with their powerful ancestral associations, could be fatal. Awe and fear are understandable emotions in such circumstances.

In areas where religion depends more on ritual dances or oratory than on objects, expression of the visual arts may be channeled (as in Samoa and much of Micronesia) into an exquisite refinement of craftsmanship, often in the making of utilitarian objects. In these circumstances, the quality of an object often becomes a symbolic reference to social status.

Oceanic visual art, then, although rarely baldly pictorial in a Western manner, is replete with references to both religious and social values. It may even, it has been suggested, be a material means by which values are transmitted nonverbally to those qualified to understand the messages involved, thus becoming a mode of communication that reinforces and is vital to society.

EARLY STYLES

The history of Oceanic art falls into two major phases, corresponding to the periods before and after Western contact. This is due not so much to the changes ensuing from contact—decisive as they have been—as to the preservation of otherwise ephemeral material by Western collectors and researchers. The total loss of early works and the paucity of archaeological discoveries renders the comprehension of ancient Oceanic art fitful and incomplete. In fact, there is not enough known about early Micronesian art to warrant discussion here. Nevertheless, what has survived elsewhere hints at the antiquity of art traditions in Oceania and sometimes illuminates the origins of more recent styles.

Australia. The Australian continent is liberally dotted with thousands of rock-art sites. They include rock shelters, outcrops of rock, and surface sheets of rock and are decorated with painted, pecked, or engraved figurative and nonfigurative forms in a wealth of styles. These are the main testimonials to the prehistoric art of the Aborigines; the only portable works from early periods that have been discovered are some elaborate items used for personal decoration. Long necklaces and chaplets made of animal teeth and lizard vertebrae, bone beads, and stone pendants have been found in burials and elsewhere dating from 15,000 BP (before the present) and later. Long bone pins indicate the existence of garments, probably cloaks made of animal skins.

The early use of colour for various purposes is attested by the inclusion of red ochre in burials at Lake Mungo in New South Wales, dated 32,000 BP. While this is not necessarily evidence of any specifically artistic activity, it shows the ritual value of the colour and of the material, which was imported from sources many miles away. Paintings for which human blood was the medium have been found and proved to be more than 20,000 years old.

The chronology of the rock-art styles is established largely

Holle Bildarchiv, Baden-Baden, Ger



Figure 2: Rock painting of a lizardlike animal. From Hawker, South Australia.

Rock-art styles

by the classic method of tracing the superposition of works in one style over works in another; but current theories are also based on such factors as known climatic and geologic events, the presence or absence in the paintings of certain animals or equipment that are now extinct or obsolete, and the degree to which modern Aborigines are familiar with the sites and the meanings of the art. One factor that decisively marks the end of the early period is the representation of European or (in the north of the continent) Indonesian cultural elements, such as ships and introduced animals.

One of the earliest known styles is the Panaramittee. It was widespread, mainly through southern Australia, central Australia, and Tasmania, and dates from about 30,000 BP onward. It is characterized by small pecked designs, both figurative and nonfigurative, on rock surfaces. The nonfigurative designs include circles, crescents, and radiating lines; the figurative are almost all of footprints and bird and animal tracks.

Another early style, dated to 20,000 BC, is represented in Koonalda Cave under the Nullarbor Plain in South Australia. Certain areas of the cave walls, which are composed of a soft rock, are densely covered with engraved or finger-marked geometric designs. Most of the designs consist of no more than parallel lines or herringbone patterns, but they cover several thousand square feet. It is possible that their significance lies as much in their placement at specific points in the cave as in their now undiscoverable symbolism.

Both rock engravings and paintings in the Simple Figurative style are widely found at sites in the north, east, and west of Australia but rarely in the interior. The style apparently followed the Panaramittee, but it cannot be dated precisely. It is characterized by somewhat loose silhouettes of human and animal forms and has remained influential until recent times.

In northwestern Australia, in both coastal and hinterland areas, there are at least two sequences of painting styles. In Arnhem Land, rock painting has been divided into a sequence of four styles, partly on the basis of apparent references to environmental changes. The earliest, the Mimi (a clan of spirit beings) or Dynamic style, is notable for linear human stick figures that wear ornaments, carry spears and boomerangs, and are occasionally endowed with animal heads. They are associated with paintings of now-extinct animals, such as the Tasmanian wolf (thylacine). The style is presumed to date from 18,000 BP to pre-9000 BP. It is followed by the Estuarine style, which developed during a period when saltwater conditions prevailed: a situtation reflected in the use of crocodiles as subjects in paintings in the X-ray style (in which the internal organs are shown). A subsequent Freshwater phase is characterized by representations of ceremonial fans made from feathers of marsh birds. Finally, there are paintings from a "contact" period, which began with the arrival of Indonesian fishers of trepang (sea cucumber) at the end of the 18th century and continued, after 1880, with the arrival of Australian drovers on horseback. The visits of both are pictured in the rock art.

A parallel sequence has been traced in paintings from the Kimberly region, to the west. An early period is manifested by the Bradshaw style of small human figures, mostly in red, perhaps dating from before 3000 BC. The Bradshaw style is succeeded by the Wandjina style, which takes its name from the ancestor spirits depicted in the paintings. The large white spirit figures are outlined in black and have mouthless, circular faces that are framed in red, rayed halos. This style has persisted to the present.

Melanesia. The first indication of the existence of any form of art in Melanesia is shown by the use of pigments, probably for personal decoration, in the eastern Highlands of Papua New Guinea by 15,000 BC. Other examples of early art in New Guinea include the stone carvings, including pestles, animal and human figures, and mortars, that have been found in the central Highlands, where most seem to have been made. Some were exported to eastern Papua New Guinea. The carvings are as yet undated, although it is known that plain stone bowls were in use about 2000 BC. Much of the sculptures' imagery



Figure 3: Stone figure, igneous rock. From the Ambum River valley, Papua New Guinea. In the Australian National Gallery, Canberra. Height 20 cm.
Collection: Australian National Gallery, Canberra

has been repeated in the recent art styles of the Sepik area and elsewhere. Rock art, in the form of paintings and petroglyphs, is abundant in Papua New Guinea but also remains undated.

Lapita art

The most important evidence of art in the early western Pacific is the ceramic style called Lapita, after a site in New Caledonia. It is the most prominent material aspect of a culture that flourished from approximately 1900 BC to the beginning of the modern era and that achieved an astonishingly wide distribution. Lapita sites, or other evidences of Lapita influence, are found from the northern coast of Papua New Guinea throughout the major island groups of Melanesia to Fiji, Tonga, and Samoa in the east. The Lapita culture complex involved intensive exchange of ceramics, stone tools, and other goods over long distances. Obsidian for tools, in particular, was traded from the Admiralty Islands and New Britain as far as New Caledonia.

The Lapita ceramics include a range of handsome vessel forms: flat-bottomed dishes, shallow and deep bowls, and small-mouthed, carinated vessels. It is their decoration, applied with toothed stamps, that makes them so distinctive. Much of it is applied in stacked horizontal zones; most of the design units are constructed from simple arcs or right angles, but some are intricate interlocking patterns. There are also some complex curvilinear designs incorporating faces that are the earliest dated human representations in the Pacific Islands. The early western ceramics are the most elaborate; designs become increasingly simple to the east until, after about 500 BC, vessels made in the Fiji-Tonga-Samoa area retain some Lapita shapes but lack decoration. Little else remains of this elegant artistic tradition. Wood carving was practiced, judging by the remains of suitable tools, but no examples have survived.

By courtesy of R.C. Green



Figure 4: Lapita pottery, reconstructed two-dimensional anthropomorphic design (*c.* 1000 BC).

It is possible that Lapita art was fundamental to the later development of art in the Pacific. Some Lapita design motifs, especially the more complex, can be shown to have survived in Melanesia until the present day, as have Lapita principles of design layout. Lapita art was also in all probability ancestral to early phases of much Polynesian art and even to the tatooing and tapa decoration styles of recent times.

No other surviving early art from Melanesia approaches the accomplishment of Lapita, but a few contemporary and subsequent prehistoric ceramic traditions deserve mention. Some, as at Sohano on Bougainville Island in the northern Solomons (*c.* 500 BC) and Yule Island off the southern coast of Papua New Guinea (*c.* 1000–2000 BP), have incised designs that may derive from Lapita. A more elaborate and impressive style is that of the Mangaasi culture of Vanuatu, which dates from 700 BC to AD 1200. Early Mangaasi ceramics include spherical pots and are decorated with bold triangles outlined with applied fillets, within which are further arrangements of incised triangles. Handles were modeled in bird and animal forms.

The apparently early New Guinea tradition of stone carving has parallels in other parts of Melanesia. Stone bowls have been found in New Britain, and stone flared ax heads with flanges shaped like birds' and animals' heads have been found in the northern Solomons.

Polynesia. In the prehistoric cultures of Polynesia, two conspicuous themes figure largely: the ceremonial ground (the marae/ahu complex, known by varying local terms) and personal ornaments. The ceremonial ground was a place of worship. It usually took the form of an enclosure (marae), which was raised or walled or in some other way delineated, with a raised platform (ahu) across one end. A row of upright stone slabs along the ahu were backrests for the gods, while other stones indicated the places of human officiants. The grounds went through various phases of development in the island groups and were the Polynesians' most conspicuous architectural achievements.

The marae/ahu complex

Early Polynesian cultures shared a number of traits deriving from a common tradition. Types of adzes, fishhooks, and certain ornaments recur, including reel-shaped necklace units and pendants of whale teeth, unshaped or shaped by carving a sliver from the lower end. Shaped whale-tooth pendants are found in the earliest phase of Marquesan culture (AD 300–600), as are small perforated shell disks that might have been attached to the coronets typical of later periods. A few simple stone figures belong to a "developmental" phase (AD 600–1300); one closely resembles small stone figures from Necker Island, the most northerly of the Hawaiian group. These are posed frontally, have circular faces with clumsily delineated features, and may date from about the 10th century. They would seem to be representative of an ancestral Polynesian carving style and are the earliest sculpture from Hawaii. Monumental stone figures of gods, in a style that persisted into the 19th century, were being carved and installed on marae in the Marquesas about 1500.

Ernest Manewal—Shostal/EB Inc.



Figure 5: Sculptures cut from volcanic rock, Easter Island.

Easter Island, remote and isolated, is the site of the most famous monuments of the Pacific. Among the monuments are some 300 stone platforms, some of which were used for burials and some of which supported the island's spectacular colossi. Work on the statues, which were carved from a soft volcanic stone, seems to have begun about AD 900. The first figures were relatively small, about 2 metres high; later statues were as much as 12 metres high. The statues' heads and torsos are in an extremely rigid frontal style, with the slender arms and elongated hands carved down the sides and across the belly. Necks are barely indicated; the faces have deep-set eyes, long pointed noses, and massive chins. The statues originally had barrel-shaped topknots of red stone and eyes of white shell and black stone. The Easter Island tradition of statue carving came to an end by about 1600, probably as a result of a serious breakdown of the culture caused by internecine wars.

The earliest New Zealand Maori culture had strong relationships to the contemporary art of eastern Polynesia, whence the Maori migrated about the 9th century. The use of tapa cloth was presumably common, and tattooing was practiced. Fishing lures (some carved as fish), fishhooks, and adzes follow Polynesian types, and the *patu* type of club in whalebone existed in both areas. In this early phase, the whale-tooth pendants and reel-shaped ornaments of Polynesia became in New Zealand massive stone versions, which were used as pendants or strung as necklaces. Other stone pendants were divided spheres and plaques with stylized fish or zoomorphs carved in relief. Wood carving has not survived, although suitable stone chisels have been found.

Early Maori art

The following phase shows the inception of specifically Maori styles. One indication is an increasing complexity exemplified by the elaboration of whale-tooth pendants. The original simple forms of central Polynesia became, by the 14th century, the so-called chevron pendants, which were probably worn in symmetrical pairs. They retain the tooth form but are flat and bordered with series of chevrons representing human limbs. A few small wood carvings from this period exist, as well as one major piece, the decoration for the roof of a house from Kaitaia. Although the roof decoration shows some Polynesian influence, it also powerfully states a major theme of Maori art: a human figure flanked by figures in profile, prototypes of the later *manaia* monsters. It is identical in composition to the lintel panels of later Maori art. Among other surviving carvings are a remarkable 16th-century stern piece and a canoe prow cover, both from the North Island; the bow cover is the oldest known work to be decorated with pecked spirals—the most dominant feature of later Maori art.

A series of combs found in a sacred deposit at Kauri Point Swamp on New Zealand's North Island illuminates the development of forms in the 16th to 18th centuries; the combs progress from square panels with engraved geometric designs to rounded forms with near-figurative decoration. Some of the later engraved features have spurs projecting from edges of parallel lines and are highly reminiscent of the carving on a canoe prow and stern post from Doubtless Bay and a relief panel from Awanui, both sites in the far north of North Island. In general, all these objects show a move away from the simple forms and plain surfaces of the earliest Maori art to more complex forms that are variegated with small areas of intensive bas-relief. This trend reached a culmination in a series of chests, for the bones of high-ranking people, carved in human form.

Following this, a highly vigorous revolution of Maori art took place. Cloaks, the primary garments, were still given geometric patterns on their borders, but otherwise there

was a new emphasis on flowing, curvilinear designs and a wealth of surface decoration. Pendants of whale teeth persisted, but only with minimal carving of a human face at the tip; and jade, from the mountains and streambeds of the South Island, became the most prestigious material for blades, weapons, and a wide range of ornaments.

EFFECTS OF EUROPEAN CONTACT

Interests of early Western explorers

The earliest Westerners to explore the Pacific Ocean undertook their dangerous voyages with firm determination to discover wealth in one form or another. Except in rare cases, they had no lively curiosity about the islanders. Neither their interests nor their skills were engaged in recording island life; indeed, their visits were often too brief to leave time for observations. Tense and even hostile encounters were frequent and must have further distorted the mutual views of those involved. The few comments and drawings to be found in the records of the expeditions of Álvaro Mendaña de Neira (1567–68, 1595), Pedro Fernández de Quirós (1605–06), Luis Váez de Torres (1606), and Don Diego de Prado (1606), all in the service of Spain, Jakob Le Maire (1616) and Abel Janszoon Tasman (1642–43) of the Netherlands, and others are thus all the more telling. They reveal that in general the material culture they witnessed was approximately the same as was recorded and collected by the late 18th-century explorers and their successors.

Thus it seems that, in many parts of the Pacific, material culture had reached a settled state by the 16th century and remained conservative for two more centuries. Then, in the 18th century, contact with European collectors began to stimulate changes in some areas of Oceania, including the outright faking of antiquities and the adaptation of traditional forms to European fancy. The islanders also began the manufacture for sale of favourite collectibles, such as stone adzes and weapons. Such adaptation of production has continued and remains in practice today.

AUSTRALIAN VISUAL ART

The art styles of the Australian Aborigines fall into three groups, which follow to some extent the ecological zones of the continent. The first group is identified with the heart of Australia; this region, which covers most of the continent's landmass, is arid desert surrounded by a belt of savanna. The second zone extends from the central desert region to the southeastern coast and includes sections of open eucalyptus forest and belts of tropical jungle. The third zone is similar to the southeastern zone, but it extends to the northeastern and northern coasts (including Arnhem Land and Cape York). It is thought that, at the time of European contact, the Aboriginal population (about 300,000 people) roughly corresponded with these divisions, the north-northeast region having the greatest numbers and the desert the least.

The material culture of all three groups was limited in types of objects but versatile and highly efficient in its adaptation to the peoples' hunting-and-gathering economy. All material objects were necessarily portable and often served more than one purpose. For example, wooden bowls were used as both food carriers and cradles; and boomerangs, which were used primarily for fighting and hunting, could also be used, in conjunction with shields, to make fires. The most consistently decorated objects were shields, spears, spear-throwers, clubs, and boomerangs of various forms.

The central desert. The art of the central desert area features arrangements of primarily curvilinear and rectilinear designs engraved on flat surfaces. Stylized depictions of birds, snakes, and human figures occur infrequently. Paint was used sparingly and was usually restricted to red and white.

Curvilinear designs mostly appeared in the eastern and central parts of the area in the form of concentric circles, arcs, semi-ovals, and wavelike patterns. Probably the most striking examples are found on the engraved churingas of the Aranda tribe. These oval or disk-shaped sacred boards were made of wood or stone and painted with red ochre. Each design element refers to a specific object or situation; but the application of the reference in its context of the general design, and its relation to myth, is known only to its clan proprietors. It is this relationship that is revealed, in whole or in part, at initiation rituals.

Ritual structures

The most elaborate design creations in the central desert area were the settings devised for totemic rituals. The ground was painted with large designs featuring the characteristic circles and serpents, in red or black on an ochre field with white dots. Arrangements of decorated poles and symbolic structures completed the settings. The participants had their bodies painted and then were covered with bird down adhered with the wearer's blood. In the northern central desert, headpieces, worn transversely or vertically, were constructed of spears sheathed in red and white bird down and represented totemic fauna and flora. In the south and the west, the totemic emblems were smaller panels of string and down worked on stick frames.

Objects made for daily rather than ritual use, such as spear-throwers and boomerangs, were typically engraved. The engraved designs are characteristically curvilinear in the central area, but the engraving of the west and northwest tends to consist of angular grooved key or diamond patterns against a background of parallel grooves (which were sometimes painted alternately red and white). Similar key patterns were engraved onto mother-of-pearl shells by tribes living along the coastal waters of the northwest. Highly prized as ornaments, the shells were traded far into the interior.

The southeast. Living in the immediate path of European colonization, the Aborigines of the southeast were the first to suffer from its effects; their culture was extinguished with some rapidity, and the area was practically depopulated. Their culture had been relatively rich. The temperate climate and the natural resources of the great Murray and Darling river systems stimulated a number of regional variations in art and material culture. In response to the cool winters, the Aborigines built fairly substantial wood shelters, covered with bark sheets and animal skins. They made large cloaks by sewing together opossum pelts incised with decorative patterns on the inner sides. For transport and fishing on the rivers, they built simple bark canoes.

Throughout the area, the basic designs were geometric. Objects were frequently engraved with dense patterns of solid or dotted zigzags and parallel lines. The rich texture thus created served as a background for other carved geometric designs, such as squares or diamonds, as well as for painted elements. Local variations in style are best seen in shields, of which there were four main types. From roughly north to south, the first type was an elongated oval with a convex surface. The second, used for parrying, was extremely narrow and was triangular in section. In the Murray River area, shields were thin, flat, broad ovals with a projecting tab at each end. The fourth type of shield, found east of the Murray River, was a narrow elongated oval pointed at both ends. Other weapons included long spear-throwers and a remarkable range of club types, with spatulate, hooked, or knobbed heads. In areas where the second and third types of shields were made, decorated boomerangs were also used for fighting, but they were engraved with uncharacteristic designs.

The ritual art of the northern area included abstract and representational designs channeled into the ground and large-scale earth effigies. Bark effigies and paintings on bark are recorded but have not survived. In the northwest, a unique form of monument was created: the dendroglyph, an engraving on a living tree trunk. Carved in the usual geometric style, dendroglyphs featured clan designs or made references to local myths. They were used to mark the graves of notable men or to indicate the perimeters of ceremonial grounds.

The dendroglyph

The north. From the rain-forest country of northeastern Queensland comes an unusual type of shield, a large flat oval with somewhat asymmetrically curved sides. Most have a raised central boss. Designs above and below the boss radiate away from it and are outlined in black and infilled with red, white, and yellow. As usual, they refer to mythological beings and episodes. Paddles and cross-shaped boomerangs were painted in the same manner for ceremonial use.

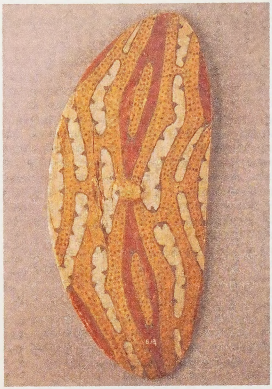


Figure 6: Shield, ochres on softwood. From the Cardwell River, northeastern Queensland. In the Australian museum, Sydney. Height 91 cm.
Australian Museum, Sydney (E13430); photograph, Ric Bolzan

The lavish use of colour on these objects is indicative of the emphasis placed on painting among the areas to the north, especially around the Gulf of Carpentaria and on its islands, in Cape York and Arnhem Land, and on Melville and Bathurst islands and Groote Eylandt. In Arnhem Land, paintings on bark sheets included both figurative images and the geometric designs typically used in sacred contexts. Paintings from western Arnhem Land and some adjacent islands were often in the X-ray style, in which animals are painted on dark monochrome backgrounds with their internal organs showing. In paintings of northeastern Arnhem Land the field was completely filled with both representational and geometric images depicted in fine-line cross-hatching. These images referred to ancestral myths and are programmatic, even narrative, in content.

By courtesy of the Metropolitan Museum of Art, New York City, The Michael C. Rockefeller Memorial Collection, bequest of Nelson A. Rockefeller, 1979



Figure 7: X-ray-style bark painting of kangaroos, earth pigments on eucalyptus bark. From Arnhem Land, Australia. In the Metropolitan Museum of Art, New York City. Height 1.03 m.

Unlike the rest of Australia, the northern zone is rich in three-dimensional wood sculpture. The Tiwi people of Melville and Bathurst islands created tall poles in abstract forms by carving, removing, or leaving in their original dimensions alternate sections of a tree trunk. Each pole was then painted in flat areas of colour interspersed with bands of cross-hatching. Such poles were planted in clusters as grave markers in elaborate funerary ceremonies, and boldly painted bark containers for offerings were placed on the poles. Throughout the northern region, small carvings of birds, animals, and plants were typical sacred emblems; but in northeastern Arnhem Land, as nowhere else in Australia, large figures of human beings also were used in ritual and sometimes as grave markers. This use of human figures has been attributed to the influence of Indonesian fishermen who visited the area for shell and sea cucumber, but it is also possible that it resulted from contact with the Torres Strait islanders to the north.

MELANESIAN VISUAL ART

The Torres Strait. The small islands of the Torres Strait, between northern Australia and southern Papua New Guinea, were inhabited by groups of people who, generally speaking, shared a common basic culture. Religious life revolved largely around male initiation cults of various creative and peregrinatory heroes, fertility cults, and funerary ceremonies. Stage settings and decorated masks were used for all of these rites. The settings were typically screens before which dancers appeared in reenactments of myths.

On the southern islands the chief material of major works of art was tortoiseshell, which was perhaps nowhere else in the world used on a comparable scale for masks and effigies. The tradition was evidently an old one, having been observed by the Spanish explorers Torres and Prado in 1606. The masks and effigies were built up of small plates of shell lashed together. Masks were painted red, with white details; some sparse decorative engraved detail was filled in with white, and carved wooden accessories, seed rattles, and feathers were added. The masks are of three types. Two, used for the hero cults, were to be worn horizontally on the top of the head and represent fish or combinations of creatures, such as the head of a crocodile or hawk with a fish's tail. Sometimes a subsidiary human face was added on top of the head. The masks for funerary ceremonies were more naturalistic with somewhat elongated faces and slightly elongated earlobes, embellished with wigs and beards of human hair. On the western islands, large shield-shaped human masks may have been worn and were certainly employed as shrines with trophy skulls attached to them. Large effigies of human beings, crocodiles, and sharks—some as large as life-size—were constructed for initiations and kept in sacred fenced enclosures. Only on Saibai Island, off the southern New Guinea coast, were masks consistently carved in wood. They are extremely elongated, with long pierced ears and crescent-shaped, toothed mouths, and were worn in harvest times. Wood sculpture was otherwise restricted to representations of human heads, which were attached to canoe prows, and to small figures of humans, turtles, dugongs (sea cows), and other animals used for sexual and fertility magic. Magic for rainmaking involved small stone figures.

The use of tortoise-shell

New Guinea. The visual arts of New Guinea are rich and highly complex. Fortunately, the great number of styles that exist can in many cases be subsumed into larger groupings corresponding to geographic areas. Moving clockwise from the extreme northwest of the island, the primary style areas are Geelvink Bay; Humboldt Bay (Teluk Yos Sudarso) and Lake Sentani; the prolific Sepik River region, which is subdivided into numerous smaller groups; Astrolabe Bay; the Huon Gulf; the Massim area; the Gulf of Papua; Fly River; the Marind-anim region; and the southwestern coast. The central Highland ranges of the island also constitute a major style area.



Figure 8: Canoe prow, wood, paint, cassowary feathers. From Geelvink Bay, Irian Jaya. In the Museum of Ethnology, Basel, Switz. Height 200 cm.
Museum fur Volkerkunde, Basel, Switz. (Vb 5960); photograph, P. Horner

Geelvink Bay. The Geelvink Bay area, including several offshore islands, is located at the northwestern end of New Guinea in the Indonesian province of Irian Jaya. Its style of sculpture seems closely related to those of such eastern Indonesian islands as Tanimbar and Leti, probably as a result of relatively recent influences. The most famous works from the area are the *korwar* figures, small statues embodying the spirits of ancestors; they were used by shamans to divine the outcome of important ventures, illnesses, and other critical situations. They generally depict standing or seated males with disproportionately large heads, befitting their function as either representations of or actual containers for skulls. The heads are as a rule strongly rectangular, with sharply cut brows, small round eyes, and long anchor-shaped noses. The figures usually incorporate local features, such as subsidiary figures, headgear, figures of snakes, or openwork shields.

The *korwar* style

The carved designs of scrolls and spirals found on the *korwar* shields were also frequently used to decorate the vertical panels forming the prow ornaments of canoes and a multitude of small objects, including headrests and bamboo tobacco containers.

Humboldt Bay and Lake Sentani. The area around Humboldt Bay and Lake Sentani is one of intensive stylistic interaction. A striking example of this interaction can be seen in the diffusion, in the early 19th century, of a pyramidal type of ceremonial house from the eastern coast to Humboldt Bay and subsequently inland to Lake Sentani. The houses had human-shaped finials roughly carved of fern wood and, projecting from their walls, long poles terminating in bird and fish figures. Variations existed, of course, and in general the Sentani ceremonial houses were less elaborate, but the houses of chiefs were equipped with figures standing on short posts protruding up through the floor. The central posts supporting the ridgepole were also carved in human form.

The western edge of the area is best known for its small carvings of human figures and for its carved canoe ornaments. The figures are squat, even bulbous in limb and body; they have neckless, globular heads, with long projecting, sharply pointed noses. The canoe carvings basically consisted of two horizontal rectangular panels united in a point at the front, on which stood a carving of a human head, sometimes with a bird perched on it. Tied behind the head was a separate carving: a vertical rod with crosspieces terminating in either human heads or bird figures. Relief or incised details were picked out in colour.

To the east, closer to Humboldt Bay, the prow carvings were S-shaped and depicted the body, neck, and head of a long-billed bird; subsidiary figures of fish and other creatures were arranged on the bird's body. Carved human figures were columnar, with the arms and shoulders in low relief and displaced forward almost to cover the chest. The figures had ovoid heads with down-slanted brows, circular eyes, and toothy crescent-shaped mouths.

Painting on tapa was common throughout the Humboldt-Sentani area, largely for women's skirts. At Lake Sentani the style was somewhat linear, using double spirals (also a common carving motif) and fish or bird forms with exaggerated V-shaped tails. Humboldt Bay tapas were denser in design, with the entire field covered with larger and bolder forms.

The Sepik River regions. Roughly 200 separate groups speaking distinct languages live around the Sepik River. As might be expected, the variety of artistic styles found among these groups is bewildering, but three visual elements seem to be basic to nearly all the styles in varying degrees: (1) designs in which two triangular forms are connected at their bases or apexes, often with further design elements in the angles so formed, (2) sculpture based on vertical series of hooklike forms that can be either unidirectional or in opposed groups, and (3) naturalistic representation of natural objects. The interplay of these three elements in various styles suggests that the first two elements preceded the third. The Sepik areas treated in this discussion are, moving clockwise, the northwestern coast, the central coast, the eastern coast, the southern tributaries, the South Sepik Hills, and the upper Sepik.

The northwestern Sepik area

The styles of the northwestern Sepik area are closely related to those of its western neighbour, the Humboldt-Sentani area. Forked-tailed zoomorphs, used on canoe prows and paddles, and pyramidal houses are common in both regions. The art of the northwestern Sepik groups, however, is based predominantly on the triangular design described above. Sculpted figures are rare in the area. The most conspicuous works are shields, which show many variants of the triangular design. Among the Olo tribe, for example, the triangles are formed from a group of scrolls.

Collection: Australian National Gallery, Canberra



Figure 9: Double figure from a house post, wood. From Lake Sentani, Irian Jaya. In the Australian National Gallery, Canberra. Height 176.6 cm.



Figure 10: Memorial board, wood. From the Sawos people, Sepik central coast, Papua New Guinea. In the Museum of Ethnology, Berlin. Height 208 cm.

By courtesy of the Museum fur Volkerkunde Staatliche Museen PreuBischer Kulturbesitz, Berlin (VI 46172); photograph, Dietrich Graf

Triangular designs can also be found painted on bark sheets used by various groups for initiations and on huge conical masks used by several groups in healing rituals. The Telefomin carved the designs onto tall boards used as house entrances. Similar boards were used to create whole facades by neighbouring tribes. Some tribes used the triangular motif in conjunction with an S-shaped double-spiral design on tobacco pipes, hand drums, and bark paintings.

The north-central section of the Sepik region stretches from the coast to Lake Chambri just south of the Sepik River. The major groups in the area are the Boiken, the Abelam, and the Sawos and the Iatmul.

The Boiken styles, which appear to have been numerous, are relatively little known. Their most conspicuous monuments are the ceremonial houses, which follow on a smaller scale the pattern of Abelam houses to the west. Masks of the coastal Boiken were in a long-nosed style; others were made in basketry. Basketry was also used for a variety of small masks, bird figures, and abstract forms that were attached to large turbo shells used as valuables. Figure sculpture was rare, but the Boiken had rich traditions of pottery making. Food and cooking pots were elaborately decorated with engraved designs and were widely distributed, especially to the river people.

The art of the Abelam tribe, which lived in the Prince Alexander Mountains, was tied to a vigorous ceremonial life. It thus presents a far more spectacular scene. Their pyramidal ceremonial houses, centres for cults of yam growing and initiation, were built on the grandest scale known in New Guinea. They featured vast painted gables and lintels, to which carvings of hornbills, parrots, and lizards were attached. The carvings were in every instance augmented by paint, which indeed the Abelam considered magical in itself.

There are three basic styles of Abelam sculpture. The figure sculpture of the north consists of simple, bulbous forms in massive but sketchy conformations, with detail largely supplied by painting in yellow, black, and white over a predominantly red ground. The eastern style is now similar to that of the north, although somewhat less dependent on polychromy. At an earlier period, the eastern sculptures were elongated, with a human head at one end and the rest of the figure consisting of clusters of bird heads. In both the north and east, major sculptures were often monumental in scale, some 20 feet long. Large openwork panels were also carved, showing humans, animals, and birds. Figures in the southern, or Wosera, style are generally standing; they have ovoid heads that are often surmounted by birds.

Painting styles also varied. Bark paintings found on the ceremonial house gables of the northern Abelam are broad, large-scale depictions of spirit faces, figures, and animals. Paintings by the southern Abelam tend to be smaller in scale and painted not in flat areas of colour but with much fine line and cross-hatching.

Masks, which were worn for initiations, were generally confined to basketry hoods with elaborate openwork eye panels and noses. Small basketry masks were attached to yams during rituals, and men wore pointed basketry crests as hair ornaments. This pointed form was repeated among the Wosera on a huge scale as a ritual headpiece made of feathers.

The Abelam made a wide range of small decorated objects, including cups, spoons, whistles, and spinning tops in coconut shell; arm ornaments, daggers, and gouges in bone; spears, digging sticks, hand drums, and stirrers in wood; and pots in clay. All were incised with human faces or with closely spaced, complicated designs incorporating the typical Abelam scroll and oval patterns.

Sawos-Iatmul art

The Sawos and the river-dwelling Iatmul, who historically derive from the Sawos, worked in styles totally different from those of the people to the north. Their ceremonial houses were long rectangular structures, with upper stories elevated on posts often carved with ancestral faces and figures. The gables were not of exaggerated size but had masks in wood or basketry. King posts, which had female figures carved at their bases, extended high above the house roofs and were topped with carvings of human beings grasped by eagles.

Human-figure sculpture was a major theme in Iatmul and Sawos art. Human figures and faces and a wealth of curvilinear ornament adorned numerous sacred objects, including flutes, slit gongs, trumpets, drums, and an array of less familiar musical instruments that simulated the voices of spirits. They were also to be found on such mundane equipment as stools, headrests, bowls, palettes,

Copyright © 1990 The Metropolitan Museum of Art, New York City, The Michael C. Rockefeller Memorial Collection, bequest of Nelson A. Rockefeller, 1979 (1979.206.1480)



Figure 11: *Mei* mask, wood, paint, shell, reed. From the western Iatmul people, Papua New Guinea. In the Metropolitan Museum of Art, New York City. Height 71 cm.

tools, weapons, and canoes. As a rule the figures were naturalistic within the limits of certain standard conventions, which varied between the eastern (Parambei) and the western (Nyaura) Iatmul. The figures of the east tended to be more gracile than those of the west, which were frequently stocky and burly. The profiles of faces on eastern Iatmul figures often had a graceful S-curve, while those from western Iatmul and the Sawos had heavy jaws, high cheekbones, and sunken eyes under horizontal brows. These same features characterized the long-nosed wooden *mei* masks of the Iatmul. Other types of masks, however, represented mythological birds, crocodiles, fish, and other animals. These were generally constructed of basketry and painted bark and were often of great size.

Perhaps the most striking material used in Iatmul-Sawos art was human skulls. These enthusiastic headhunters covered the skulls of victims and ancestors with clay and painted them in the patterns used in life. The skulls were then displayed on racks made of painted bark sheets or were mounted on puppets for use at initiations and funerary ceremonies.

In the far-eastern section of the Sepik region, around the Ramu River, the peoples living along the coast and on offshore islands engaged in extensive cultural exchanges, trading dances, masks, slit gongs, and carvings. The Murik people at the mouth of the Sepik River were particularly active in this regard. Tribal styles thus spread widely. In some areas local styles incorporated or were supplanted by imported styles, but in many localities a multitude of distinct styles existed side by side.

Although styles vary, most figure sculpture from the eastern Sepik depicts standing males (females exist but are unusual). The figures range in size from miniature to larger than life. They have ovoid heads that droop forward and limbs that are slightly flexed. Some are fitted with actual human skulls. Both figures and masks frequently feature enormously exaggerated noses, which signify masculinity (females have short noses). Besides ritual objects, a wide range of utilitarian equipment, from canoe prows to bowls, was decorated with carved representations of humans, birds, and animals. The carvings were often augmented by the tight geometric patterning that was characteristic of Murik art in particular.

Through the flat, swampy country west of the eastern coastal hills, several tributaries flow northward to the lower Sepik, each associated with a particular artistic style. These rivers are, from east to west, the Porapora, the Keram, and the Yuat. The art of the Porapora area is related to lower Ramu styles but is less elaborate and profuse. Ceremonial house posts were carved with figures in a plain, almost geometric style—a style that was also used in carving stafflike figures with dishlike receptacles for ancestral skulls.

By courtesy of the Museum fur Volkerkunde Staatliche Museen PreuBischer Kulturbesitz, Berlin (VI 38609); photograph, Dietrich Graf



Figure 12: Featherwork panel, feathers tied to a wooden panel. From the Kambot people, Keram River, East Sepik province, Papua New Guinea. In the Museum of Ethnology, Berlin. Height 88 cm.



Figure 13: Flute stop in the form of a male figure, wood, cowrie shells, fibre, cane, cassowary feathers, human hair, mother-of-pearl, polychrome. From Yuat River, lower Sepik, Papua New Guinea. In the Baltimore Museum of Art. Height 49.53 cm.
Baltimore Museum of Art, gift of Alan Wurtzburger (BMA 1955.251.36)

The southern tributaries

The Kambot tribe of the Keram River, on the other hand, combined sculpture and painting in complex, ambitious designs to decorate their ceremonial houses. The houses' long, horizontal gables were filled with painted compositions of an ancestral hero with his wives and animals. Paintings also adorned the interiors, and the gable painting was often replicated on a grand scale in feather mosaics on wood slabs—a unique technique in the Sepik. Sacred objects included large panels of basketry that had human skulls attached and were decorated with clay, shells, and boar tusks. Small versions of the panels were attached to sacred flutes. Wood carvings included rectangular shields, which were engraved and painted, and small-featured hemispheric or oval masks. Huge figures of crocodiles were constructed of painted bark sheets for initiations.

The Yuat River people, especially the Biwat (Mundugumor), carved slit gongs, shields, masks, and various types of figure sculpture. Masks, like those of the Kambot, were usually hemispheric. Small figures used as flute stops had grossly enlarged heads that projected forward; they were often carved in conjunction with parrots and other creatures. Masks, as well as wood snakes used in sorcery and other such objects, often bristled with spike forms, which are a common motif in Biwat art. In relief carving, such as can be seen on shields, almost every line or band is serrated, creating a dazzling effect. The same technique was used in enormous paintings of crocodiles that were displayed at yam-harvest ceremonies.

The southern Sepik Hills

The opposed-hook style of Sepik sculpture was predominant along the middle reaches of the Sepik River and among the hills ranging across the southern border of the Sepik valley, including the Hunstein Mountains. The most spectacular works in this style were figures carved by the Alamblak in the eastern Sepik Hills. The figures, known as *yipwon,* represent patron spirits of hunting and war. They are topped by a downcurved hook; directly beneath

this is a human face, and below that is a vertical series of downcurved hooks. An oval element, representing the heart, appears next at about the centre of the figure; below the heart is a series of upturned hooks, and the whole is supported on a single leg. Small examples of these *yipwon* were personal amulets; larger figures, up to two metres high or more, were clan-owned property kept in ceremonial houses.

A second type of carving has also been recovered, usually from burial rock shelters of the Ewa, a now much-diminished group south of the Alamblak. These figures are related in general form to the *yipwon,* but their bodies are expressed as panels and scrolls rather than hooks. Other flat figures are of females in frontal positions with raised arms and hands.

The Bahinemo west of the Alamblak carved opposed-hook objects with no head or leg. They also made masks, for display only, which incorporated hooks and human features; these represented bush and water spirits. Groups farther west made hook carvings of the Bahinemo type and also carved hook patterns on shields and slit gongs. Other hook carvings are of uncertain provenance. They often have right-angled rather than curved hooks.

That the hook style was once predominant throughout much of the Sepik area is suggested by traces of it in still other styles. South of the river, masks of some Yuat, Keram, and middle Ramu rivers groups are framed in series of hooks. More remarkable, from far north of the Sepik, a number of the Abelam's carvings incorporated opposed hooks in the form of bird beaks.

The upper Sepik

A number of small groups lived along the upper reaches of the Sepik River. The most productive were the Kwoma. Like the Abelam, they celebrated yam cults in ceremonial houses that were basically roofs supported on posts, without walls. The ridgepoles of the houses were carved with mythical characters, human and animal. The ceilings were covered with bark paintings with semiabstract designs recalling characters and incidents in myths; the finials on the gables were also carved with mythical figures and birds. Similar designs were used on pottery feast bowls and on daggers made from human bones.

The main nonarchitectural carvings—*yena,* human heads; *mindja,* long boardlike carvings with a head at one end; and *nogwi,* figures of women—were made for the three main rituals of the yam cult. The carving style is simple and massive, with heads having straight brows above a slightly concave facial place on which appear conical eyes, a long, heavy nose, and a small V-shaped mouth.

Copyright © 1990 The Metropolitan Museum of Art, New York City. The Michael C. Rockefeller Memorial Collection, purchase, Nelson A. Rockefeller gift, 1965 (1978.412.854)



Figure 14: *Yipwon* figure, wood. From the Alamblak people, East Sepik province, Papua New Guinea. In the Metropolitan Museum of Art, New York City. Height 128 cm.



Figure 15: Dance paddle, carved wood, paint. From the Trobriand Islands, Papua New Guinea. In the Field Museum of Natural History, Chicago. Height 69 cm.
Field Museum of Natural History, Chicago (111604c)

Astrolabe Bay. The people of Astrolabe Bay, southeast of the coastal Sepik-Ramu area, carved as their most important works large ancestor figures, few of which now remain. Most of the figures are standing males, posed frontally. Their shoulders are hunched well forward of the torso, their arms hang straight down, and their hands are placed horizontally on the hips. The face is triangular with the chin extending below the chest; ornaments project downward from the mouth. A long, narrow nose hangs down from rigidly horizontal brows; the protruding eyes are round and staring.

Other works from the area include oval masks with teardrop-shaped eyes, enormous arched noses, and elaborate openwork ears. Shields were disk-shaped and were decorated with relief carvings of X-shapes and circles. Bowls, arm shells, drums, and bull-roarers were incised with small geometric patterns. The same angular patterns were used for tapa paintings.

The Huon Gulf. The cultures and art styles of the Huon Gulf have strong links to those of both Astrolabe Bay and southwestern New Britain, especially in architecture and carving, which was made in large quantities. The main theme, the human figure, was expressed in blocky, almost cubist forms, with the nearly rectangular head sunk deeply between the shoulders. Large freestanding examples of such figures can be found in both standing and kneeling positions. Smaller figures were used as supporting poles for ceremonial houses, as the shafts of suspension hooks, as ladle handles, and as the supports for headrests and betel mortars. Masks, in bark cloth or wood, used the same convention. Other types of objects, however, were decorated with geometric patterns.

Tami Island bowls

The most famous products of the area are the large, shallow, basically oval bowls that were made on Tami Island and traded to the mainland and New Britain. Most have a human face carved at one end, with the rest of the bowl serving as an elaborate headdress; others were carved in the forms of birds and fish. The designs were incised and filled in with lime to stand out against the black background.

The Massim area. The islands off the extreme southeastern tip of New Guinea were linked by the kula trading cycle, which distributed not only shell valuables—the ostensible motive of the transactions—but also quantities of other goods. Notable among these were carvings in dark hardwood, which was the special product of Kiriwina, the largest of the Trobriand Islands.

The great variety of design motifs ranged from abstract shapes to both stylized and naturalistic bird, human, and animal forms. Incised designs most frequently featured curvilinear patterns, which could be easily adapted to represent stylized snakes or birds; the incisions were usually filled in with lime to make the design stand out. Among the items carved were mortars and spatulas used to prepare betel nut; long, flat war clubs; splashboards and decorative panels attached to the prow and stern of seagoing canoes;



Figure 16: Ancestral board, wood, paint. From the Gope people, Wapo Creek, Gulf province, Papua New Guinea. In the Metropolitan Museum of Art, New York City. Height 82.82 cm.

Copyright © 1990 The Metropolitan Museum of Art, New York City.
The Michael C. Rockefeller Memorial Collection, gift of Nelson A. Rockefeller, 1969 (1978.412.794)

and dance paddles (two semicircular panels connected by a handhold bar). Dance paddles were sometimes painted, but, in general, painting of wooden objects was minimal. Painting was mainly used to decorate the gables of yam storehouses and on convex oval war shields.

The Gulf of Papua. The succession of cultures situated along the vast Gulf of Papua and in the deltas of the rivers flowing into it produced one of the richest complexes of art styles in New Guinea. In general, the people believed that they owed much of their basic culture to Kiwai, the large island at the mouth of the Fly River to the west, even though their societies showed important local variations. The groups who lived in the swamplands of the west were cannibalistic and practiced orgiastic rites; those who dwelt on the beaches of the east were not given to either practice. All, however, built huge longhouses—in the west these were communal dwellings, in the east they were reserved for men. Most of the groups shared certain types of masks, as well as carved sacred boards with ancestral or supernatural representations in relief. Much of the carving is, indeed, two-dimensional.

The Bamu-Turama area

At the far west of the western area, the art of the Bamu and Turama rivers is largely a somewhat geometrized version of Kiwai sculpture, including some oversize human figures. Two other types of objects were universal in the area from the Bamu River to Goaribari Island at the midpoint of the gulf. They are a dome-shaped basketry mask, which was usually covered with clay and painted and featured a long protruding nose, and sacred boards in quasi-human form. The oval-shaped sacred boards have a face at the top, the indication of a neck, and vertical slots on the body that suggest arms flanking vertical uprights or drawn-up legs. The boards were kept in shrines, as male and female pairs, with human skulls suspended from the uprights. Three-dimensional works from the western gulf had simpler forms—often retaining the natural shape of the wood—but they were carved in elaborate relief.

Farther east, around Wapo Creek, the Era River, and Uramu Island, sacred boards had faces, but, instead of bodies, the field shows a vertical sequence of floating abstract designs, which can be read as extremely stylized anatomic elements. The small silhouetted human figures had upraised arms but were not used as skull racks. Masks in the area were sometimes dome-shaped, but, unlike those from the Bamu-Turama section, they had protuberant jaws. The eastern section of the gulf also developed a flat, oval mask made of basketry covered in bark cloth; the masks were almost identical in shape with the sacred wooden boards and were painted with similar designs.

The lower Fly River. On Kiwai, the large island at the mouth of the Fly River, initiation was marked by the display of naturalistic, almost life-size figures of ancestral men and women. Their heads were virtually identical with the masks made on Saibai Island in the Torres Strait. Small pendant figures from Kiwai were, however, extremely stylized and flat and were covered with a chevron pattern. Much the same facial design was used on canoe splashboards. Large figures were also carved on the coast along the river in a distinctive style in which faces have squared-off jaws. This style of face was repeated on the shafts of the great numbers of arrows made for trade.

The Marind-anim. The people of the coast and hinterland areas of New Guinea northwest of the Torres Strait and east of Frederik Hendrik Island (Pulau Yos Sudarsa), in what is now the Indonesian province of Irian Jaya, included the large tribe of the Marind-anim. Their material culture was limited, except in one respect: the ephemeral art produced for the celebrations of their initiatory cults. The most elaborate performances of the three main cults were held by groups living along the coast.

Dema

The central figures of Marind myth are the *dema;* these were considered not only the ancestors of the present clans but also the creators of all the elements of the world. The *dema* were represented at initiations, which could take from several days to many months to perform, by costumed men and by effigies. The costumes were, if not naturalistic, highly allusive accumulations of objects that recalled the *dema* and their creations. The wearer's basic disguise was a fibre costume. He carried on his back or head at least one large effigy carved in softwood; the effigy was partially painted, but it was mainly decorated with white, red, and blue seeds. The costume was also hung with flat semiabstract panels covered with seeds. Small masks and feather headdresses completed the assembly.

This complexity of style did not carry over into Marind sculpture, which was simple, sometimes crude. The largest works were tall posts with animals and geometric designs carved in high relief; these were used for temporary feast

Natural History Museum of Los Angeles County (A. 1931.31-61)

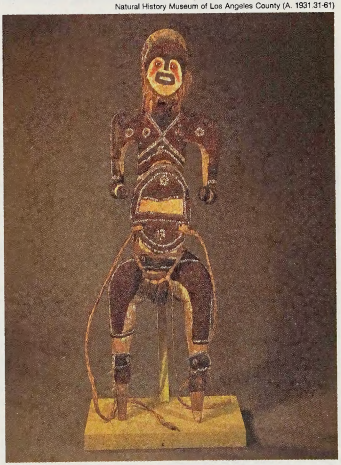


Figure 17: *Deme,* coconut god, wood, seeds, paint. From the Marind-anim people, Irian Jaya. In the Natural History Museum of Los Angeles County. Height 152.5 cm.

houses. Sometimes large posts, with upright winglike projections on either side, were erected as grave markers. These may be akin to carved forked posts on which the Marind hung head-hunting trophies.

The southwestern coast. The two main groups living on the southwestern coast of New Guinea between the Vogelkop Peninsula (Jazirah Doberai) and Frederik Hendrik Island are the Mimika (Kamoro) to the west and the Asmat to the east. Their styles have much in common.

Carving was the major art among the Mimika and was usually closely associated with initiatory and other rituals. Human figures were the primary subject. They were typically depicted standing with slender limbs. The legs were usually slightly flexed; the hands were held at chest level or under the chin, and the head was a short cylinder with miniscule features. Incised designs were used to decorate the figures, but the repertoire of patterns was limited. The most important design element was a pointed oval, representing the navel, a symbol of birth and, by extension, of life itself.

Mimika funerary poles

The largest works of the Mimika are tall poles carved as funerary memorials. These generally consist of a vertical series of figures, with the uppermost figure holding a large triangular panel that projects outward and is carved with the standard designs in openwork. Larger-than-life-size figures of pregnant women used in rites devoted to increasing human fertility were usually displayed at the ceremonial house during a ritual celebrating the creation of the cosmos.

Canoe prows, hand drums, boards, and other objects were decorated with openwork or low-relief carving for ceremonial purposes. Masks were not carved but were made of netting. They were usually conical or hoodlike and sometimes had wooden eyes attached but were generally featureless. The netting section was attached to cane epaulets and bands around the chest. Painting and the decoration of domestic objects were relatively uncommon among the Mimika.

Asmat art, including the art of some neighbouring groups, presents a similar, but somewhat richer, picture. Four style areas can be distinguished, two on the coast (northwest and central) and two inland, with many local variations. During the 20th century an increasing naturalism has affected figure sculpture, particularly among the central

University Museum of Archaeology and Anthropology, Cambridge, Eng. (1914.231.254)



Figure 18: Female figure, wood, paint. Mimika people, Utakwa River, Irian Jaya. In the University Museum of Archaeology and Anthropology, Cambridge, Eng. Height 183 cm.



Figure 19: Mask, rattan, sennit, and other materials. From the Asmat people, Upper Pomatsj River tributary, Irian Jaya. In the Metropolitan Museum of Art, New York City. Height 203.2 cm.

Copyright © 1990 The Metropolitan Museum of Art, New York City, The Michael C. Rockefeller Memorial Collection, gift of Nelson A. Rockefeller and Mrs. Mary C. Rockefeller, 1965 (1978.412.1274)

Asmat; size, too, has tended to increase. Sculpture is the primary art form; indeed, it has a mythical character in that humanity itself is said to derive from carvings made by a creator hero. Most carvings are made in conjunction with a cycle of feast-ceremonials held in a set sequence. Painting is unknown except on carvings, where white tends to predominate, with some red and black. Seeds of various colours are used decoratively, particularly on the skulls preserved by these cannibalistic headhunters.

The central Asmat are known for their large memorial poles, called *mbis* or *bisj*. The poles are similar in design to those of the Mimika. The openwork carvings on the triangular projections incorporate crescent shapes, S-shapes, stylized hands, birds' heads, and other symbols of headhunting. The figures on the poles represent or commemorate clan members who were killed by enemy tribes. Smaller versions of the poles were used in the interiors of the ceremonial houses as supports for crossbeams. The poles have a symbolic relationship to canoes and sometimes stand on a canoe-shaped base. Canoe prows were often decorated with large carved figures and head-hunting motifs similar to those on the *bisj* poles. Head-hunting symbols were also frequently carved onto the sides of canoes and onto memorial carvings in the form of stylized crocodiles. Wooden war shields in the central Asmat area are rectangular in shape, with a small phallic projection at the top. The front surface is carved with a bold design of crescent shapes, outlined in low relief by narrow bands.

In contrast, the shields of the northwestern Asmat are rounded at the ends, with a human face, or sometimes an openwork human figure, carved at the top. The front is carved with a vertical row of symbols (usually flying foxes), and the field is densely covered with small subsidiary designs. The largest works by the northwestern Asmat are enormous "soul ships" displayed at initiations. They follow the general form of canoes, except that they do not have bottoms. They do have elaborate prow and

Asmat "soul ships"

stern carvings, and a number of carved figures representing water spirits sit in the body of the canoe.

Asmat masks, like those of the Mimika, are woven. They often feature wooden eyepieces or are crowned with a flat wood silhouette of a tortoise. They cover not merely the head and shoulders but also much of the torso. The rest of the body is screened with a bushy mass of fibre strips. The masks were particularly used in a ceremony to expel spirits of the dead from the village.

Carving in the inland areas was restricted; it was used for the ornamentation of spears and digging sticks but was prominent only on shields. The designs are more geometric but are otherwise similar to those of the central Asmat.

The Highlands. The central cordillera of New Guinea is inhabited largely by agriculturalists. Although the area has the highest population density of the island, it is also the least productive of works of visual art of a permanent nature.

In the eastern Highlands in Papua New Guinea, shields are painted with geometric designs related to those of the Telefomin and other groups to the north. Figure representations are limited to images made of coiled basketry. Such figures are carried at festivals in the southern Highlands. The most remarkable are the squat fertility figures (*yupin*) of the western Enga. Thin boards carved in openwork are used to represent the dead in festivals farther east.

Masks are fairly common throughout the area, especially in the eastern Highlands. They are generally made of gourds, with hair, feathers, teeth, and other materials applied for features. The Asaro River "Mudmen" are particularly well known for the grotesque imagination they display in making their clay masks.

Body decoration

The most important manifestation of art in the New Guinea Highlands is body decoration. Although decoration usually involves an individual's entire body, the main focus is the head, which is adorned with a great variety of hats and wigs. The most striking materials used are feathers. The area is famous for its many species of birds of paradise, and it is their magnificent plumage that the Highlanders exploit. Each major tribal group has its own local style, which is rich in symbolic content.

The Admiralty Islands. The traditional culture of the Admiralty Islands, which lie northeast of the Sepik River area of New Guinea and are administratively part of Papua New Guinea, is now practically extinct. Although the population consisted of many different language groups, the people have been popularly divided into three artistic style groups: the Usiai, who lived in the interior of Manus Island (Great Admiralty Island), the largest of the Admiralty Islands; the Matankor, who lived on the small islands to the north, east, and southeast of Manus; and the largest group, the Manus, who lived on the southern coast of Manus as well as on some offshore islands. Each group relied on the others for some items of food and manufactures; trade was active and frequent.

The Matankor produced wood carvings and decorated objects, each island having its own specialties. For example, the people on Baluan made bird-shaped bowls, ladles, and spatulas; on Lou, obsidian was carved into great hemispheric bowls; on Rambutyo figures and anthropomorphic lime spatulas were common; and the people on Pak made beds (used nowhere else in Melanesia) and slit gongs. Although the Matankor were neither culturally nor linguistically homogeneous, their art style shows a considerable uniformity. Surface designs consisted largely of repeated triangles, diamonds, rectangles, and opposed curves, often in bordered bands, sometimes in openwork or relief. These busy, if repetitious, patterns were often accented in black and white on expanses of red background; generally they were employed as strips or in small areas rather than covering an entire object.

Human or crocodile figures are common themes and were used to top house ladders, bed frames, and bedposts and to decorate canoe prows and sterns. Freestanding oversize male and female figures, probably of ancestors, were placed on either side of the doors of men's houses. Human figures also appeared on slit gongs, with the upper body carved as one projecting lug and the legs at the gong's other end. The human figure was usually shown standing, with the arms hanging straight down and the hands either free or placed on the hips. The torso and limbs tended to be square in section, but the calves were angular. The head was ovoid in shape, and the mouth was often set at a right angle to a prognathous, muzzlelike face. A characteristic male hairstyle consisted of a cylinder capped with a sphere rising from the top of the head.

Among the most impressive Matankor objects are the hemispheric bowls on four short legs. These were equipped with a pair of spiral handles representing crocodiles' tails, and bands of designs were carved around the rims. The standard repertoire of abstract and representational motifs was also adapted to small carved objects, such as ladle handles. A favourite ornament was the *kapkap,* a breastplate consisting of a carved tortoiseshell plate mounted on a giant clam shell.

New Ireland. The long, narrow island of New Ireland shows three distinct style areas: the northwest, the centre, and the southeast. The first area is celebrated for its *malanggan* carvings and masks, which share their name with a series of religious ceremonies held primarily as funerary celebrations but also (by extension) for the validation of land claims, the establishment of subclans, and other important events. In this most elaborate style of Oceania, the usual form of the face has horizontal brows, with the deep-set eyes inlaid with bright sea-snail opercula; the nose is strongly arched and massive; and the jaws, broad horizontal rectangles, show a formidable array of serrated teeth. The fully three-dimensional figures usually have added attributes, including bird and animal forms; they often clutch frameworks of rods, which enclose them. Flat areas are pierced in intricate patterns, a technique that was probably fostered when the islanders acquired steel tools. All works were painted for fullest effect in sharply defined areas of black, white, red, and yellow; small sections of black cross-hatching and other patterns often further enhanced the design.

Malanggan carvings

Malanggan carvings on poles display either individual figures or several figures stacked vertically. Freestanding carvings often illustrate mythological incidents and can be of great size. Pigs, birds, and fish are the subjects of other carvings. Seated figures used for rainmaking ceremonies were constructed from tree trunks, bamboo, and other materials; they had raised hands and were fitted with carved heads.

Some *malanggan* masks are almost indescribably complicated, with the basic style of carved face adorned with long vertical tusks and other protrusions, fitted with openwork side panels of birds and fish, surmounted with birds, snakes, and figures, and enclosed within a lattice of bars. Some simpler types have seminaturalistic faces.

Other *malanggan* carvings are horizontal panels. There are several types, including one with a fish head at each end and human figures between them, another with apertures in the middle through which men put their heads, and one with representations of the moon. One type of carving that depicted a bird struggling with a snake was sometimes mounted on the head of a figure.

Uli figures

In central New Ireland the primary objects of mortuary cults were carvings known as *uli.* These are standing figures with female breasts and male genitals; they sometimes have raised hands and may support smaller figures in front of them or on their shoulders. The head is usually large and is topped by a thin, upright crest; the eyes are inlaid with shell, the nose is hooked, and the

By courtesy of Musee Barbier-Mueller, Geneva (4402-A); photograph, P.A. Ferrazzini

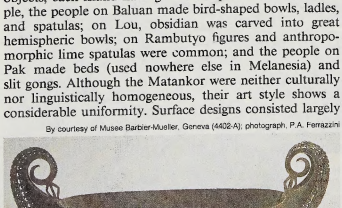


Figure 20: Bowl, wood. Admiralty Islands, Papua New Guinea. In the Barbier-Mueller Museum, Geneva. Height 63 cm.

wide mouth exposes the teeth above a triangular chin. The body of the *uli,* like that of the *malanggan,* is often enclosed in sweeping bands. The intricate polychromy of the *malanggan* is absent, however; white is the main colour, with touches of red and black. Small *uli* were perched on conical constructions; large ones were housed in similarly conical huts. The ceremonies associated with the *uli* were elaborate, but their significance—apart from a relationship to fertility and warfare—is obscure. Wood figures in the same powerful style were topped with skulls over which clay had been modeled; these were used in rainmaking as well as in mortuary ceremonies. Among some central groups, mortuary ceremonies also featured a large bark and cane disk with a central aperture framed by petallike projections. The disk was painted red and yellow and was kept in a hut with posts carved with the same emblem, apparently of the sun. Skulls were displayed in the disk's central aperture. The design was echoed in the *kapkaps,* which were worked with exquisite delicacy in central New Ireland.

The southeastern style area of New Ireland includes small nearby islands and the northern Gazelle Peninsula of New Britain. In the north of this area, figure sculpture takes the form of small chalk figures of males and females, with rounded faces, round eyes, straight noses, and wide, toothy mouths. The hands are joined in front of the torso. The white chalk is accentuated with touches of black and red—a colour scheme prevalent throughout southeastern New Ireland. Sometimes said to have been made for mortuary ceremonies, the chalk figures more probably were used by secret societies such as those on the Gazelle Peninsula.

Several types of masks were made in the area. The masks of the Tanga Islands were ephemeral constructions of bark and fibre over bamboo frames. They were semiconical in shape, with long backswept ears, thin upturned noses, and extended chins or beards. On the neighbouring mainland, masks were made of the same materials but were more naturalistic. Masks from the southwest were made of wood and had faces similar to those of the chalk figures. The end walls of houses were frequently screened with planks that were incised or painted with small and sparse units of design, often showing stylized animals and fish. Architectural sculpture, however, was rare except in the Tanga Islands and on the southwestern coast.

New Britain. Knowledge of art in New Britain has largely been limited to the coastal areas and to the Gazelle Peninsula in the northeast. Masks, dance shields, and other ceremonial objects are the primary works.



Figure 22: *Uli* figure, wood. From the northeast coast of New Ireland, Papua New Guinea. In the Museum of Ethnology, Basel, Switz.

Museum fur Volkerkunde, Basel, Switz.; photograph, Hans Hinz, Basel.

The Tolai people on the coast of the Gazelle Peninsula probably emigrated from southeastern New Ireland and thus share certain style characteristics, such as boomerang-shaped canoe prows, with that area. The human figure is a common subject of Tolai art and is almost always depicted standing, with arms bent and hands held to the ears. Carved faces are naturalistic, sometimes with long beards, but in paintings the face is often reduced to round eyes and a crescent-shaped mouth. Other common motifs are disks with long triangles below them and spirals.

Tolai art



Museum fur Volkerkunde, Basel, Switz.; photograph, Hans Hinz, Basel

Figure 21: Cult house with *malanggans.* The house is made of wood, bamboo, small palm leaves, and croton leaves. The figures are carved in wood and painted with earth pigments and yellow oil colour. The house was erected to commemorate 14 deceased persons from the village. From Medina, New Ireland, Papua New Guinea. In the Museum of Ethnology, Basel, Switz.

Much Tolai art was incorporated into the ritual of two male secret societies, the Iniet and the Dukduk. Iniet initiations were held in walled enclosures lined with paintings of human figures. Long panels of openwork carving showing human figures, animals, and abstract designs were carried in one initiation dance, while the frontal bones of human skulls, which had been over-modeled, painted, and embellished with hair and beards, were worn as masks in other dances. Wooden human figures of various sizes, as well as small chalk or soft stone figures (usually human but sometimes of an animal), were also used.

The Dukduk society used male (*dukduk*) and female (*tubuan*) masks. Both types are cone-shaped and were constructed of cane and fibre. The *dukduk* is taller than the *tubuan* and is faceless. The *tubuan* has circular eyes and a crescent-shaped mouth painted on a dark background. Both masks have short, bushy capes of leaves.

The mountains south of the Tolai's coastal area are inhabited by the Baining, who consist of several groups of seminomads. Virtually their only works of art were masks and other objects carried in dances; these, however, being constructed of light materials (bamboo covered with bark cloth), were often of great size. The most remarkable came from the Chachet (northwestern Baining), who constructed figures up to 40 feet high for daytime mourning ceremonies. The Chachet figures had essentially tubular bodies with rudimentary arms and legs and tall heads with gaping mouths and painted eyes. Among other Baining groups, the best-known type of mask consisted of a flat upper panel, which was either circular or divided into two lobes, and a gaping mouth, from which hung a chinlike or tonguelike form. Two enormous circular eyes were painted on the flat panel. All the Baining groups used, in addition to masks, dance headpieces made of painted vertical panels or poles topped with images painted on bark cloth. The masks and the painted designs represented many items of the natural world.

Sulka masks

The masks of the small Sulka group on the southeastern coast of New Britain were, like those of the Baining, made of ephemeral materials—in this case, narrow strips of pith bound together into a cone shape. The colour scheme of the Sulka masks is brilliant: white, black, yellow, and green designs over a bright pink background. On masks representing the human head, a swelling at the top indicates the brow, while the eyes, nose, mouth, and chin are indicated with paint or attached elements. On featureless masks, the main cone supports either another cone with a

Museum fur Volkerkunde, Basel, Switz.; photograph, Hans Hinz, Basel



Figure 23: Mask, tapa (bark) cloth. From the Baining people, northern New Britain, Papua New Guinea. In the Museum of Ethnology, Basel, Switz. Height 79.4 cm.



Figure 24: *Art from the Solomon Islands.* (Left) Inlaid shield, basketry, mother-of-pearl, clay, paint. From the central Solomon Islands. Height 83 cm. (Right) Canoe paddle, wood, paint. From Buka Island, Northern Solomons province, Papua New Guinea. Height 170.2 cm. In the Metropolitan Museum of Art, New York City.

(Left) Copyright © 1981 The Metropolitan Museum of Art, New York City, The Michael C. Rockefeller Memorial Collection, gift of Nelson A. Rockefeller, 1972 (1978.412.730), (right) copyright © 1990 The Metropolitan Museum of Art, New York City, The Michael C. Rockefeller Memorial Collection, purchase of Nelson A. Rockefeller, 1966 (1978.412.1491)

painted underside or a large, flat, painted disk. The Sulka used wood to carve female figurines; headpieces in the shape of dogs, praying mantises, or women; and convex, oval battle shields.

The western part of New Britain presents a scene of overlapping styles influenced both by other areas of the island and by nearby New Guinea. The wood-carving style of the Kilenge, for instance, was almost identical with that of the Siassi and Tami islands in its themes and patterns. Carvings on drums, a number of small objects, and, in particular, the most important type of wooden mask exemplify this affinity.

The Solomon Islands. The Solomon Islands lie southeast of New Ireland. The most important islands artistically are Buka and Bougainville in the west; Choiseul, Vella Lavella, Santa Isabel, and the New Georgia group in the centre; and Guadalcanal, Malaita (Mala Mara), Ulawa, and San Cristobal in the east. Although there are various regional and local styles, in general the art of the Solomon Islands is characterized by the predominant use of the colour black, with white and red used primarily for details; the use of borders or patterned bands of design; the use of mother-of-pearl inlay; and an emphasis on the human head.

Figures on the Solomon Islands were usually depicted in a sitting or squatting position, but on Buka Island standing figures can be found. The Buka figures are somewhat perfunctorily carved, with squared-off forms and engraving on the faces. Females have slightly flared, flat-topped coiffures, while those of males are mitrelike and pointed. The human figure was a common subject on Buka. Ceremonial houses were built with posts shaped like human figures, and forked posts carved with human heads stood in front of the houses to receive the corpses of war victims. Buka-language speakers living on the northern coast of Bougainville Island sculpted comparable figures but in a more natualistic style. These were painted glossy black, with some red and white detail representing scarification.

A constant motif in Buka-Bougainville two-dimensional

The *kokorra*

art is the *kokorra,* a silhouette of a squatting or standing human figure with upraised hands and the male mitre-like coiffure. This figure—or the head alone—was painted and carved in bas-relief upon a great variety of objects, including canoes, paddles, slit gongs, dance clubs, and architectural elements.

The central Solomon Islanders were vigorous head-hunters, building canoes of great size for their expeditions. The general model throughout the archipelago had tall, upcurved prow and stern posts, which were decorated with rows of shells and, at the waterline of the prow, a small carving (*musumusu*) of the head and arms of a guardian spirit. The *musumusu* sometimes incorporated bird characteristics. Human figures were usually depicted with the lower half of the face jutting forward boldly. Shields in the area were normally made of plain wickerwork and had a tapered oval outline. Some were decorated with shell inlay outlining elongated anthropomorphic beings and other motifs. In New Georgia, dancing human figures, rings, and other designs were carved in openwork on panels made of tridacna shell. The panels were used to enclose small shrines for ancestral skulls.

A major focus of southern Solomon culture was bonito fishing, with its symbolic relationship to sea spirits and ancestors. The roofs of canoe houses, which were the centres of male activities, were supported on huge posts carved with full-length figures of bonito, sharks, and ancestors. Model canoes and large carvings of bonito were kept in these houses, and ancestral skulls were enshrined there. Fish and animal motifs can also be found on a variety of smaller objects, including finials, finely worked *kapkaps,* and bowls. One of the most common forms of bowls, which were made in large numbers for both religious and daily use, represents a bird holding a fish in its beak and standing on a shark. Figures that combined human and animal features, such as a shark with human legs, were also common, especially on Malaita and Ulawa. Some human figures and heads featured the prognathous face characteristic of central Solomon sculpture, but the general style in the southern islands was more naturalistic in human proportions and features. Shell inlay was applied to bowls and other objects, including ceremonial clubs and dance staffs, though less lavishly than in the central Solomons.

Vanuatu. Following great disruption and depopulation in the 19th century, practically nothing remains of the cultures of northern and southern Vanuatu. The central islands of the archipelago—the large island of Malakula, Pentecost (Pentecôte) to the northeast, and Ambrym to the east—are the most significant artistically. Although this group comprises several distinct style areas, it shares certain cultural characteristics. The major socioreligious institution of Vanuatu was the graded society, a hierarchical system of ceremonies performed by men and women to attain greater prestige. The number of grades and their names, as well as the name of the system itself, varied from place to place, but in all areas the grade ceremonies included the sacrifice of pigs or boars, feast giving, and the production of physical memorials. The memorial sculpture was prescribed in form and was made in wood or from the trunks of tree fern, an easily worked material resembling compacted pine needles.

The grade system of Vanuatu

On the northeastern Malakula coast and on some small neighbouring islands, large dance grounds surrounded by stone monoliths and dolmens were set up at stages of the grade ceremonies. The stones were covered by simple shelters with the main wood post carved into human form. The head on such figures was half the height of the whole post; it was shaped like an elongated diamond and had circular eyes, a long nose, and a crescent-shaped mouth. These posts supported a ridgepole consisting of a tree trunk, the spreading roots of which were roughly shaped into the likeness of a hawk. The same image was also used for the ridgepole of ceremonial houses and on huge ceremonial banners of bark and bamboo. The bodies of these hawks are shaped like diamonds, their wings are triangular, and their heads resemble the stylized carvings of birds made for canoe prows.

The eastern coast of Malakula, Pentecost Island, and Ambrym Island together form another style area. The slit gongs of Ambrym are the largest and most elaborately carved of any in Vanuatu. As in all parts of Vanuatu where they were used, the Ambrym slit gongs were set around the dance grounds, standing upright or at a slight angle. They are found in two main forms. In the north of the island, the gongs terminate in huge three-dimensional heads with puny arms below them. The forehead and the back of the head are scored with deep cross-hatching; the eyebrows are arched, with disk-shaped eyes filling the space below them; and the concave oval facial plane is almost filled by a massive carved nose. The slit gongs of western Ambrym, on the other hand, taper upward to an even greater height and have two to three faces in the elongated-diamond style carved on them in relief. The northern Ambrym style of face is also found on the fern-wood figures made for grade ceremonies. These figures were sometimes of humans, with oversize heads, globular torsos, and slight limbs, but they could also have human heads with zoomorphic bodies. All were originally coated with clay and painted with vivid polychrome body-painting designs. Masks from the area include one type carved in hardwood that was common to both Ambrym and southern Pentecost; it has a bulging forehead, prominent cheeks, pierced eyes, a grinning mouth, and a huge, arched nose. Brilliantly coloured masks made from softwood were apparently usually comedic in intent.

The southern coastal peoples of Malakula, including some in the southern mountains, employed a wealth of objects not only for grade ceremonies but also for farming and initiation rituals. A major technique involved the use of a vegetable compound moistened with coconut milk to model heads or oversize seated figures over coconuts, bamboo frames, or other hard materials. The skulls of important men were also covered with the compound and then mounted on bodies of fern wood and bamboo adorned with beaded armlets and the tusks of sacrificed pigs. Smaller heads were sometimes fastened to the shoulders of these figures. All the modeled works were fully painted, with a great use of deep blue, white, red, and yellow. Most of the modeled heads, figures, and masks have boars' tusks protruding from their mouths. Helmet masks are janiform or have quadruple faces or are surmounted

(Left) Copyright © 1990 The Metropolitan Museum of Art, New York City. The Michael C. Rockefeller Memorial Collection, bequest of Nelson A. Rockefeller, 1979 (1979.206.1697). (right) by courtesy of Musee de l'Homme, Paris (10131)



Figure 25: *Melanesian masks.*
(Left) Helmet mask, wood (tree fern), straw, compost, paint, tusks, glass. From southern Malakula Island, Vanuatu. In the Metropolitan Musuem of Art, New York City. Height 65 cm. (Right) Mask, wood, basketry, human hair. From New Caledonia. In the Museum of Man, Paris.

with seated figures. Carving was less significant in the southern areas than in the rest of Vanuatu, and figures were somewhat flat and squat, with broad faces indicated only by a summary series of horizontal grooves.

In the mountainous northern region, sculpture appears to have been confined to large heads in fern wood and small janiform heads on spear shafts. These were carved in a powerful geometric style; the deeply cut, sharp-edged ridges and hemispheres approach total abstraction.

New Caledonia. The art of New Caledonia consists almost entirely of wood sculpture; painting was used only to accentuate, in red and white, details of the carvings, which were otherwise entirely blackened. The central feature of every village was a circular ceremonial house with low walls and a towering conical thatched roof. These houses were equipped with a wealth of architectural sculpture. On either side of the door stood a massive jamb with a stylized face at the top; the rest of the jamb was ornamented with geometric designs. The jambs were spanned by a lintel with abstract patterns, while a carved head stood between them as a sill. Inside the house, opposite the doorway, additional large figures served to demarcate special areas. The roof was crowned with a tall finial, which combined a face with stylized elements representing limbs and a headdress. Independent heads and figures of ancestors were carved on the tops of stakes planted in the ground near the ceremonial houses, at spots where their commemorative rituals had been carried out. All these architectural carvings could be claimed and carried off by the maternal relatives of the clan for which they were made; consequently, both the actual objects and their styles were widely dispersed.

Ceremonial houses of New Caledonia

The north of New Caledonia is the richest area of the island in sculpture. The masks made there are particularly striking. The long wooden faces have strongly arched eyebrows, glaring eyes (sometimes round but usually bean-shaped), and grinning, crescent-shaped, toothy mouths. The nose springs forward from between the cheeks like a great parrot beak. Other masks of the same general style have equally prominent bulging and rounded noses.

Masks from farther south are almost square; the features are flatter and angular, and the horizontal, slotted mouths are edged with red seeds. A beard of human hair hangs from the chin, and a cylindrical basket topped with a mop of human hair crowns the head. A cloak of black feathers completes the ensemble. The dancers who wore these costumes are said to have impersonated ancestral water spirits.

Similar faces and clothing can be seen on some of the figures carved onto the posts of ceremonial houses. Other figures have faces in the same styles but have domed, neckless heads; their bodies and limbs are cylindrical or ovoid and are markedly nipped in at the joints. A few small carvings show infants lying on baby carriers; they were used in magic for human fertility. Small carved heads used as a kind of currency and small heads carved on spear shafts complete the New Caledonian range of figure sculpture.

War clubs had stylized phallic, tortoise, and bird-head ornamentation. The ceremonial adz, an object unique to the island, consisted of a nephrite disk mounted on a decorated shaft; it served both as a symbol of chiefly office and as a cosmic symbol.



Figure 26: Doorjamb, wood. From Ouebia, New Caledonia. In the Museum of Ethnology, Basel, Switz. Height 1.65 m.
Holle Bildarchiv, Baden-Baden, Ger.

POLYNESIAN VISUAL ART

By comparison with the art of Melanesia, Polynesian art appears modest in scale and relatively simple in both form and imagery. This is a somewhat deceptive impression, however; the rapid adoption of Christianity in Polynesia led to the obsolescence and destruction of many types of objects now known only through the records of early explorers. The conception of Polynesian art that is gained from extant objects is therefore incomplete and must be corrected and amplified through the study of early writings and drawings. These make it clear that large-scale works were common and that Polynesian imagery was at times as dynamic and creative as Melanesian.

Limited knowledge of Polynesian art

Fiji, Tonga, and Samoa. Ethnogeographically, Fiji is usually considered to be the easternmost island group of Melanesia. As a result of trade links and direct migration in both prehistoric and historic times, however, it has much in common culturally and socially with its Polynesian neighbours to the east, Samoa and Tonga. In fact, the arts of these three island groups can be discussed together.

The endemic nature of warfare in Fiji led to the production of great numbers of wooden clubs, which were the principal weapons. At least 10 types of clubs were made, each with several subtypes. Considerable care was lavished on the engraved designs that decorated the clubs; sometimes the designs were even inlaid with whale ivory, probably by Tongan craftsmen. Fewer types of clubs were made in Tonga, but they were often completely covered in fine geometric patterning interspersed with tiny silhouettes of human or animal figures. Samoa shared a few club types with Fiji but is more notable for its wooden spears, which have graceful, elaborate barbs.

Tapa cloth was made in vast quantities and in lengths up to hundreds of yards. It was generally left plain white for daily use and decorated for special occasions. But in Fiji black, brown, and reddish dyes derived from various barks were used to decorate even everyday cloth with bold, dense geometric designs. Several techniques were used; stenciling was the most prevalent in Fiji, but freehand drawing, rubbing, and block printing were also practiced.

In historic times, pottery was made only in Fiji. Bowls carved out of wood usually had four legs in Fiji but a dozen or so in Samoa. Fijian bowls, in particular, show considerable variety of form. Large food bowls were often in the form of turtles. Small, shallow, footed dishes used by priests were usually shaped like hearts, crescents, or abstract forms, but a few resemble canoes or highly stylized humans or flying ducks.

Surviving examples of figure sculpture from all three island groups are extremely rare, and all known examples are small in size. Wooden deity figures from Tonga generally have a stocky torso, which is set well forward of protruding buttocks and short sturdy legs. The arms are stiff and straight. The head is disproportionally large; the face is flat, with the features barely indicated. The whole surface is smooth. The Tongans also carved small female figures in whale ivory; the earliest examples known are tiny, indicating a sparing use of rare material. Some small examples were strung in groups as necklaces. Later pieces, probably carved after whaling made ivory more common, are somewhat larger. A few double figures were made as suspension hooks, and some were exported to Fiji. Wooden figures from Fiji and Samoa are in much the same style, but they are less clearly defined in form.

The use of whale ivory



Figure 27: *Carved ivory ornaments.*
(Top) Breastplate, ivory, sennit. From Fiji. (Bottom) Necklace with eight human figures, sperm whale ivory, sennit. From Fiji. In the University Museum of Archaeology and Anthropology, Cambridge, Eng.

University Museum of Archaeology and Anthropology, Cambridge, Eng. (top, Z.2749; Z.2752)

The Cook Islands. The Cook Islands group includes Aitutaki, Atiu, Mitiaro, Rarotonga, and Mangaia, among others. Religious effigies are the main surviving works from the area, and both abstract and semirepresentational sculptural styles can be found on all the islands.

On Mangaia, the surface of virtually every decorated object was incised with the so-called K-motif, a dense pattern of cross-hatching interspersed with zigzags and concentric diamonds. Mangaian deities were represented by long cylindrical shafts with flared ends cut to form a group of vertical fins. The fins were pierced and carved into a series of arches, each perhaps representing a human figure with an arched back.

Similar emblems of divinity were made on Aitutaki, but in a simpler style; instead of having flared ends, the shafts were topped by flat panels that were engraved and partially pierced with geometric designs. The edges of some panels were serrated. Sacred staffs from Atiu and Mitiaro were topped by a dome flanked by pointed ovals, the lower end of the staff was spatulate or cylindrical, and the shaft itself supported vertical rows of arches. Gods in human shape were also carved on Aitutaki; they resemble Tahitian figures in posture, but their hands are more oval, their features are mere slits, their bellies droop and protrude, and their limbs are puny and square in section.

Some of the finest examples of Polynesian sculpture are from Rarotonga. Small figures of gods, originally placed on the prows of canoes, were depicted in a deep squatting stance. Their heads make up about half the total height, with the facial features reduced to simple forms—the nose is expressed only by the upper lip. The figures all have exaggerated phalli, and some were painted with black geometric designs. Rarotongan staff gods have similar facial features, but their heads are flattened into essentially two profiles. The midsection of the staff god consists of alternating full-face and profile figures. Below this is a long shaft, which is swathed in a quantity of painted tapa and terminates in a phallus.

Rarotongan staff gods

The Society Islands. The few remaining examples of traditional art from the Society Islands come mainly from Tahiti, the largest island of the group. Figure sculptures, all in wood, resemble Tongan forms in general, but they exhibit a flowing curve of the back and buttocks rather than the Tongan rigidity and protuberance. The head tends to be longer and is somewhat broader across the cheeks, the jawlines more sharply defined.

Figures of the gods took several forms; most were kept at the marae (sacred enclosure), often in special wooden containers housed in portable shelters. Hollow, life-size figures made of basketry or wood were used to hold red and yellow feathers, which were highly valued—even sacred—in these islands. Smaller divine emblems include shaped lengths of wood that were partly wrapped in braided sennit; lengths of sennit were sometimes sewn onto these figures to indicate features and limbs. Large numbers of thin, tall wooden slabs were set up on the marae; they were carved with openwork geometric designs and topped with figures of birds, human beings, or spiked projections. They closely resemble the god symbols of the Cook Islands.

Small figures were also used to decorate the prows and high, upward curving sternposts of smaller canoes. Large war canoes, which were up to 100 feet long, were equipped with towering sternposts carved with a vertical series of human figures. Carved posts of a similar model were also set up as boundary markers, as were small, crude stone figures.

As a rule, personal property and household equipment in the Society Islands were simple and unadorned, but fly whisks, which were necessary to keep off the swarms of flies that plagued and disgusted the islanders, usually had some ornamentation. The handles were generally carved of wood and were frequently topped with a single figure, which was sometimes depicted perched on one leg. A few handles were assembled from ivory segments that had been carved in openwork and tied together with sennit; these usually terminated in an arched-back ("acrobatic") human figure.

Garments worn on the Society Islands, including large ponchos, were of painted or printed tapa. In battle, men of high status wore tubular headdresses with protruding crests and gorgets decorated with bands of feathers, shark teeth, and dog hair. The most extraordinary costumes were those of mourners; they consisted of masks and aprons made of mother-of-pearl, crescent-shaped breastplates decorated with mother-of-pearl shells, and feather cloaks.

Bernice Pauahi Bishop Museum, Honolulu (1971.198.01)



Figure 28: Mourner's dress, pearl shell, turtle shell, coconut shell, feathers, bark cloth, sennit. Society Islands. In the Bernice Pauahi Bishop Museum, Honolulu. Height 220 cm.



Figure 29: Wooden image of the god A'a creating man and lesser gods. From Rurutu, Austral Islands. In the British Museum. Height 1.11 m.

By courtesy of the trustees of the British Museum

Decorative patterns used in the Austral Islands

The Austral (Tubuai) Islands. The surfaces of works from the southern Austral Islands were often incised with dense patterns of triangles, crescents, stars, and cross-hatching. The edges of such works were often notched in rows. Such lavish decoration covers carvings from Raivavae, including a few female figures with extremely summary facial features and indications of gorgets and headdresses. The same motifs cover small bowls, long-handled ladles, and broad-bladed ceremonial paddles—which exist in such numbers as to make it likely that many were made for sale soon after the arrival of European collectors. The most remarkable carving from Raivavae is found on tall and slender standing drums. The lower halves of the drums are carved in openwork, with rows of minute dancing figures alternating with rows of crescent shapes, which in some cases represent the dancers' skirts. The same repertoire of patterns was also used on tapa and to ornament wooden elements of houses.

The style of Rurutu, to the north of the group, uses the star design and chevrons but is otherwise less ornate. Some objects were traded to other islands, the most common being fly-whisk handles, which were exported to Tahiti. Each handle was topped by a pair of figures placed back to back. The shaft below was incised with chevrons or, more characteristically, consisted of a vertical series of spools. Slender spearheads were carved with miniature stylized pigs, resembling phalli. The same images, as well as stylized testicles, birds, and geometric forms, were carved in ivory and strung as necklaces exported to Mangaia. Carved wooden shafts partly covered with woven sennit were sacred objects on Rurutu, as elsewhere in Polynesia.

Perhaps the only surviving example of figure sculpture from Rurutu is one of the most impressive Polynesian sculptures: an image of the god A'a in the act of creating men and other gods. The primary figure, in Society Islands style, has 30 small stylized figures arranged symmetrically on its torso, limbs, and face, 10 being placed as the facial features. The figure has a hollow back and when found contained 24 small figures (now lost).

The Marquesas (Marquises) Islands. The most characteristic feature of Marquesas art is a strict conventionalization of the human face. It has huge eyes (circular or pointed ovals), with a continuous curved brow line that is connected to a nose shown as two small, broad semicircles; the mouth is shaped like a horizontal oblong. The design is admirably suited to works in both two and three dimensions.

Marquesan figure sculpture, in wood and stone, represented deified ancestors. The head on such a figure was typically shaped like a dome or a vertical cylinder; the almost featureless torso showed the familiar Polynesian forward arch of the back but placed no emphasis on the buttocks; the legs were ponderous, carved rather than bent, and the arms were slight, with the hands resting on the stomach. Life-size and oversize figures were kept on the platforms of the sacred enclosures. The walls of such platforms also often incorporated stone slabs with faces carved in relief. Stone figures appear to have been used as votive offerings or in fishing magic.

Other small figures in wood, usually engraved with tattoo patterns, were lashed as steps to stilt poles used in competitive races at commemorative festivals for the dead. Tiny figures in the same convention also appear on the various types of ivory ear ornament, on small bone cylinders worn in the hair or used as toggles, on the wooden or ivory handles of semicircular fans, which were plaited from coconut leaflets or pandanus leaves, or on small ivory tobacco pipes. Two types of spectacular head ornament were worn by men. One was a headband with a mother-of-pearl shell supporting an openwork tortoiseshell plaque, somewhat analogous to a Melanesian *kapkap,* the other a coronet of concave strips of white shell alternating with engraved strips of tortoiseshell.

The usual Marquesan weapons were spears and clubs, the latter having flared semicircular upper ends carved with faces and geometric designs. Decorated household objects included engraved bowls and gourds with engraved lids; stone pounders had shafts carved with human heads.

Easter Island figure sculpture

Easter Island. Besides the great prehistoric stone figures already described, the Easter Islanders in more recent times created a remarkable body of small sculpture in wood. The best-known are two types of male figures and one type of female figure, presumably all of ancestral significance. Some of the male figures are naturalistic, with an upright stance and somewhat slack and paunchy bodies; the hands, placed at the hips, have the typical elongated fingers of the stone colossi. The second, better-known type is one of the more extraordinary images in Oceanic art: it represents a bowed, skeletal figure with sunken abdomen, protruding rib cage and spine, and emaciated limbs. The face is skull-like, with a jutting nose and bared teeth. Both the naturalistic and the skeletonic male figures were worn ceremonially as pendants. In contrast to these fully three-dimensional figures, the female figures are frontal and flattened, except for the head; they have one arm placed across the torso and the other across the belly. All the figures wear goatee beards and have mythical creatures carved in low relief on the craniums. A few of both the male and the female figures have the double heads found elsewhere in Polynesia (*e.g.,* Tahiti). As in all Easter Island wooden sculpture, the glaring eyes were indicated by tiny disks of obsidian set in rings of bird or fish bone. Other carvings are of bird-headed human figures associated with the cult of a bird-god; still other human figures have the heads of lizards. Besides these there is a range of small, largely inexplicable grotesques.

Crescent-shaped wood carvings with a bearded human head on each tip were worn as pectorals; carvings of fish worn in the same manner served as emblems of authority. Women wore small spheres carved with relief designs as charms.

For commemorative funerary ceremonies, enormous bark-cloth effigies, painted with tattoo designs, were set up in front of the image platforms. Only a few miniature bark-cloth figures of this type have survived.

Long clubs with human heads carved at one end were

used for fighting. Graceful paddlelike objects, actually extremely stylized human figures, were carried in dances.

Easter Island abounds in works engraved on exposed rock surfaces, including outlines of turtles, fish, and above all the bird-headed men of the bird cult. In addition to the petroglyphs, paintings of birds, dance paddles, and other subjects exist in caves or on the interior surfaces of stone house walls.

New Zealand. It is estimated that by the beginning of the classic period of Maori art and culture about 90 percent of the population lived in the North Island of New Zealand. The smaller groups of the South Island were apparently more conservative, and it was thus largely in the north that a spectacular florescence of sculpture and architecture took place.

Maori carving styles

Two major carving styles existed. The northwestern style (especially in reliefs) featured undulating, serpentine shapes; human bodies and limbs were tubular or triangular in section and often intertwined. Heads were peaked, with gaping mouths that often had hands and arms looped through them. Large areas of surface remained smooth, while small sections and details, such as lips and eyebrows, were finely patterned. The eastern style of human figures was basically naturalistic, apart from oversize heads; the stances of figures resembled those of central Polynesia, with short legs, swaybacks, and hands on the torso, but the neck was well defined. Some faces were naturalistic portraits; others were ferociously stylized, with slanting oval eyes, distorted mouths, and outstretched tongues. Both styles of faces were usually covered with tattoo designs, while scrolls and other designs accentuated the shoulders, hips, and knees.

The northwestern style is thought to be older than the eastern, partly on the basis of oral tradition and partly because the "central-Polynesian" aspect of the eastern style suggests that it may have developed as a result of later immigration. There was a considerable mobility of local styles. Innovation, however much desired, was nevertheless restrained to some extent by the existence of a recognized repertoire of named patterns, predominantly scrolls, which sustained a continuity of style.

National Museum of New Zealand, Wellington, New Zealand (B17580)



Figure 30: Storehouse doorway, wood, red lead paint. From the Maori people, New Zealand. In the National Museum of New Zealand, Wellington. Height 92 cm.

The continual quest for prestige in Maori society encouraged men of high status to commission and own important works. The choice of such works changed throughout Maori history. It appears that war canoes were the most prestigious works in the 18th century. Communal war canoes, which were up to 100 feet long, were lavishly decorated with carving and painting. In most parts of the country the attached prow carving had a figure leaning forward with arms stretched behind it; a thin panel carved with openwork spirals ran from the figure to a transverse slab, on the back of which was yet another figure, which looked down the length of the canoe from a squatting position. The sternpost was a high vertical slat with openwork scrolls and a small figure seated at the foot. More figures were carved in relief along the hull; their eyes were sometimes inlaid with mother-of-pearl. The whole canoe was painted red, with black-and-white details, and had feather streamers hung from the sternpost.

In the first half of the 19th century, following population changes resulting from intensified tribal warfare, the introduction of firearms, and the spread of Western diseases, a number of local styles were extinguished, and, after the European suppression of fighting, the decorated storehouse came into prominence. As a precaution against vermin, these food storehouses were elevated on posts, which were often in human shape. The houses had pitched roofs and deep porches. The outer gable was fitted with bargeboards, usually carved with a complex mythological scene of several figures hauling ashore a whale (a symbol of plenty), which was represented by an abstract pattern of spirals indicating its jaws. At the peak of the gable stood a mask or small figure, and at the foot of the bargeboards were sills with reliefs of ancestors. In some large storehouses, vertical panels between the ends of the sill and the bargeboards were carved with copulating figures—another reference to fertility—while still more figures were carved on a small door and on panels at the back of the porch.

Maori meetinghouses

Largely because of the influence of the artist Raharuhi Rukupo, the meetinghouse subsequently became the central object of local tradition and pride, as it remains to this day. The meetinghouse follows the same principles of design as the storehouse, but it is built at ground level. The exterior carvings stress not the aspect of abundance but ancestral power: indeed, the whole building symbolizes the tribal founding ancestor, with the gable mask representing his face, the bargeboards his arms, and the ridgepole his spine. The porch is undecorated, except for large lintels above the door and the frame around a single window. The interior, however, is fully furnished with relief panels of frontal ancestral figures alternating with panels of laced reed in polychrome geometric patterns. The lower sections of the house posts are carved as small ancestral images. Beams and rafters are painted in prescribed red, white, and black designs.

The prevalence of warfare in the later phases of Maori history led to the construction of defensive earthworks and palisades around hilltop villages. The gates were massive planks, carved above the entrance with major ancestral figures. The palisades incorporated posts with further ancestral figures, often much bigger than life-size.

Maori material culture differed from that of the rest of Polynesia in the absence of certain types of objects (such as stools and neck rests), in the addition of others, but above all in its ornateness. Images of divinities are rare outside of architecture, except for "god sticks" (rods with heads at the upper end). Weapons included a range of short handclub types, in wood, nephrite, or whalebone, reflecting early Polynesian models. Some had human figures carved in relief near the grip (overall engraving of the blade was a late development). Staff clubs had ends carved as faces with sharp protruding tongues. Chiefs owned carved adzes with fine nephrite blades as insignia of status. Lavish relief carving covers the entire surface of small flutes, paint cups, and, above all, lidded treasure boxes.

The highly valued nephrite, classed in several grades of beauty, was employed not only for clubs and hooks but also for a variety of pendants. The best-known depict small frontal human figures in contorted poses. Cloaks, the usual garment of both sexes, were woven of flax, bor-



Figure 31: *Hawaiian featherwork.*
(Left) War god Kukaiilimoku, wicker framework covered with netting into which feathers are knotted. In the British Museum. Height 81.3 cm. (Right) Royal cloak, wicker covered with netting into which feathers are knotted. In the British Museum. Height 1.75 m, overall diameter 2.89 m.
By courtesy of the trustees of the British Museum; photograph (left) J.R. Freeman & Co. Ltd.

dered with geometric patterned bands. Feathers (pigeon and kiwi) and dog's hair ornamented others. Like other Maori arts, weaving was a highly ritualized process.

Hawaiian feather-work

The Hawaiian Islands. Like those of the Marquesas Islands and Easter Island, the artists of the Hawaiian Islands developed their own variants on Polynesian style. This resulted partly from isolation, partly from the structure of religious belief, and partly from the existence of a warlike aristocracy. Featherwork, for example, was made and used in other parts of Polynesia, but no other group produced anything as spectacular as the feather cloaks, capes, and helmets of Hawaiian chiefs. These were worn not only for important ceremonial occasions but also for actual combat. In fact, the relatively small trapezoidal capes, thought to be an early form, were designed so that the fit of the straight edge around the neck would afford some protection from attack. Larger cloaks have rounded necks and rounded lower edges with flaring sides; they were constructed of mostly red and yellow feathers, with some black and green. The feathers were attached in bunches to a netted base. The cloaks were decorated with triangles, lozenges, circles, squares, and sweeping crescents. With the cloaks, chiefs wore wicker helmets, shaped as caps with crescentic crests, which were also covered in feathers. Heads of the war god were also made of wickerwork covered with red feathers; the mouths on such heads were set with dog's teeth, and the eyes were made of large mother-of-pearl plates with wooden knobs for the pupils. Some were equipped with locks of human hair, others had crests resembling those of the helmets. These formidable images were carried into battle on long poles.

Wooden figures of divinities fall into several types and styles. The largest formed part of the settings of the sacred enclosures. Posts, carved at the top with rudimentary human forms, including heads with slanting eyes, wide mouths, and chevron-shaped brows, were parts of fences. Flat figures or faces topped with panels decorated with chevrons or silhouettes were kept inside the enclosures. Three-dimensional, full-length figures or busts on posts stood at the gate of the enclosure, at points within it, and in a semicircle facing a sacrificial altar. They were often well over life-size. The majority that survive are in the so-called Kona-district style dating from the late 18th to early 19th century; like the feathered war-god heads, they are associated with the reign of Kamehameha I (1782–1819). The bodies of the figures are massive assemblages of swelling conical or tubular segments, often sharply demarcated. The heads are proportionately large and have thrusting chins; the extremely exaggerated hori-



Figure 32: War god Kukaiilimoku, wood. From the Hawaiian Islands. In the British Museum. Height 76.8 cm.
By courtesy of the trustees of the British Museum

zontal mouths are shaped like a figure eight and are filled with menacing teeth and outlined with ridges. The heads are crowned with voluminous knobbed coiffures; plaits sweep down and back, incorporating the ridged eyes.

The conventions of Kona-style bodies occur in some earlier figures, including smaller figures of gods mounted on props, apparently personally owned by chiefs. Many differences in detail exist. Some of the figures have crested helmets, while others have elaborate tiered headdresses, triple-peaked headdresses, or none at all. Several were clearly made as pairs. Small figures without props were privately owned by families. Many of these have Kona-style bodies, helmets, crests, and other such features, but some female figures were carved in a fleshy, naturalistic style and were adorned with human hair. The small figures embodied protective gods and spirits and were used as containers for sorcery materials. Human figures adorn a number of other items, including bowls and racks for spears and poles. They are sometimes positioned with raised hands or in a headstand.

The Hawaiians made many types of personal ornament. The best-known is probably the hook-shaped whale ivory pendant, which was traditionally strung on coils of human hair. For clothing, especially for loincloths, skirts, and cloaks, the Hawaiians impressed and painted tapa with geometric designs in red and brown; the manufacturing tradition continued long after Western contact, with subsequent changes in designs and use of colour.

Polynesian influence in Melanesia and Micronesia

The Polynesian outliers. Small populations speaking Polynesian languages live within the geographic areas of Melanesia and Micronesia, on islands in the Caroline, Solomon, and Vanuatu groups. They are apparently immigrants, largely from western Polynesia, who arrived at various times in the latter 1st millennium AD. Although in many cases the culture of these groups is generally Micronesian, their art often resembles Polynesian works. Figure sculpture, for example, often exhibits the characteristic protruding buttocks. Some figures have flat faces with horizontal brows and pointed chins similar to those found in western Polynesia.

A marked feature of decorative design in much of this area is the repetition of small triangles in rows. This theme is also expressed in three dimensions as rows of pyramids or truncated pyramids. The design is found on dance paddles, canoe prows, house posts, bowls, headrests, stools, and other small objects.

MICRONESIAN VISUAL ART

Micronesia can be divided into two style areas. Western Micronesia consists of the island groups in the western Caroline archipelago, including Palau and the states of Yap and Truk (Chuuk). Eastern Micronesia includes the eastern Caroline Islands, the Marshall Islands, and Kiribati. The Mariana Islands, which lie north of the Caroline Islands, do not figure in this account, because of the obliterating effects of late 17th-century Spanish occupation on their population and its culture.

Throughout Micronesia, forms are exceedingly basic; decorative detail is largely geometric, consisting of bands of solid colour and rows of triangles in various patterns. Zoomorphs and anthropomorphs are relatively rare and are often expressed as silhouettes. The most frequent use of even this amount of decoration is found in the western Carolines, diminishing to none in eastern Micronesia.

The most frequently decorated items in the western Carolines were buildings and canoes, reflecting their importance in Micronesian society. Indeed, double-hulled canoes with crescent-shaped sterns and prows served as models for shrines constructed in Palau, Truk, and other smaller islands. Carved birds were sometimes placed in the shrines. The actual canoes of Truk were given elegant prow and stern ornaments with highly stylized sea swallows standing beak to beak. The most impressive buildings in the western region are the men's houses of Palau, which incorporate the greatest incidence of representational art in Micronesia. The gables were screened with horizontal planks that were engraved and painted with mythical and historical scenes. A life-size figure of a seated woman was often attached at the lower centre of the gable. The carved face of a god appeared at the peak of the gable. Other architectural elements, including the posts and beams, featured relief carvings similar to those on the gable. Palauan carving was otherwise restricted to large ceremonial bowls and covered boxes in avian form and to stools. The wood was treated to produce a dark red surface and was inlaid sparingly with tridacna shell in abstract patterns. Some modeling was used to decorate ceramic oil lamps.



Figure 33: Female figure from a clubhouse gable, wood. From Palau. In the Metropolitan Museum of Art, New York City. Height 65.4 cm.

Copyright © 1984 The Metropolitan Museum of Art, New York City. The Michael C. Rockefeller Memorial Collection, gift of Nelson A. Rockefeller and purchase, Nelson A. Rockefeller gift, 1970 (1978.412.1558)

In Yap the ceremonial houses were less lavishly adorned. Posts and beams were painted with silhouettes of dugongs and with black-and-white patterns of triangles or were carved with human and animal figures. Crude figures of birds and animals were hung in front of the house gables. The houses stood on platforms of coral blocks, onto which faces were sometimes carved. A trident, sometimes with a bird perched on the central prong, was a characteristic feature of Yapese art and is found on prow and stern canoe carvings, headdresses, and house decorations.

Eastern Micronesia

Eastern Micronesia produced no significant carving, apart from paddle-shaped dance staffs and plain, fan-shaped canoe prows. Material culture as a whole was simple and utilitarian. Kiribati warriors wore armour—trousers and jackets with high protective panels at the back of the neck. These suits were woven of sennit and embroidered in human hair with geometric designs.

Weaving of banana and hibiscus fibre on backstrap looms was practiced throughout the Caroline Islands, except in Palau. The fine yarns, which were dyed black, brown, and red, were woven into loincloths, sashes, skirts, and burial shrouds. Their geometric patterning paralleled the designs used in carved decorations and tattooing. In the Marshall Islands, pandanus and coconut strips were plaited into square mats worn as clothing. These were decorated with borders of checks, stripes, and rectilinear designs, which were closely related to social rank and degrees of sanctity. (D.Ne.)

Music and dance

THE ROLE OF MUSIC AND DANCE

Music and dance in Polynesia and Micronesia are audible and visual extensions of poetry, whereas in Melanesia they are aimed more at spectacular display during times of life crises and as a part of secret-society rituals. The differences between Melanesian and Polynesian-Micronesian music and dance reflect more basic differences between the Big Man sociopolitical structure found in many parts of Melanesia and the dynastic chiefdoms of Polynesia and Micronesia.

Melanesia. The leader, or Big Man, in many Melanesian societies is often a self-made man; he becomes a leader by creating followers, succeeding because he possesses skills that command respect in his society, such as oratory, bravery, gardening prowess, and magical powers. He amasses goods and has great public giveaways, often in connection with the erection of a Big Man's dwelling or a

men's house, the purchase of higher grades of rank in secret societies, the sponsorship of funeral or other religious ceremonies, or the installation and consecration of slit gongs (or slit-drums, percussion instruments made from hollowed-out logs or living tree trunks). These ceremonies occasion spectacular performances of music and dance as well as extraordinary displays of visual art.

There are basically two kinds of dance in these Melanesian ceremonies: dances of impersonation and dances of participation. In the first type, the dancer impersonates mythical or ancestral beings; the dance-actor becomes someone else, and his attire is usually distinctly unhuman or supernatural—consisting often of huge masks and a full otherworldly costume. The dance movements are dictated by the two considerations that the impersonated beings are not human and that the dancer's attire is difficult to move in. Thus, the movements are those of legs and swaying bodies; the arms are often covered and frequently used to steady the costume and mask or perhaps hold a drum to accompany the dance. The movements do not interpret recited poetry; however, the accompanying sounds of musical instruments may represent the voices of the supernatural beings.

Impersonation of otherworldly beings

The dances of participation are often extensions of these dramatic ceremonies, for individuals who do not impersonate spirits often join in and dance with them, imitating the steps of the supernatural. In dances celebrating head-hunting, warfare, funeral rites, or fertility—in which the entire community sometimes participates—the same movements are used, often to the accompaniment of drumming and communal singing. The dances have a character of spontaneity and do not require long and arduous training. Their aim is not the simultaneous flawless execution of music and intricate movements but, rather, the creation of a mass rhythmic environment that might be characterized as a visual extension of rhythm. If words are associated, they are repetitious and seem not to tell a story; they may even be unintelligible. Although the specific structure of any single dance tradition in Melanesia is not yet known, it seems probable that the isolated units of movement would be primarily those of legs and body.

Polynesia. The entirely different world of Polynesia stands in contrast. Polynesian dance is a visual extension of poetry that uses chant or heightened speech as a vehicle for the praise and honour of high-ranking chiefs or visitors. In Polynesia, power resides in chiefly office, and traditional oral texts tell of a chief's deeds and his descent from the gods. Genealogical rank is a distinctive feature of Polynesian societies, and music and dance pay allegiance to the rank-based sociopolitical structure, reflecting and validating the system of social distinctions and interpersonal relationships. In these societies, where power resides in the office and the regime is long and enduring, specialists compose poetry, add music and movement, and rehearse the performers for many months before a public ceremony. Movements are primarily those of hands and arms, and interpretation is that of a storyteller. The dancers do not become characters in a drama, and their stylized gestures do not correspond to words or ideas as they do in literature-inspired dance traditions of Indonesia and Southeast Asia. In Polynesia the dancer interprets a story orally, usually chanting or reciting metred poetry, and accompanies the words with actions. Yet, although Polynesian dance texts are based on traditional stories, legends, or myths, a story is not "told" in the usual sense: traditional literature is referred to in a roundabout way, but the poetry is often the vehicle for saying something else, usually something relevant to the occasion at which it is presented. In addition, the order of the dances and the choice and placement of the dancers often supply further information about the social structure.

The structure is known for at least three Polynesian dance traditions—Tongan, Tahitian, and Hawaiian—and the basic units of movement are primarily those of the arms. The only Polynesian dance tradition, however, that has been thoroughly studied is the Tongan. Tongan dance is a visual extension of poetry and is closely intertwined with social organization. This sung poetry is a series of references to mythology, chiefly genealogies, famous scenic places, and contemporary events. The dances, which are performed either standing or sitting, interpret selected words of the text with hand and arm movements. The distinguishing characteristics of Tongan dance are the emphasis on the rotation of the lower arm and the flexion and extension of the wrist, as well as a quick sideward tilt of the head. The legs are used mainly to keep time with sideward stepping movements, and there is a marked absence of hip or torso movement. In pre-European times an important dance was the *me'etu'upaki*—a paddle dance performed by a large group of men in accompaniment to singing and a slit gong, which was often played by a high-ranking chief. This dance is still performed today. Group dances called *me'elaufola* were performed by men or women separately in accompaniment to singing, long bamboo stamping tubes, and percussion sticks. An evolved form of this dance, which flourishes today, the *lakalaka*, is performed by men and women together in accompaniment to sung poetry only. Solo and small group dances performed by one, four, or eight women often follow the large group dances and are more concerned with beautiful movements than with interpretation of poetry, although the same movements are used. In the 20th century, Polynesian dances can be classified into six genres, three of which have survived from pre-European times. The most acculturated dance type, the *tau'olunga*, is a combination of Tongan and Samoan movements accompanied by Western-

Tongan dance movements

(Left) © Axel Poignant; (right) Kal Muller—Woodfin Camp





Figure 34: (Left) Dance of participation in which the "Mudmen" of the Asaro River area, Papua New Guinea, reenact their ancestors' defeat of their enemies. (Right) Dance of spirit impersonation, Ambrym Island, Vanuatu.

style singing in conjunction with stringed instruments.

Micronesia. Music and dance in Micronesia, though certainly not the same as their Polynesian counterparts, are closely related to them. With the exception of Truk in the central Carolines, which displays traits of Melanesian and possibly Indonesian influence, the music structure of all parts of Micronesia is predominantly word-determined, as is that of Polynesia. Dance movements are mainly of hands and arms in accompaniment to poetry. In some islands, such as Yap (in the western Carolines) and Kiribati, there is a similar concern for rank in the placement of dancers, as well as the emphasis on rehearsed execution of songs and movements. But, although movements and types of dance have a superficial similarity to those of Polynesia, there are differences. In the Yap empire, for example, dancers from Ulithi, Woleai, and other islands performed and taught their choreography and texts to the Yapese as tribute, even though the dance texts were in languages unintelligible to the Yapese dancers; the function of movements was not to illustrate a story but to decorate it. Instead of acknowledging a chief's deed or genealogy, the Yapese dancers demonstrated the overlordship of Yap to the other islands. Even in Ifalik, where texts were in their own language, the movements did not interpret poetry but were apparently abstractly decorative. The same is true for the Kiribati. Thus, Polynesian dance could be characterized as illustration of poetry and Micronesian dance as decoration of poetry, while music in both areas serves as an elevated form of audible performance for poetry.

Dance as ornament to poetry

In many parts of Micronesia, dance and music were associated with tattooing, and with the decline of tattooing has come the virtual demise of these genres. The importance and dependence of the Micronesians on the sea is illustrated in poetry, music, and dance. In some areas, dances were performed on a platform of canoes; canoelike paddles were used in other dances, and, in some areas, performers wore head ornaments modeled after canoe parts.

Again, the specific structure of any of the Micronesian dance traditions is not known, but apparently the basic units are primarily those of hands and arms and, if early descriptions are to be believed, the head.

Australia. Music and dance of Australian Aborigines are important elements of sacred ceremonies that reenact mythological origins of the tribes and ensure the continued supply of foods through the propitiation of totemic plants and animals. Little is known about the internal structure or basic movements of the various Aboriginal dance traditions; however, in general terms there are often mimetic movements that involve the entire body to add a visual interpretative extension to the oral tradition of the tribe.

MUSICAL INSTRUMENTS

Oceanic cultures have developed a large variety of sound-producing instruments. Some are unique, such as the friction blocks of New Ireland: three to four plaques carved out of a wooden block are rubbed with the hands to produce shrieking or hollow-resonant sounds, depending on size (8 to 80 inches for the entire instrument). Many instruments are used not in musical contexts but for other purposes—for example, to produce the voices of supernatural beings (in Melanesia), as lures (shark rattles), as toys, and for communication.

Musical instruments proper generally lack provision for musical efficiency and easy handling. Flutes have no or few finger holes and no air ducts. Tuning devices with drums are rare, as are fixed resonators with chordophones (stringed instruments), which are represented only by simple types of musical bows and zithers. In general, instrumental music is culturally less important than vocal music, and in some areas it is absent altogether. In some other areas, however, such as the Solomon Islands, there are highly developed pan-flute orchestras.

Although some types of instruments—*e.g.*, conch trumpets and slit-drums—can be found in many parts of Oceania (excepting Australia), others occur only locally or are distinguishing features of certain musical areas. Open, hourglass-shaped drums with one membrane are typical of New Guinea and Melanesia, while Polynesians use drums of cylindrical shape that are technically kettle-drums. Flutes of various types are usually blown with the mouth in Melanesia, with the nose in Polynesia (nose flutes), and both ways in western Micronesia. In contrast to their simple technical structure, which is the more demanding on the skills of the player, some instruments display elaborate ornamentation related to their function as cult objects in Melanesia, or they may be highly carved and finished befitting their function of honouring gods and chiefs in Polynesia.

REGIONAL STYLES AND TRADITIONS

Melanesia. Melanesia, including New Guinea, houses a multitude of regional musical styles, few of which have been thoroughly investigated. The diversity, which parallels the linguistic situation, is assumed to be a result both of migrations and of the relative isolation of ethnic groups due to geographic conditions. Intergroup contacts, including European influences, mostly since the late 19th century, have had a clear influence on present styles.

Generally speaking, Melanesian music tends to be less word-determined than Polynesian and Micronesian music. Melody, rhythm, and form appear to be shaped by intrinsically musical principles rather than by the structure and meaning of accompanying text. In fact, song texts in Melanesia often use archaic or foreign languages and are thus frequently unintelligible to all participants in a performance.

Melodic characteristics. Musical scale and melodic movement have been primary criteria in the Western

(Left) Burt Glinn—Magnum; (right) © Axel Poignant



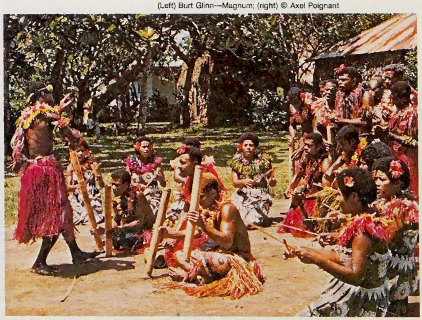


Figure 35: (Left) Festival dance, Yap, Caroline Islands. (Right) Welcome dance, performed as a mark of respect to visitors, Viti Levu Island, Fiji.

The triadic and pentatonic forms

analysis of Melanesian music. The main types of melodic form are triadic, in which the melody moves exclusively or predominantly on the steps of a triad (three tones, each a third apart, as C-E-G); and pentatonic, which uses five steps within an octave, the melodic structure typically emphasizing seconds, fourths, and fifths. Other types include "narrow," in which melodic movement is restricted to an ambit (range) of a third; and "tiled," in which the melody consists of a sequence of short narrow phrases on different tonal levels, always in a descending order.

Several attempts have been made to link these types of melodic form with specific cultures within Melanesia. In the mid-20th century the Dutch scholar Jaap Kunst attributed the tiled type found in the interior of western New Guinea, the Torres Strait, and Australia to "a people who, without doubt, emigrated from Asia to Australia—where the majority of them finally settled—by way of New Guinea and Torres Strait." Triadic melody style has been connected with speakers of non-Austronesian (Papuan) languages in New Guinea and elsewhere in Melanesia, while pentatonic structure is described as an element of Austronesian culture in Melanesia. But, whenever new data become available, previous hypotheses on the distribution of stylistic characteristics and their attribution to ethnic and cultural groupings usually have to be revised.

Also, rarely is the music of an ethnic group found to be based on one single principle of melodic structure. In most cases a mixture of several principles is apparent, with one or the other prevailing.

Musical style and cultural context. This mixture of musical structures holds true for the few New Guinea groups whose music has been studied in its cultural context: the Monumbo, the Kate, the Watut, and the Kaluli. A more detailed discussion of Kate music illustrates the stylistic heterogeneity of the Kate, who live in the hinterland of the Huon Peninsula of northeastern Papua New Guinea and speak a non-Austronesian (Papuan) language, while some of their neighbours on the coast and on adjacent islands speak Austronesian (Melanesian) languages. A Lutheran mission was established in that area in 1886.

Music of the Kate people of New Guinea

Before the mission terminated their non-Christian religious activities, the Kate shared with their neighbours, specifically the Melanesian-speaking Jabem, Bukawa, and Tami, a secret initiation cult that provided for an exchange of music and dances among participants. The mission introduced "Christian songs" with texts in Kate language and European church tunes; but missionaries also created "Christian" adaptations of traditional Kate melodies, which were more readily acceptable than Lutheran hymns.

By about 1910 the Kate experienced a twofold cultural change resulting from the continuing contact with their Melanesian-speaking neighbours and from the impact of European colonial culture. But many aspects of their precolonial culture were then still functioning or in fresh memory. Music and dance activities were connected with childbirth, children's games, initiation, hunting, agriculture, ceremonial bartering of pigs, warfare, and death—the latter occasion being the only one to prohibit dancing. Consequently, initiation ceremonies, which usually extended over two years, had to be interrupted whenever a death occurred. In addition to music and dance in ceremonial context, there were songs for entertainment and expression of individual sentiments or experiences. Most of the common social dances and dance songs were adopted from the offcoast Siassi Islands, including texts that were unintelligible to the Kate.

Structural characteristics of Kate music include triadic and pentatonic melody, both pure and in various degrees of blending; monophony (music having a single voice part); reverting and progressive strophes (stanzas) of varying lengths; basically isorhythmic organization (*i.e.,* using recurring rhythmic patterns); and a wide range of tempi. Analysis of stylistic characteristics in relation to context and historical data revealed that triadic melody and short progressive strophes are associated with the old, non-Austronesian (Papuan) stratum of Kate culture. Pentatonic melody and longer, reverting strophes seem to represent Melanesian influence, while a specific melody style characterized by the use of several pentatonic modes in succession ("modulating pentatony"), very wide ambit, and extended strophic form can be attributed to European culture contact.

For the Kaluli, a group of rain-forest dwellers in the Southern Highlands province of Papua New Guinea, the American anthropologist Steven Feld has demonstrated the integration of diverse musical structures and natural sounds under one aesthetic ideology. The concept of "lift-up-over sounding," which calls for a continuity of overlapping sound qualities and the avoidance of unison, governs all Kaluli musical expression, including recently acquired Christian church tunes and pan-New Guinea popular music, and thus projects Kaluli ethnic identity.

The Solomon Islands. While the music of New Guinea and western Melanesia—particularly the Bismarck Archipelago—is predominantly vocal and monophonic, the music of the Solomon Islands is largely determined by use of highly developed panpipes. These instruments have three to nine closed tubes, usually doubled by open tubes that sound the higher octave. New instruments are tuned by comparison with old "masters," at the occasion of special ceremonies. Panpipes are used in orchestras alone and in conjunction with song. Vocal melodic style, which is characterized by triadic structure, wide melodic leaps, elaborate polyphony (several simultaneous voice parts), a specific timbre, or tone colour, and frequent change of register, is apparently an imitation of panpipe music. Types of polyphony include parallel melodic movement a third or a sixth apart, a practice most likely a result of overblowing the double panpipes, which may produce such parallels automatically.

Panpipe ensembles

The 'Are'are people of Malaita, in the Solomon Islands, distinguish four types of panpipe ensemble and more than 20 musical types. The music of panpipe ensembles enjoys the highest prestige among the 'Are'are. They have an extensive system of thought about music that centres on *'au,* their word for bamboo, the material of which panpipes are made. Their system of musical concepts is so far not written, however; instead, it is passed on orally among specialists as a music theory that is in some respects homologous to European music theory.

The triadic melody style of the Solomon Islanders seems to have spread into adjacent western Polynesia, where similar melodic types are found in Tuvalu and Futuna.

Micronesia. From the Carolines in the west to Kiribati in the east, most traditional music is accompanied by dancing in standing or sitting posture. Group singing with rhythmic accompaniment by body and ground beats or concussion sticks is the prevailing type of musical performance. Purely instrumental music was performed on nose and mouth flutes in the Carolines and Marianas.

Except in the central Carolines (Truk), where musical influences from Melanesia and eastern Indonesia are prominent, elements of chanting and metred declamation are the most conspicuous characteristics of musical structure, underlining the importance of poetry versus intrinsically musical principles. Vocal polyphony takes the forms of drone (sustained note heard against a melody) and parallel movement in a variety of intervals, with fourths most common.

Micronesia is the least-known part of Oceania, as far as music and dance are concerned. On some of the atolls in Kiribati, consecrations of assembly halls, races of boat models, and meetings between local groups were or still are connected with performances of dances in which both women and men participated. Active participation and choreographic role are determined by individual proficiency as dancer and singer and by social rank. The *ruoia* is a sequence of standing dances in which movements are slow and mainly those of the arms and hands. In introductory and main dances, up to six leading dancers, male or female, pose as "gliding frigate birds" in front of the other dancers, who are lined up according to status in their patriclans (kinship groups). All dancers participate in chanting long poems that are rehearsed beforehand. Endings of dance songs are frequently shouted, and texts of the final dance in a sequence are recited in heightened speech throughout. There are also sitting dances, with arm and hand movements similar to those of the standing

Importance of social status in Micronesian dance

dances, and stick dances. Dance gestures are not illustrative of the song texts, which are not generally understood by performers or audience.

The texts of traditional dance songs were "received" by composers from ancestor spirits (*anti*) in special rituals and probably in trance. Since the early 20th century, multipart singing of European church tunes has spread throughout the area. In consequence of a general culture change, social dances based on traditional movement patterns but accompanied by adaptations of Western music have become dominant.

Polynesia. Since the second half of the 18th century, all parts of Polynesia have undergone a drastic cultural change that affected music and dance traditions severely. At present, adaptations of Western musical forms are predominant throughout the area. Both European church singing introduced by various missions and secular styles, ranging from the whalers' chanteys to modern international entertainment music, have participated in this process. Yet, remnants or elements of precontact Polynesian music have survived almost everywhere, either alongside of or within acculturated styles—because of a marked traditionalism, as in Hawaii and New Zealand, or because of a delayed acculturation process, as in many of the smaller and remoter islands.

Common traits. The first useful descriptions of Polynesian music and dance come from Captain James Cook and his companions on his exploration voyages (1768–80). Such reports of early travelers agree with 20th-century research in suggesting that, despite regional variance, the concepts and structural characteristics of music and dance are highly similar throughout Polynesia. Music serves as a vehicle for Polynesian poetry, as dance is its illustration. The central role of the word explains why Polynesian music is primarily vocal. The only noteworthy traditional instruments used independently from song are the nose flute and the musical bow. Accompaniment of song includes body percussion (*e.g.,* slaps, claps), drums, and various idiophones (instruments the bodies of which vibrate to produce sound, such as rattles and slit-drums).

Emphasis on words in Polynesian songs

The most obvious stylistic characteristics of songs, common to all parts of Polynesia, result from word orientation. Most traditional songs can be classified as chant, as recitation in heightened speech, or as a blending of both. In some areas—for example, Tuvalu—the same text may be performed in either style. Chanting uses a limited number of tone levels, mostly a third or a fourth apart, and numerous tone repetitions. Rhythmic organization varies from recitative bound to the accents of the words to strict repetition of rhythmic patterns. Vocal polyphony is widespread throughout Polynesia and undisputedly indigenous. It was described by 18th-century explorers when foreign influence could hardly have affected the style of music. The most common form of polyphony, reported from almost all Polynesian groups with the marked exception of the Maori (of New Zealand) and Hawaiians, is drone produced by a second part that follows the melody rhythmically while repeating one—usually the basic—tone. All authorities agree that pure drone is a precontact element of Polynesian music. Other forms of polyphony occur locally and are believed to be a result of European influence; especially notable among these is imitational counterpoint (simultaneous, interwoven melodies using melodic imitation among the voice parts).

Beyond these basic common traits, Polynesian peoples have developed musical forms and stylistic characteristics that are distinctive for individual groups of islands.

Hawaii. What is generally known as "Hawaiian music" is the result of the acculturation that began in the early 19th century and that was greatly enhanced by the introduction (*c.* 1820) of Christian hymn tunes. The ukulele, so closely connected with this almost entirely Western style of singing, is a local version of the Portuguese *bragha,* a small guitar imported to Hawaii about 1879. The Hawaiian, or steel, guitar is a metal-stringed adaptation of the European instrument that is played by stopping the strings with a metal bar.

Despite the predominance of Western—and, more recently, Asian—influences, some evolved forms of precontact Hawaiian music and dance have been preserved. Their stylistic characteristics fall well within the limits of what has been described as common Polynesian elements.

The Society Islands. The inhabitants of the Society Islands, whose ritual and profane dances accompanied by polyphonic chanting, nose flute, and drum playing were admired by 18th-century explorers, experienced a particularly rapid and thorough Westernization of their music. Still, the modern Tahitian *himene,* contrapuntal compositions in as many as six voices, retain some indigenous elements of music structure derived from polyphonic chant. The term *himene* is derived from "hymn," but the style includes both traditional Christian hymn texts and traditional Polynesian texts. The *himene* represent a highly developed form of a hybrid Polynesian-European music style.

The Tahitian *himene*

The Maori. The Maori of New Zealand have lost most of their instrumental music in the process of acculturation but have preserved many of their traditional chants and dances, which are classified according to function and contents of the text. Among the more prominent types are the lullabies (*oriori*), the laments (*tangi*), the incantations (*karakia*), the love songs (*waiata aroha*), the historical or genealogical recitations (*patere*), and the dance songs (*haka*). They are either recited in heightened speech or sung on narrow melodic lines undulating around a central tone, *oro.* Rhythm is largely word-bound. Any polyphony is considered a fault of performance. One important aesthetic concept requires a performance to be uninterrupted even by breaks for breathing. Consequently, chants are usually performed by two or more singers who take breaths at different moments. As in all of Polynesia, the younger generations favour adaptations of Western music.

Western archipelagoes. The musical traditions of western Polynesia are better known than those of any other part of Oceania. Descriptive monographs are available for Samoa, Tonga, Bellona Island (a Polynesian outlier in the Solomon Islands), Tokelau, Wallis and Futuna, and Tuvalu. There is a considerable degree of stylistic and terminological coherence that characterizes western Polynesia as a distinct musical province within Polynesia.

Before Western contact, music in Tuvalu was closely connected with social rank, religion, and magic. There are no detailed descriptions of dances; vocal styles included recitation in heightened speech and chant with drone polyphony (common to most of Polynesia) and triadic melodies resembling those of the Solomons. The Samoan emissaries of the London Missionary Society who converted the people of Tuvalu to Christianity (1861–76) destroyed the traditional social hierarchy and suppressed dances and songs either related to non-Christian beliefs or simply not fitting for their concepts of morality. They introduced pentatonic Christian songs characterized by two-part contrapuntal polyphony resulting from overlapping antiphony (contrasting groups of singers). This "pentatonic antiphony" is believed by some authorities to have developed in Samoa under European influence. By 1900, it seems to have become the predominant musical style in Tuvalu for both religious and secular topics. From 1914, church hymns and school songs in four-part European harmony began to replace "pentatonic antiphony" as the favourite style. By 1960, four-part harmony was the almost exclusive style of church, school, and dance tunes. International "Hawaiian" music is gradually penetrating the islands as mass media and Western musical instruments such as guitars and ukuleles become available. Remnants of the earlier traditions persist only with members of the older generations, although outside interest has stimulated a modest revival movement. (Di.C./A.Ka.)

Literature

THE ROLE OF THE AUTHOR

Mana

Creativity in many parts of the Pacific is perceived in terms of mana, a word of Pacific origin that is commonly used in Melanesian and Polynesian languages to express the power or force that is concentrated in objects or persons. This includes creative energy, which emanates from supernatural sources. Even though creativity has divine

origins, mana is never exclusively the possession of supernatural beings. It is available to chiefs because of their closeness to supernatural powers, and ordinary people can also increase their creativity by performing appropriate rituals or by studying and acquiring skills. Outstanding creative individuals have mana in greater abundance than ordinary individuals, but every individual has the potential for some form of creativity.

Each individual may express his creativity in different ways, however, and three broad categories of literary artists can be described. In Polynesia the traditional artist was the priest-poet, who served both as a priest or magician and as an entertainer. The priest-poets were in the employment of chiefs and had well-defined functions. They sometimes formed a distinct hereditary caste; in other cases they were drawn from chiefly clans and were occasionally even high chiefs themselves. The priest-poets became obsolete as the traditional chiefdoms gave way to Western colonization and nationhood, but in most parts of the Pacific there are still what has been called "free-lance," or "unattached," practitioners. These include wandering poets and minstrels who, unlike the priest-poets, do not go through any formal rigid training and do not claim to have special mana. They learn their trade by listening to other artists. Sometimes their function is socially recognized and they receive payment for their performances. As in most cultures, there are also amateur artists, including occasional versifiers, choral singers, and individuals such as grandmothers who tell stories and thus contribute to oral literature.

Types of oral texts

This classification of artists as professional, free-lance, or amateur emphasizes the significance of individual talent in both oral and written literature. Because the names of individual composers or poets do not feature prominently in oral literature, it has led to the popular belief that oral literature is communally created. There are, of course, instances of communal or group involvement in the creation of certain kinds of oral literature; at the centre of all creative activity, however, is a chief composer who bears ultimate responsibility for the creation. In *Return to the Islands* (1957), Sir Arthur Grimble vividly relates how oral poems were composed in Kiribati. He describes the first stirring of poetry as a "divine spark of inspiration," which gives the poet his mana. This mana, in turn, causes the poet to remove himself from society into his " 'house of song,' wherein he will sit in travail with a poem that is yet unborn." After the poem is formed in the poet's mind, he calls together some of his close friends and recites the "rough draft" of his poem. Grimble observes,

> It is the business of his friends to interrupt, criticise, interject suggestions, applaud, or howl down, according to their taste. Very often, they do howl him down, too, for they are themselves poets. On the other hand, if the poem, in their opinion, shows beauty, they are indefatigable in abetting its perfection. . . . When all their wit and wisdom has been poured out upon him, they depart. He remains alone again—probably for several days—to reflect upon their advice, accept, reject, accommodate, improve, as his genius dictates.

The status of written literature

Although the author of a written work is easier to identify and recognize, the role of such writers in Oceania is more difficult to define. Written literature depends greatly on readership, and in much of Oceania the writer's language (usually English) and major genres (*e.g.*, novel, short story) are nonindigenous to the region. Consequently, written literature from the Pacific is better known internationally than in Oceania itself. The level of literacy in English—especially in Papua New Guinea, where about 700 traditional languages are spoken—is not high. Interest in local literatures is generally confined to the emerging educated elite in the urban centres. For the majority of people throughout Oceania, traditional oral literatures, such as storytelling and dance and musical performances, still satisfy their aesthetic needs. (S.)

ORAL TRADITIONS

Understanding of the content of Oceanic oral literature is limited. Nevertheless, a few general comments about traditional forms, types, and themes can be made. First, because the purpose of literature is to communicate, it demands an audience. In the case of an oral literature, communication depends first on memory, and this usually means that such memory aids as rhythm and stock formulas and phrases are an important element of all texts. The majority of Oceanic texts keep closely to traditional forms and appear to be committed to memory and then are communicated in a strictly unvarying manner. This is, however, only approximately the case, because the various techniques of formalization can allow for a rather fluid text. The tradition can be made evident at the lexical level, with the possibility of a great freedom of syntax. It only prevents prosodic elements from taking on primary importance.

The literary occasions of the Oceanian peoples are, as in other cultures, reflected in sacred literature, political literature, and frivolous—even erotic—texts. This division, however, should not be taken to represent an attempt at classification; any such pigeonholing would be inconvenient indeed, because such a large number of texts defy neatness and straddle two categories. But there are certainly two poles between which the various forms of literary expression can be placed. On the one hand, there is a body of works that appeals to Western readers and is made accessible to them by its use of the poetic image. On the other hand, there are many texts, often brief, in which each word is frequently a complete image. This kind of text is part and parcel of the culture that has produced it and requires a veritable arsenal of commentaries of others to interpret the key words and to unravel its significance to nonnative readers. Texts are, basically, of two kinds: (1) recitatives, whose form is rigid; they can be expanded but not transformed; in this category belong all the songs or chants that accompany dances (whether the performers be standing or seated), funeral chants, songs that accompany children's games, and those with an erotic significance; and (2) public orations, in which the elements are formally but roughly organized, giving the speaker the right to vary the presentation within certain limits established for this literary genre. Such discourses, which can aptly be delivered as a high-level political oration and as a funeral eulogy or remembrance speech, can also, in a more simple form, commemorate such events as a birth or a marriage.

Oceanic themes are those that appear in other literatures of the world: love and death, defiance and hatred, nostalgia for the past, and the pleasure of the moment. Nature provides the necessary images. There is, nevertheless, a barrier between Oceanic literature and that of other cultures. Although it presents a familiar mental universe, it does so in what is often an allusive manner that demands an intimate knowledge of local place-names, local political geography, and land division before its meaning can be understood: the owl, for example, is symbolic of a given place, the lizard of some other, and the sea eagle associated with a third.

Melanesia. Melanesian oral literatures are better known to contemporary scholarship than are Polynesian because over a longer period they remained untampered with by Western anthropologists and scholars and were, therefore, not distorted by them. Because of this, they are of great interest, although they are only just being recovered. Researchers have amassed a body of astonishingly rich oral literature and have been able to establish just what the function of each text was in the life of the society they studied. The results of their work have obliged scholars to reexamine their conceptions of the oral literary tradition. It now emerges that not even words, symbols, or places have any fixed significance: the vocabulary is at once coherent and diversified, for it is used in a way that takes into account the momentary interest of the parties present.

Audience response

The traditional speech. A traditional speech may be "hurled" by a speaker to a crowd, which typically responds with a muffled cry at the end of each "sentence." It is the quality of accord reached between the crowd and the speaker that comprises the "new content" of the familiar, traditional speech (henceforth becoming part of its tradition), while the accord also conveys what other cultures would call the "message"—the new stage in a rising political career, perhaps, or a declaration denoting peace between combatants or the beginning of a war. Anyone analyzing the speech itself has to search for and consider carefully all the possible interpretations and temporal con-

ditions it might have. The traditional form of the speech is respected all the more because it allows this variation of content—within acceptable limits—to be conveyed by nuances that may easily escape observers from other cultures.

The form of the speech also presents a somewhat thorny problem. On the surface, it is a simple enumeration of the local groups and their symbols (that is, the portion of the land that is theirs, together with its animals, plants, weather conditions, and so on). The recitation of these physical realities is an affirmation of their very society, and they are stated one after the other, linked by some stock connecting phrase. Each listed item is given its precise geographic location; of special importance are the names of special places where authority is exercised or where rites are practiced. The whole constitutes a world vision, or system, in which the individual society and its members have their place. The native audience is perfectly familiar with the spatial affirmation of their society contained in the speech, and, from the dry enumeration of its components, they are able to supply for themselves a history of alliances and wars and to remodel the traditional text until it fits the conditions of the present. The orator's delivery, the nuances he maintains or introduces into the speech, are a sign of his success.

The myth. Another kind of oral literature, less easily defined, is the one commonly understood as myth (a term borrowed from Western classical culture). Involved here are all formalized recitatives, straightforwardly delivered, whose apparent aim is to give the account of a rite, for example, or to describe the situation of some group of family lineage (perhaps in connection with land ownership or political status) or to tell of the origin of man and his culture. The recitative may involve as little as 10 lines, or it may take three hours to narrate. In any given local culture, the rules that govern the way in which the texts' content is formalized and those that govern the way in which they are recited are consistent. The symbolic vocabulary, formally identical with that used in public speeches, carries elaborate but acknowledged references. A text may be established on the basis of a single symbol, but, in general, the symbolic pattern is so complex that other cultures have great difficulty in understanding it. Indeed, it is only possible to decipher its meaning if the cultural significance of every place mentioned in the text is understood, if it is known what creature or being is worshiped where and to the benefit of which group, and often only if the itineraries that are the subject of a majority of the myths are entirely familiar to the interpreter, so that old place-names—and thus their meaning—that are referred to in the text can be identified.

Polynesia and Micronesia. Polynesian and Micronesian oral literatures are similar in structure, but they vary in detail from island to island. Little is known about the Micronesian texts, however, and work on the interpretation of Polynesian oral literature has been hindered. Authentic oral texts were not collected when the opportunity was at hand, and it is now often too late to do so because the traditions—especially of religion—that maintained them have not survived. The few studies of Polynesian and Micronesian oral literatures that have been published have generally analyzed simplistic summaries rather than authentic texts. Nevertheless, they give a glimpse of how widespread certain legendary cycles were. The most frequently recurring legend, from a geographic point of view, is that of the trickster figure Maui-tiki-tiki, who was a fisherman of the islands and who discovered fire. He can be recognized, on the fringes of the Polynesian area, as the god of the first fruits of the yam harvest. He was sometimes revered under a symbolic manifestation or sometimes as a less abstract figure.

The trickster figure in legend

The few scraps of knowledge available about the indigenous literatures of Australia and Micronesia, as well as those of Polynesia, indicate that the figures of their great mythological cycles were simultaneously general symbols and local divinities. The people saw no contradiction in this double manifestation: great cultural heroes were naturally assigned to a specific place when an individual within that culture would establish a reverential dialogue with any one of them. (J.Gt./Ed.)

MODERN LITERATURE

Development of written literature. Although literacy is a precondition for the development of written literature, the relationship between literacy and literary growth is not simple. In many areas of Oceania, literacy in the vernacular languages existed for more than 100 years before significant literature in the local languages appeared. The slow emergence of written literature can be explained by such factors as the strength of traditional oral literatures, the lack of serious oppression (which might have stimulated nationalistic or protest writings), limited resources, a condition of dependency, and the indifference of missionaries and the ruling elites toward indigenous literatures.

The growth of written literature did not begin in earnest until after the establishment of the University of Papua New Guinea (1965) and the University of the South Pacific (1968). The most significant works have been written in English and have come from the regions served by the two universities (Papua New Guinea and the Cook Islands, Fiji, Kiribati, Nauru, Niue, the Solomon Islands, Tonga, Tuvalu, Vanuatu, and Western Samoa). This can be attributed to the combination of forces that converged in these regions in the 1960s and '70s: the agitation of political independence, the growth of regional consciousness, the establishment of universities, a growing awareness of other Third World writing, the use of Pacific-oriented curricula in schools, an increased confidence in the English language and its use as a language of creative expression, the widespread growth of English-language newspapers, the emergence of small commercial printers, an overall improvement of material condition throughout the area, the emergence of educated elites, and greater democratization and Westernization of social and political institutions.

Although some of these factors were also present in parts of Micronesia, they provided little or no impetus to literary development. By the late 20th century, Micronesia still could not claim to have a written literature of its own. In French Polynesia and New Caledonia, some indigenous writing in French has developed, but with remarkably different variables: the main orientation is not toward the other Pacific states but toward France. Easter Island, separated from other Pacific island states by distance, has remained relatively isolated from recent cultural developments in the Pacific. In this Spanish-speaking Polynesian island, the beginning of indigenous writing is focused on biographies and histories.

The first creative efforts of many indigenous writers were published in *Unispac,* the University of the South Pacific student newspaper. In 1973 several of the writers featured in the newspaper formed the South Pacific Creative Arts Society, which then established the literary magazine *Mana.* At about the same time, Ulli Beier, who had already played a significant role in the development of Nigerian literature, established *Kovave,* a journal of New Guinea writing, and *Gigibori,* a journal focusing on New Guinea cultures. Since then the number of literary journals in Oceania has proliferated and includes *Bikmaus* (formerly *Papua New Guinea Writing*) and *Ondobondo* in Papua New Guinea and *Moana, Faikava,* and *Sinnet* elsewhere in the South Pacific. Although the most significant writing has been in English, there are signs of a renewed interest in writing in local languages. The Samoan journal *Moana* is devoted to writing in the Samoan language, the bilingual *Faikava* includes writing in Tongan and English, and *Sinnet* has attempted to publish writings in English, Hindi, and Fijian. In Papua New Guinea and to a lesser extent in the Solomon Islands, some writers have experimented with Melanesian Pidgin, but no writing of distinction has appeared in that language.

Literary journals

Early writings. The earliest written works by Pacific Islanders were mission-inspired autobiographical writings in the vernacular languages. No fictional or imaginative writing emerged during this early period. The first published work with some literary merit was Florence Frisbie's autobiography, *Miss Ulysses from Puka-Puka* (1948), originally written in English, Rarotongan, and Pukapukan.

The first published novel from Oceania was *Makutu* (1960) by Thomas Davis, a Cook Islander, and Lydia Henderson, his New Zealand-born wife. Like their earlier

autobiography, *Doctor to the Islands* (1954), it was written in English. The novel, which deals with the cultural conflict between Pacific and Western values in an imaginary land called Fenua Lei, has more in common with the "South Seas" literature of Stevenson, Maugham, and other European writers than with modern Oceanic writing. Although *Makutu* has few of the stereotypes of South Seas literature, it is nonetheless written with the European reader in mind.

Later writings. In the 1970s, reacting against the distortions in the European vision of the Pacific, writers such as Albert Wendt of Western Samoa argued for a literature written by Pacific Islanders. In Wendt's novella *Flying Fox in a Freedom Tree* (1974), the protagonist-narrator explains that he has "decided to become the second Robert Louis Stevenson, a tusitala or teller of tales, but with a big difference. I want to write a novel about me." Similarly, Epeli Hau'ofa of Tonga in his poem "Blood in the Kava Bowl" maintained that it is only the insider who has real access to a culture's deeper consciousness. These writers were echoing what was said in Africa, the West Indies, and other former colonial countries about literature: a culture must be written about from the inside, and the literature should be for the benefit of the local people.

Political unrest in the late 1980s, particularly in Fiji, Vanuatu, and Papua New Guinea, profoundly altered Pacific writers' perceptions about power and authority. Hau'ofa, the playwrights Vilisoni Herenkiko and Larry Thomas, the short-story writer Subramani, and others were no longer willing to accept the view that life in Oceania is ordered and culturally cohesive. Their works, which explore the subjects of individual and ethnic identity and the tensions and fragmentation of multicultural society, have ushered in a new phase in Oceanic literature.

The influence of oral traditions. The role of oral traditions in modern Oceanic literature is extremely significant and has been recognized by both writers and critics. In the early phase, the task of collecting and translating oral texts was an integral part of overall literary development and often inspired the creation of an original work. In addition, a variety of oral texts has been incorporated by Oceanic writers in short fiction, ranging from a Fijian cautionary tale in "The Taboo" (1980) by Akanisi Sobusobu and the legend about the Fijian shark-god Dakuwaqa in "A Childhood Experience" (1972) by Sitiveni Kalouniviti to anecdotes, jokes, risque tales, and tall tales in Epeli Hau'ofa's *Tales of the Tikongs* (1983). Oral forms have also been used in written works for specific thematic purposes. For example, Vincent Eri in his first novel, *The Crocodile* (1970), tried to give a sense of the spiritual world of the precontact society of Papua New Guinea, and he used traditional myths, legends, and tales of magic to express the life of a village where the sacred and secular coexist.

Perhaps the author who has been the most successful at integrating oral and written literatures has been Albert Wendt. Wendt began writing in the 1960s when there was no commonly recognized Oceanic literary style or store of techniques. Consequently, he invented for himself a mode of expression rich in Oceanic imagery, mythology, and colloquialism; and he invested and enriched the borrowed forms of the novel and the short story with a variety of narrative styles derived from traditional oral texts. He adopted the techniques of the raconteur and conversationalist, who freely mingle jokes, gossip, legends, and proverbs in their discourse, and fused them with the novelistic techniques of interior monologue and deployment of symbolism. Wendt's particular achievement has been his ability to absorb the history, myths, and other oral traditions of his country and to synthesize them with contemporary realities and the idiosyncracies of written fiction, imposing upon it all a vision that is his own. (S.)

BIBLIOGRAPHY. General characteristics of Oceanic civilizations, with analyses of culture and religion as the root of the arts, are explored in PETER BELLWOOD, *Man's Conquest of the Pacific: The Prehistory of Southeast Asia and Oceania* (1978); and DOUGLAS L. OLIVER, *Oceania: The Native Cultures of Australia and the Pacific Islands*, 2 vol. (1989).

Visual arts: Well-illustrated studies of the region as a whole include JEAN GUIART, *The Arts of the South Pacific*, trans. from French (1963); ALFRED BUEHLER, TERENCE BARROW, and CHARLES P. MOUNTFORD, *The Art of the South Sea Islands: Including Australia and New Zealand*, rev. ed. (1968); CARL A. SCHMITZ, *Oceanic Art: Myth, Man, and Image in the South Seas*, trans. from German (1969); SIDNEY M. MEAD (ed.), *Exploring the Visual Art of Oceania: Australia, Melanesia, Micronesia, and Polynesia* (1979); GEORGE A. CORBIN, *Native Arts of North America, Africa, and the South Pacific* (1988), a general introduction to aboriginal art; and PETER GATHERCOLE, ADRIENNE L. KAEPPLER, and DOUGLAS NEWTON, *The Art of the Pacific Islands* (1979).

The Aboriginal art of Australia is examined in DACRE STUBBS, *Prehistoric Art of Australia* (1974); ROBERT HUGHES, *The Art of Australia*, rev. ed. (1970); and RONALD M. BERNDT and E.S. PHILLIPS (eds.), *The Australian Aboriginal Heritage: An Introduction Through the Arts*, 2nd ed. (1978). For a survey of Melanesian art, see WALDEMAR STÖHR, *Kunst und Kultur aus der Südsee: Sammlung Clausmeyer Melanesien* (1987); and DOUGLAS NEWTON, *New Guinea Art in the Collection of the Museum of Primitive Art* (1967). On Polynesia as a whole, see EDWARD DODD, *A Pictorial Peregrination Through the Shapely and Harmonious, Often Enigmatical, Sometimes Shocking Realms of Polynesian Art* (1967; also published as *Polynesian Art*, 1969); ALLEN WARDWELL, *The Sculpture of Polynesia* (1967); and TERENCE BARROW, *Art and Life in Polynesia* (1972). Smaller areas are discussed in J. HALLEY COX and WILLIAM H. DAVENPORT, *Hawaiian Sculpture*, rev. ed. (1988); TERENCE BARROW, *The Art of Tahiti and the Neighbouring Society, Austral, and Cook Islands* (1979); and SIDNEY M. MEAD (ed.), *Te Maori: Maori Art from New Zealand Collections* (1984). On Micronesia, see JEROME FELDMAN and DONALD H. RUBINSTEIN, *The Art of Micronesia* (1986), a brief catalog with informative essays and a bibliography. (D.Ne.)

Music and dance: A general introduction is found in WILLIAM P. MALM, *Music Cultures of the Pacific, the Near East, and Asia*, 2nd ed. (1977). HANS FISCHER, *Sound-Producing Instruments in Oceania: Construction and Playing Technique, Distribution and Function*, rev. ed. (1986; originally published in German, 1958), provides a comprehensive study of the subject. For Melanesian music, see HUGO ZEMP, "Àrèare Classification of Musical Types and Instruments" and "Aspects of Àrèare Musical Theory," two original studies of music, musical instruments, and concepts in a Solomon Islands culture, both in *Ethnomusicology*, 22:37–67 (January 1978) and 23:5–48 (January 1979), respectively; JAAP KUNST, *Music in New Guinea*, trans. from Dutch (1967), covering western New Guinea; and STEVEN FELD, *Sound and Sentiment: Birds, Weeping, Poetics, and Song in Kaluli Expression* (1982), a penetrating study of the symbolic and emotional dimensions of music in a small-scale culture of Papua New Guinea. For Polynesia, see ADRIENNE L. KAEPPLER, *Polynesian Dance: With a Selection for Contemporary Performances* (1983), a description of dances and songs from several island areas; JOHANNES C. ANDERSEN, *Maori Music, with Its Polynesian Background* (1934, reprinted 1978), covering Polynesia in general with an emphasis on New Zealand; E.G. BURROWS, *Native Music of the Tuamotus* (1933, reprinted 1971), a study of Tuamotu chant in its cultural context, and *Songs of Uvea and Futuna* (1945, reprinted 1971), examining two small western Polynesian islands; JANE MINK ROSSEN, *Songs of Bellona Island*, 2 vol. (1987), a description of musical categories and styles of a Polynesian population in the Solomon Islands; E.S. CRAIGHILL HANDY and JANE LATHROP WINNE, *Music in the Marquesas Islands* (1925, reprinted 1971); HELEN H. ROBERTS, *Ancient Hawaiian Music* (1926, reprinted 1971); and JANE FREEMAN MOULIN, *The Dance of Tahiti* (1979), a detailed analysis. Micronesian dance is described in a short study, MARY BROWNING, *Micronesian Heritage* (1970). (Di.C./A.Ka.)

Literature: A general survey of literature is provided in NORMAN SIMMS, *Silence and Invisibility: A Study of the Literature of the Pacific, Australia, and New Zealand* (1986). Emerging literatures are further analyzed in H.H. ANNIAH GOWDA (ed.), *Essays in S. Pacific Literature* (1977); SUBRAMANI, *South Pacific Literature: From Myth to Fabulation* (1985); and SUBRAMANI (ed.), *The Indo-Fijian Experience* (1979), an anthology with the editor's descriptive introduction. Other useful anthologies with commentary include JOHANNES C. ANDERSEN, *Myth & Legends of the Polynesians* (1928, reprinted 1986); ALBERT WENDT (ed.), *Lali: A Pacific Anthology* (1980); ULLI BEIER (ed.), *Black Writing from New Guinea* (1973), and *Voices of Independence: New Black Writing from Papua New Guinea* (1980); BERNARD GADD (ed.), *Pacific Voices: An Anthology of Writing by and About Pacific People* (1977); and *Mana* (semiannual). (S.)

Oceans

When viewed from space, the predominance of the oceans on the Earth is readily apparent. The oceans and their marginal seas cover nearly 71 percent of the Earth's surface, with an average depth of 3,795 metres (12,450 feet). The exposed land occupies the remaining 29 percent of the planetary surface and has a mean elevation of only 840 metres (2,756 feet). Actually, all the elevated land could be hidden under the oceans and the Earth reduced to a smooth sphere that would be completely covered by a continuous layer of seawater 2,686 metres deep. This is known as the sphere depth of the oceans and serves to underscore the abundance of water on the Earth's surface.

The Earth is unique in the solar system because of its distance from the Sun and its period of rotation. These combine to subject the Earth to a solar radiation level that maintains the planet at a mean surface temperature of 16° C (61° F), which varies little over annual and night-day cycles. This mean temperature allows water to exist on the Earth in all three of its phases—solid, liquid, and gaseous. No other planet in the solar system has this feature. The liquid phase predominates on the Earth. By volume, 97.957 percent of the water on the planet exists as oceanic water and associated sea ice. The gaseous phase and droplet water in the atmosphere constitute 0.001 percent. Fresh water in lakes and streams makes up 0.036 percent, while groundwater is 10 times more abundant at 0.365 percent. Glaciers and ice caps constitute 1.641 percent of the Earth's total water volume.

Each of the above is considered to be a reservoir of water. Water continuously circulates between these reservoirs in what is called the hydrologic cycle, which is driven by energy from the Sun. Evaporation, precipitation, movement of the atmosphere, and the downhill flow of river water, glaciers, and groundwater keep water in motion between the reservoirs and maintain the hydrologic cycle.

The large range of volumes in these reservoirs and the rates at which water cycles between them combine to create important conditions on the Earth. If small changes occur in the rate at which water is cycled into or out of a reservoir, the volume of a reservoir changes. These volume changes may be relatively large and rapid in a small reservoir or small and slow in a large reservoir. A small percentage change in the volume of the oceans may produce a large proportional change in the land-ice reservoir, thereby promoting glacial and interglacial stages. The rate at which water enters or leaves a reservoir divided into the reservoir volume determines the residence time of water in the reservoir. The residence time of water in a reservoir, in turn, governs many of the properties of that reservoir.

This article focuses on the oceanic reservoir of the world. It discusses in general terms the properties of this body of water and the processes that occur within it and at its boundaries with the atmosphere and the crust of the Earth. The article also delineates the major features of the ocean basins, along with those of the continental margins and shorelines. Considered, too, are the economic aspects of the oceans, including some of the environmental problems linked with the utilization of marine resources.

For specifics concerning the relationship of the oceans to the other reservoirs of the Earth's waters, see HYDROSPHERE, THE. See also BIOSPHERE, THE, for coverage of the life-forms that populate the marine environment. Information about the nature, scope, and methods of oceanography and marine geology are provided in EARTH SCIENCES, THE.

For coverage of other related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 212, 221, 222, 223, 231, 232, 241, and 10/33, and the *Index*.

The article is divided into the following sections:

General considerations 135
  Relative distribution of the oceans 135
  Major subdivisions of the oceans 135
  Origin of the ocean waters 136
Chemical and physical properties of seawater 137
  Composition of seawater 137
    Dissolved inorganic substances
    Dissolved organic substances
    Effects of human activities
  Chemical evolution of the oceans 140
    The early oceans
    The transition stage
    The modern oceans
  Physical properties of seawater 142
    Salinity distribution
    Temperature distribution
    Thermal properties
    Density of seawater and pressure
    Optical properties
    Acoustic properties
    Ice in the sea
Circulation of the ocean waters 148
  General observations 148
  Distribution of ocean currents 148
  Causes of ocean currents 149
    Pressure gradients
    Coriolis effect
    Frictional forces
    Geostrophic currents
    Ekman layer
  Wind-driven circulation 151
    Equatorial currents
    The subtropical gyres
    The subpolar gyres
    Antarctic Circumpolar Current
  Thermohaline circulation 153
Waves of the sea 154
  Surface gravity waves 154
    Wind waves and swell
    Wind surges
    Waves of seismic origin
    Standing, or seiche, waves
    Internal waves
  Ocean tides 156
    Tide-generating forces
    Tidal bores
Density currents in the oceans 157
  General observations 157
  Turbidity currents 157
Impact of ocean-atmosphere interactions on weather and climate 158
  Seasonal and interannual ocean-atmosphere interactions 158
    General considerations
    Link between ocean surface temperature and climate anomalies
  Formation of tropical cyclones 160
    Conditions associated with cyclone formation
    Effects of tropical cyclones on ocean waters
    Influence on atmospheric circulation and rainfall
  The Gulf Stream and Kuroshio systems 161
    The Gulf Stream
    The Kuroshio
    Poleward transfer of heat
  El Niño/Southern Oscillation and climatic change 163
    The El Niño phenomenon
    The Southern Oscillation
Ocean basins 165
  Exploration of the ocean basins 165
  Oceanic crust 168
    Structure and composition
    Investigations of the oceanic crust
  Oceanic ridges 171
    Principal characteristics
    Distribution of major ridges and spreading centres
    Spreading centre zones and associated phenomena

Fracture zones and transform faults 173
Fracture zones
Transform faults
Deep-sea trenches 174
Types
Structure
Aseismic ridges 175
Seamounts, guyots, and abyssal hills 175
Deep-sea sediments 176
Sediment types
Sedimentation patterns
Evolution of the ocean basins through plate movements 178
Paleoceanography 179
Continental margins 180
General considerations 180
Margin types
Economic importance of continental margins
Continental shelf 182
Continental slope and rise 183
Formation of continental slopes
Formation of continental rises
Submarine canyons 184
Characteristics
Formation
Coastal and nearshore features 184
Coral reefs, coral islands, and atolls 184
Reef accumulation
Origin and development of reefs
Types of reefs
Lagoons 188
Nature of the lagoon environment
Factors involved in lagoon formation
Estuaries 189
Gulfs and bays 190
Factors that affect the characteristics of gulfs and bays
Classification of gulfs and bays
Economic aspects of the oceans 191
Medium for transportation and communications 191
Source of food and water 191
Fishing
Desalination
Energy resources 192
Tidal power generation
Ocean thermal energy conversion
Source of minerals and other raw materials 193
Petroleum
Minerals
Waste disposal and other related actions 194
Bibliography 195

General considerations

RELATIVE DISTRIBUTION OF THE OCEANS

Those conducting oceanic research generally recognize the existence of three major oceans, the Pacific, Atlantic, and Indian. (The Arctic Ocean is considered an extension of the Atlantic.) Arbitrary boundaries separate these three bodies of water in the Southern Hemisphere. One boundary extends southward to Antarctica from the Cape of Good Hope, while another stretches southward from Cape Horn. The last one passes through Malaysia and Indonesia to Australia, and then on to Antarctica. Many subdivisions can be made to distinguish the limits of seas and gulfs that have historical, political, and sometimes ecological significance (see Figure 1). However, water properties, ocean currents, and biological populations do not necessarily recognize these boundaries. Indeed, many researchers do not either. The oceanic area surrounding the Antarctic is considered by some to be the Southern Ocean.

The Southern Ocean

If area-volume analyses of the oceans are to be made, then boundaries must be established to separate individual regions. In 1921 Erwin Kossina, a German geographer, published tables giving the distribution of oceanic water with depth for the oceans and adjacent seas. This work was updated in 1966 by H.W. Menard and S.M. Smith. The latter only slightly changed the numbers derived by Kossina. This was remarkable, since the original effort relied entirely on the sparse depth measurements accumulated by individual wire soundings, while the more recent work had the benefit of acoustic depth soundings collected since the 1920s. This type of analysis, called hypsometry, allows quantification of the surface area distribution of the oceans and their marginal seas with depth.

The distribution of oceanic surface area with 5° increments of latitude shows that the distribution of land and water on the Earth's surface is markedly different in the Northern and Southern hemispheres (Figure 2). The Southern Hemisphere may be called the water hemisphere, while the Northern Hemisphere is the land hemisphere. This is especially true in the temperate latitudes.

This asymmetry of land and water distribution between the Northern and Southern hemispheres makes the two hemispheres behave very differently in response to the annual variation in solar radiation received by the Earth. The Southern Hemisphere shows only a small change in surface temperature from summer to winter at temperate latitudes. This variation is controlled primarily by the ocean's response to seasonal changes in heating and cooling. The Northern Hemisphere has one change in surface temperature controlled by its oceanic area and another controlled by its land area. In the temperate latitudes of the Northern Hemisphere, the land is much warmer than the oceanic area in summer and much colder in winter. This situation creates large-scale seasonal changes in atmospheric circulation and climate in the Northern Hemisphere that are not found in the Southern Hemisphere.

MAJOR SUBDIVISIONS OF THE OCEANS

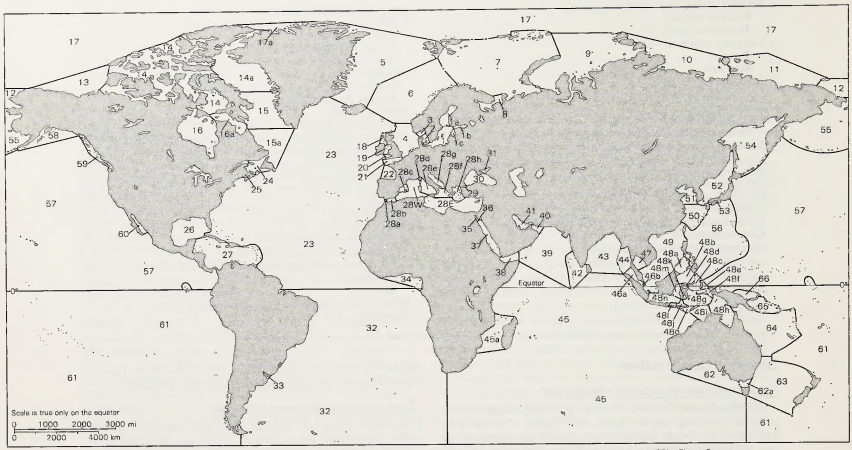
The surface areas and volumes of water contained in the oceans and major marginal seas are shown in Table 1. If the volume of an ocean is divided by its surface area, the mean depth is obtained. The data in this table indicate that with or without marginal seas, the Pacific is the largest ocean in both surface area and volume, the Atlantic is next, and the Indian is the smallest.

The Atlantic exhibits the largest change in surface area and volume when its marginal seas are subtracted. This indicates that the Atlantic has the greatest area of bordering seas, many of which are shallow.

Table 1 does not indicate how the oceanic water is distributed with depth except at the sea surface. Hypsometry can show how the area of each ocean or marginal sea changes as depth changes. A special curve known as a hypsometric, or hypsographic, curve can be drawn that portrays how the surface area of the Earth is distributed with elevation and depth (see Figure 3). This curve has been drawn for the total Earth and all its oceans. Individual oceans and seas are not portrayed, although similar curves can be constructed for each ocean and subdivision. The average depth of the world's oceans, 3,795 metres, and the average elevation of the land, 840 metres, are indicated. The highest point on land, Mount Everest (8,850 metres), and the deepest point in the ocean, located in the Mariana Trench (11,034 metres), mark the upper and lower limits of the curve, respectively. Since this curve is drawn on a grid of elevation versus the Earth's area, the area under the curve covering the 29.2 percent of the Earth's surface that is above sea level is the volume of land above sea level. Similarly, the area between sea level and the curve depicting the remaining 70.8 percent of the Earth's surface below sea level represents the volume of water contained in the oceans.

Hypsographic curve

Portions of this curve describe the area of the Earth's surface that exists between elevation or depth increments. On land, little of the Earth's total area—only about 4 percent—is at elevations above 2,000 metres. Most of the land, 25.3 percent of the total Earth, is between 0 and 2,000 metres. About 13.6 percent of the total land area is at higher elevations, with 86.4 percent between 0 and 2,000 metres when the areas are determined relative to land area only. In the oceans, the percentages of the area devoted to depth increments yield information about the typical structure and shape of the oceanic basins. The small depth increment of 0–200 metres occupies about 5.4 percent of the Earth's total area or 7.6 percent of the oceans' area. This approximates the world's area of continental shelves, the shallow flat borderlands of the continents that have been alternately covered by the oceans during interglacial



1. Baltic Sea
1a. Gulf of Bothnia
1b. Gulf of Finland
1c. Gulf of Riga
2. Kattegat, Sound and Belts
3. Skagerrak
4. North Sea
5. Greenland Sea
6. Norwegian Sea
7. Barents Sea
8. White Sea
9. Kara Sea
10. Laptev Sea
11. East Siberian Sea
12. Chukchi Sea
13. Beaufort Sea
14. Northwestern Passages
14a. Baffin Bay
15. Davis Strait
15a. Labrador Sea
16. Hudson Bay
16a. Hudson Strait
17. Arctic Ocean
17a. Lincoln Sea
18. Inner Seas (off the west coast of Scotland)
19. Irish Sea and St. George's Channel
20. Bristol Channel
21. English Channel
22. Bay of Biscay
23. North Atlantic Ocean
24. Gulf of St. Lawrence
25. Bay of Fundy
26. Gulf of Mexico
27. Caribbean Sea
28W. Mediterranean Sea (Western Basin)
28a. Strait of Gibraltar
28b. Alboran Sea
28c. Balearic Sea
28d. Ligurian Sea
28e. Tyrrhenian Sea
28E. Mediterranean Sea (Eastern Basin)
28f. Ionian Sea
28g. Adriatic Sea
28h. Aegean Sea
29. Sea of Marmara
30. Black Sea
31. Sea of Azov
32. South Atlantic Ocean
33. Rio de la Plata
34. Gulf of Guinea
35. Gulf of Suez
36. Gulf of Aqaba
37. Red Sea
38. Gulf of Aden
39. Arabian Sea
40. Gulf of Oman
41. Persian Gulf
42. Laccadive Sea
43. Bay of Bengal
44. Andaman Sea
45. Indian Ocean
45a. Mozambique Channel
46. Malacca and Singapore Straits
46a. Strait of Malacca
46b. Singapore Strait
47. Gulf of Thailand
48. East Indian Archipelago (Indonesia)
48a. Sulu Sea
48b. Celebes Sea
48c. Molucca Sea
48d. Gulf of Tomini
48e. Halmahera Sea
48f. Ceram Sea
48g. Banda Sea
48h. Arafura Sea
48i. Timor Sea
48j. Flores Sea
48k. Gulf of Born
48l. Bali Sea
48m. Makassar Strait
48n. Jave Sea
48o. Savu Sea
49. South China Sea
50. East China Sea
51. Yellow Sea
52. Sea of Japan
53. Inland Sea
54. Sea of Okhotsk
55. Bering Sea
56. Philippine Sea
57. North Pacific Ocean
58. Gulf of Alaska
59. Coastal waters of southeast Alaska and British Columbia
60. Gulf of California
61. South Pacific Ocean
62. Great Australian Bight
62a. Bass Strait
63. Tasman Sea
64. Coral Sea
65. Solomon Sea
66. Bismarck Sea

Figure 1: Boundaries of the world's oceans and seas.
By courtesy of the International Hydrographic Organization

stages and uncovered during glacial periods (see below the section *Continental margins: Continental shelf*).

At depths between 200 and 1,000 metres and between 1,000 and 2,000 metres, an area only slightly larger—6.02 percent of the Earth's total area or 8.5 percent of the oceans' area—is found. These depths are related to the regions of the oceans that have very steep slopes where depth increases rapidly. These are the continental slope regions that mark the true edge of the continental landmasses. Marginal seas of moderate depths and the tops of seamounts, however, add their area to these depth zones when all the oceans are considered. The majority of the oceanic area lies between 4,000 and 5,000 metres.

The continental shelf region varies immensely from place to place. The seaward boundary of the continental shelf historically is determined by the 100-fathom, or 200-metre, depth contour. However, 85 fathoms, or 170 metres, is a closer approximation. The true boundary at any given location is marked by a rapid change in slope of the seafloor known as the shelf break. This change in slope may be nearly at the coastline in areas where crustal plates converge, as along the west coast of North and South America, or it may be located more than 1,000 kilometres seaward of the coast, as off the north coast of Siberia. The average width of the shelf is about 75 kilometres, and the shelf has an average slope of about 0.01°, a slope that is barely discernible to the human eye. Seaward of the shelf break, the continental slope is inclined by about 4°.

ORIGIN OF THE OCEAN WATERS

The huge volume of water contained in the oceans (and seas), 137×10^7 cubic kilometres, has been produced during the geologic history of the Earth. There is little information on the early history of the Earth's waters. However, fossils dated from the Precambrian some 3.3 billion years ago show that bacteria and cyanobacteria (blue-green algae) existed, indicating the presence of water during this period. Carbonate sedimentary rocks, obviously laid down in an aquatic environment, have been dated to 1 billion years ago. Also, there is fossil evidence of primitive marine algae and invertebrates from the outset of the Cambrian Period some 540 million years ago.

The presence of water on the Earth at even earlier times is not documented by physical evidence. It has been suggested, however, that the early hydrosphere formed in response to condensation from the early atmosphere. The ratios of certain elements on the Earth indicate that the planet formed by the accumulation of cosmic dust and was slowly warmed by radioactive and compressional heating. This heating led to the gradual separation and migration of materials to form the Earth's core, mantle, and crust. The early atmosphere is thought to have been highly reducing and rich in gases, notably in hydrogen, and to include water vapour.

The Earth's surface temperature and the partial pressures of the individual gases in the early atmosphere affected the atmosphere's equilibration with the terrestrial surface.

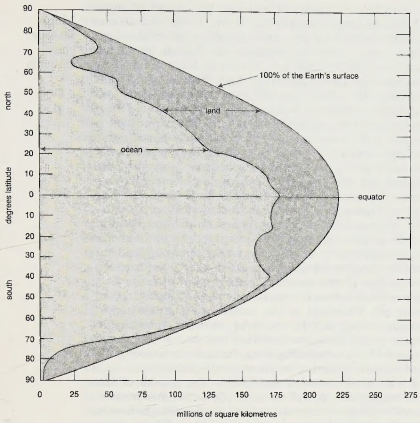


Figure 2: *Distribution of land and ocean waters by latitude.* The surface areas of the landmasses and the oceans are almost equal in the mid-latitudes of the Northern Hemisphere. By contrast, land is nearly absent in the same latitudes of the Southern Hemisphere.

From A.C. Duxbury and A.B. Duxbury, *Introduction to the World's Oceans* (1989); Wm. C. Brown Group

As time progressed and the planetary interior continued to warm, the composition of the gases escaping from within the Earth gradually changed the properties of its atmosphere, producing a gaseous mixture rich in carbon dioxide (CO_2), carbon monoxide (CO), and molecular nitrogen (N_2). Photodissociation (*i.e.*, separation due to the energy of light) of water vapour into molecular hydrogen (H_2) and molecular oxygen (O_2) in the upper atmosphere allowed the hydrogen to escape and led to a progressive increase of the partial pressure of oxygen at the Earth's surface. The reaction of this oxygen with the materials of the surface gradually caused the vapour pressure of water vapour to increase to a level at which liquid water could form. This water in liquid form accumulated in isolated depressions of the Earth's surface, forming the nascent oceans. The high carbon dioxide content of the atmosphere at this time would have allowed a buildup of dissolved carbon dioxide in the water and made these early oceans acidic and capable of dissolving surface rocks that would add to the water's salt content. Water must have evaporated and condensed rapidly and accumulated slowly at first. The required buildup of atmospheric oxygen was slow because much of this gas was used to oxidize methane, ammonia, and exposed rocks high in iron. Gradually, the partial pressure of the oxygen gas in the atmosphere rose as photosynthesis by bacteria and photodissociation continued to supply oxygen. Biological processes involving algae increased, and they gradually decreased the carbon dioxide content and increased the oxygen content of the atmosphere until the oxygen produced by biological processes outweighed that produced by photodissociation. This, in turn, accelerated the formation of surface water and the development of the oceans. (For further details on the formation and development of the oceans, see below *Chemical and physical properties of seawater: Chemical evolution of the oceans.*) (A.C.D.)

Photochemical production of oxygen

Chemical and physical properties of seawater

COMPOSITION OF SEAWATER

The chemical compositon of seawater is influenced by a wide variety of chemical transport mechanisms. Rivers add dissolved and particulate chemicals to the oceanic margins. Wind-borne particulates are carried to mid-ocean regions thousands of kilometres from their continental source areas. Hydrothermal solutions that have circulated through crustal materials beneath the seafloor add both dissolved and particulate materials to the deep ocean. Organisms in the upper ocean convert dissolved materials to solids, which eventually settle to greater oceanic depths. Particulates in transit to the seafloor, as well as materials both on and within the seafloor, undergo chemical exchange with surrounding solutions. Through these local and regional chemical input and removal mechanisms, each element in the oceans tends to exhibit spatial and temporal concentration variations. Physical mixing in the oceans (thermohaline and wind-driven circulation; see below *Circulation of the ocean waters*) tends to homogenize the chemical composition of seawater. The opposing influences of physical mixing and of biogeochemical input and removal mechanisms result in a substantial variety of chemical distributions in the oceans.

Principal inorganic constituents of seawater

Dissolved inorganic substances. In contrast to the behaviour of most oceanic substances, the concentrations of the principal inorganic constituents of the oceans (Table 2) are remarkably constant. For 98 percent of the oceans' volume, the concentrations of the constituents shown in the Table vary by less than 3 percent from the values given in columns 2 and 3. Furthermore, with the exception of inorganic carbon, the principal constituents shown in the Table have very nearly fixed ion concentration ratios (column 4). Calculations indicate that, for the main constituents of seawater, the time required for thorough oceanic mixing is quite short compared to the time that

Table 1: Surface Area, Volume, and Average Depth of Oceans and Seas

	area		volume		average depth	
	000,000 sq km	000,000 sq mi	000,000 cu km	000,000 cu mi	m	ft
Atlantic Ocean						
without marginal seas	82.440	31.830	324.600	77.900	3,930	12,890
with marginal seas	106.460	41.100	354.700	85.200	3,330	10,920
Pacific Ocean						
without marginal seas	165.250	63.800	707.600	169.900	4,280	14,040
with marginal seas	179.680	69.370	723.700	173.700	4,030	13,220
Indian Ocean						
without marginal seas	73.440	28.360	291.000	69.900	3,960	10,040
with marginal seas	74.920	28.930	291.900	70.100	3,900	12,790
Arctic Ocean	14.090	5.440	17.000	4.100	1,205	3,950
Mediterranean Sea and Black Sea	2.970	1.150	4.200	1.000	1,430	4,690
Gulf of Mexico and Caribbean Sea	4.320	1.670	9.600	2.300	2,220	7,280
Australasian Central Sea	8.140	3.140	9.900	2.400	1,210	3,970
Hudson Bay	1.230	0.470	0.160	0.040	128	420
Baltic Sea	0.420	0.160	0.020	0.005	55	180
North Sea	0.570	0.220	0.050	0.010	94	310
English Channel	0.075	0.029	0.004	0.001	54	180
Irish sea	0.100	0.040	0.006	0.001	60	200
Sea of Okhotsk	1.583	0.611	1.300	0.300	838	2,750
Bering Sea	2.304	0.890	3.330	0.800	1,440	4,720
The world ocean	361.100	139.400	1,370.000	329.000	3,790	12,430

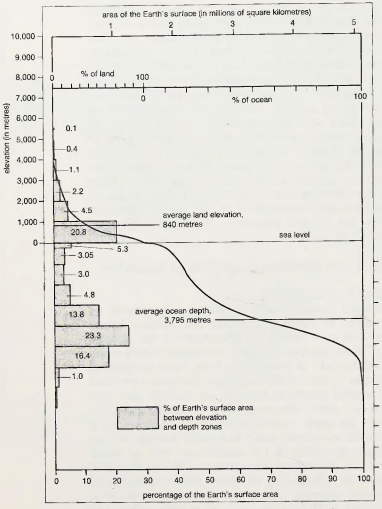


Figure 3: Hypsographic curve showing how the surface area of the Earth is distributed with elevation and depth.

would be required for input or removal processes to significantly change a constituent's concentration. The concentrations of the principal constituents of the oceans vary primarily in response to a comparatively rapid exchange of water (precipitation and evaporation), with relative concentrations remaining nearly constant.

Salinity is used by oceanographers as a measure of the total salt content of seawater. Practical salinity, symbol S, is determined through measurements of a ratio between the electrical conductivity of seawater and the electrical conductivity of a standard solution. Practical salinity can be used to calculate precisely the density of seawater samples. Because of the constant relative proportions of the principal constituents, salinity can also be used to directly calculate the concentrations of the major ions in seawater. Using the relative concentrations shown in column 4 of Table 2, ionic concentations are calculated as 0.015577 mole per kilogram multiplied by salinity multiplied by relative concentration. The measure of practical salinity was originally developed to provide an approximate measure of the total mass of salt in one kilogram of seawater. Seawater with S equal to 35 contains approximately 35 grams of salt and 965 grams of water.

Table 2: Principal Constituents of Seawater*

ionic constituent	g/kg of seawater	moles/kg†	relative concentration
Chloride (Cl^-)	19.162	0.5405	1.0000
Sodium (Na^+)	10.679	0.4645	0.8593
Magnesium (Mg^{2+})	1.278	0.0526	0.0974
Sulfate (SO_4^{2-})	2.680	0.0279	0.0517
Calcium (Ca^{2+})	0.4096	0.01022	0.0189
Potassium (K^+)	0.3953	0.01011	0.0187
Carbon (inorganic)	0.0276	0.0023	0.0043
Bromide (Br^-)	0.0663	0.00083	0.00154
Boron (B)	0.0044	0.00041	0.00075
Strontium (Sr^{2+})	0.0079	0.00009	0.000165
Fluoride (F^-)	0.0013	0.00007	0.000125

*Concentrations at salinity equal to 34.7. †Ionic concentrations are conventionally expressed in molecular units. One mole per kilogram is equivalent to 6.023×10^{23} ions or molecules per kilogram of seawater. The relative concentrations in column 4 provide the number of ions of each constituent in one kilogram of seawater as compared to the number of chloride ions in one kilogram of seawater.

Although the 11 constituents shown in Table 2 account for more than 99.5 percent of the dissolved solids in seawater, many other constituents are of great importance to the biogeochemistry of the oceans. Such chemicals as inorganic phosphorus (HPO_4^{2-} and PO_4^{3-}) and inorganic nitrogen (NO_3^-, NO_2^-, and NH_4^+) are essential to the growth of marine organisms. Nitrogen and phosphorus are incorporated into the tissues of marine organisms in approximately a 16:1 ratio and are eventually returned to solution in approximately the same proportion. As a consequence, in much of the oceanic waters dissolved inorganic phosphorus and nitrogen exhibit a close covariance. Figure 4 shows dissolved inorganic phosphorus distributions in the Pacific Ocean. The observed distributions strongly bear the imprint of phosphorus incorporation by organisms in the surface waters of the ocean and of the return of the phosphorus to solution via a rain of biological debris remineralized in the deep ocean. Inorganic phosphate concentrations in the western Pacific (Figure 4) range from somewhat less than 0.1 micromole per kilogram (1×10^{-7} mole per kilogram) at the surface to approximately 3 micromoles/kg (3×10^{-6} mole/kg) at depth. Inorganic nitrogen ranges between somewhat less than 1 micromole/kg and 45 micromoles/kg along the same section of ocean and exhibits a striking covariance with phosphate.

Minor seawater constituents

A variety of elements essential to the growth of marine organisms, as well as some elements that have no known biological function, exhibit nutrient-like behaviour broadly similar to nitrate and phosphate. Silicate is incorporated into the hard structural parts of certain types of marine organisms (diatoms and radiolarians) that are abundant in the upper ocean. Dissolved silicate concentrations range between less than 1 micromole/kg (1×10^{-6} mole/kg) in surface waters to approximately 180 micromoles/kg (1.8×10^{-4} mole/kg) in the deep North Pacific. The concentration of zinc, a metal essential to a variety of biological functions, ranges between approximately 0.05 nanomole/kg (5×10^{-11} mole/kg) in the surface ocean to as much as 6 nanomoles/kg (6×10^{-9} mole/kg) in the deep Pacific. The distribution of zinc in the oceans is observed to generally parallel silicate distributions. Cadmium, though having no known biological function, generally exhibits distributions that are covariant with phosphate and concentrations that are even lower than those of zinc.

Many elements, including the essential trace metals iron, cobalt, and copper, show surface depletions but in general exhibit behaviour more complex than that of phosphate, nitrate, and silicate. Some of the complexities observed in elemental oceanic distributions are attributable to the adsorption of elements on the surface of sinking particles. Adsorptive processes, either exclusive of or in addition to biological uptake, serve to remove elements from the upper ocean and deliver them to greater depths. The distribution patterns of a number of trace elements are complicated by their participation in oxidation-reduction (electron-exchange) reactions. In general, electron-exchange reactions lead to profound changes in the solubility and reactivity of trace metals in seawater. Such reactions are important to the oceanic behaviour of a variety of elements, including iron, manganese, copper, cobalt, chromium, and cerium.

Dissolved gases

The processes that deliver dissolved, particulate, and gaseous materials to the oceans ensure that they contain, at some concentration, very nearly every element that is found in the Earth's crust and atmosphere. The principal components of the atmosphere, nitrogen (78.1 percent), oxygen (21.0 percent), argon (0.93 percent), and carbon dioxide (0.035 percent), occur in seawater in variable proportions, depending on their solubilities and oceanic chemical reactions. In equilibrium with the atmosphere, the concentrations of the unreactive gases, nitrogen and argon, in seawater (0° C, salinity 35) are 616 micromoles/kg and 17 micromoles/kg, respectively. For seawater at 35° C, these concentrations would decrease by approximately a factor of two. The solubility behaviours of argon and oxygen are quite similar. For seawater in equilibrium with the atmosphere, the ratio of oxygen and argon concentrations is approximately 20.45. Since oxygen is a reactive gas

essential to life, oxygen concentrations in seawater that are not in direct equilibrium with the atmosphere are quite variable. Although oxygen is produced by photosynthetic organisms at shallow, sunlit ocean depths, oxygen concentrations in near-surface waters are established primarily by exchange with the atmosphere. Oxygen concentrations in the oceans generally exhibit minimum values at intermediate depths and relatively high values in deep waters. This distribution pattern results from a combination of biological oxygen utilization and physical mixing of the ocean waters. Estimates of the extent of oxygen utilization in the oceans can be obtained by comparing concentrations of oxygen with those of argon, since the latter are only influenced by physical processes. The physical processes that influence oxygen distributions include, in particular, the large-scale replenishment of oceanic bottom waters with cold, dense, oxygen-rich waters sinking toward the bottom from high latitudes. Due to the release of nutrients that accompanies the consumption of oxygen by biological debris, dissolved oxygen concentrations generally appear as a mirror image of dissolved nutrient concentrations, such as those shown in Figure 4.

While the atmosphere is a vast repository of oxygen compared to the oceans, the total carbon dioxide content of the oceans is very large compared to that of the atmosphere. Carbon dioxide reacts with water in seawater to form carbonic acid (H_2CO_3), bicarbonate ions (HCO_3^-), and carbonate ions (CO_3^{2-}). Approximately 90 percent of the total organic carbon in seawater is present as bicarbonate ions. The formation of bicarbonate and carbonate ions from carbon dioxide is accompanied by the liberation of hydrogen ions (H^+). Reactions between hydrogen ions and the various forms of inorganic carbon buffer the acidity of seawater. The relatively high concentrations of both total inorganic carbon and boron—as $B(OH)_3$ and $B(OH)_4^-$—in seawater (see Table 2) are sufficient to maintain the pH of seawater between 7.4 and 8.3. (The term pH is defined as the negative logarithm of the hydrogen ion concentration in moles per kilogram. Thus, a pH equal to 8 is equivalent to 1×10^{-8} mole of H^+ ions per kilogram of seawater.) This is quite important because the extent and rate of many reactions in seawater are highly pH-dependent. Carbon dioxide produced by the combination of oxygen and organic carbon generally produces an acidity maximum (pH minimum) near the depth of the oxygen minimum in seawater. In addition to exchange with the atmosphere and, through respiration, with the biosphere, dissolved inorganic carbon concentrations in seawater are influenced by the formation and dissolution of the calcareous shells ($CaCO_3$) of organisms (foraminiferans, coccolithophores, and pteropods) abundant in the upper ocean.

Dissolved organic substances. Processes involving dissolved and particulate organic carbon are of central importance in shaping the chemical character of seawater. Marine organic carbon principally originates in the uppermost 100 metres of the oceans where dissolved inorganic carbon is photosynthetically converted to organic materials. The "rain" of organic-rich particulate materials, resulting directly and indirectly from photosynthetic production, is a principal factor behind the distributions of many organic and inorganic substances in the oceans. A large fraction of the vertical flux of materials in the uppermost waters is converted to dissolved substances within the upper 400 metres of the oceans. Dissolved organic carbon (DOC) accounts for at least 90 percent of the total organic carbon in the oceans. Estimates of DOC appropriate to the surface of the open ocean range between roughly 100 and 500 micromoles of carbon per kilogram of seawater. DOC concentrations in the deep ocean are 5 to 10 times lower than surface values. DOC occurs in an extraordinary variety of forms, and, in general, its composition is controversial and poorly understood. Conventional techniques have indicated that, in surface waters, about 15 percent of DOC can be identified as carbohydrates and combined amino acids. At least 1–2 percent of DOC in surface waters occurs as lipids and 20–25 percent as relatively unreactive humic substances. The relative abundances of reactive organic substances, such as amino acids and carbohydrates, are considerably reduced in deep ocean waters. Dissolved and particulate organic carbon in the surface ocean participates in diel cycles (*i.e.*, those of a 24-hour period) related to photosynthetic production and photochemical transformations. The influence of dissolved organic matter on ocean chemistry is often out of proportion to its oceanic abundance. Photochemical reactions involving DOC can influence the chemistry of vital trace nutrients such as iron, and, even at dissolved concentrations on the order of one nanomole/kg (1×10^{-9} mole/kg), dissolved organic substances in the upper ocean waters are capable of greatly altering the bioavailability of essential trace nutrients, as, for example, copper and zinc.

Problems of marine pollution

Effects of human activities. Although the oceans constitute an enormous reservoir, human activities have begun to influence their composition on both a local and a global scale. The addition of nutrients (through the discharge of untreated sewage or the seepage of soluble mineral fertilizers, for example) to coastal waters results in increased phytoplankton growth, high levels of dissolved

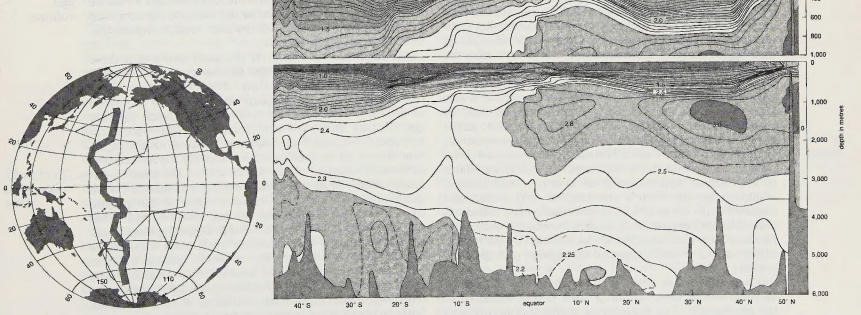


Figure 4: Vertical distribution of dissolved inorganic phosphate concentrations (in micromoles per kilogram) in the western Pacific Ocean. The vertical exaggeration is 2500:1 in the top section and 1000:1 in the bottom section. The segment of the ocean surveyed is shown in the inset at the lower left (see text).

and particulate organic materials, decreased penetration of light through seawater, and alteration of the community structure of bottom-dwelling organisms. Through industrial and automotive emissions, lead concentrations in the surface ocean have increased dramatically on a global scale compared with preindustrial levels. Certain toxic organic compounds, such as polychlorinated biphenyls (PCBs), are found in seawater and marine organisms and are attributable solely to the activities of humankind. Although most radioactivity in seawater is natural (approximately 90 percent as potassium-40 and less than 1 percent each as rubidium-87 and uranium-238), strontium-90 and certain other artificial radioisotopes have unique environmental pathways and potential for bioaccumulation. Among the most dramatic influences of human activities on a global scale is the remarkable increase of carbon dioxide levels in the atmosphere. Atmospheric carbon dioxide levels are expected to double by the middle of the 21st century, with potentially profound consequences for global climate and agricultural patterns. It is thought that the oceans, as a great reservoir of carbon dioxide, will ameliorate this consequence of human activities to some degree. (For more specific information on this subject, see HYDROSPHERE, THE: *Impact of human activities on the hydrosphere: Buildup of greenhouse gases.*) (R.H.By.)

CHEMICAL EVOLUTION OF THE OCEANS

The chemical history of the oceans has been divided into three stages. The first is an early stage in which the Earth's crust was cooling and reacting with volatile or highly reactive gases of an acidic, reducing nature to produce the oceans and an initial sedimentary rock mass. This stage lasted until about 3.5 billion years ago. The second stage was a period of transition from the initial to essentially modern conditions, and it is estimated to have ended 2 to 1.5 billion years ago. Since that time it is likely that there has been little change in seawater composition.

The early oceans. The initial accretion of the Earth by agglomeration of solid particles occurred about 4.6 billion years ago. Heating of this initially cool, unsorted conglomerate by the decay of radioactive elements and the conversion of kinetic and potential energy to heat resulted in the development of a liquid iron core and the gross internal zonation of the Earth (see EARTH: *The structure and composition of the solid Earth*). It has been concluded that formation of the Earth's core took about 500 million years. It is likely that core formation resulted in the escape of an original primitive atmosphere and its replacement by one derived from loss of volatile substances from the Earth's interior. Whether most of this degassing took place during core formation or soon afterward or whether there has been significant degassing of the Earth's interior throughout geologic time is uncertain. Recent models of Earth formation, however, suggest early differentiation of the Earth into three major zones (core, mantle, and crust) and attendant early loss of volatile substances from the interior. It is also likely that the Earth, after initial cold agglomeration, reached temperatures such that the whole Earth approached the molten state. As the initial crust of the Earth solidified, volatile gases would be released to form an atmosphere that would contain water, later to become the hydrosphere; carbon gases, such as carbon dioxide, methane, and carbon monoxide; sulfur gases, mostly hydrogen sulfide; and halogen compounds, such as hydrochloric acid. Nitrogen also may have been present, along with minor amounts of other gases. Gases of low atomic number, such as hydrogen and helium, would escape the Earth's gravitational field. Substances degassed from the planetary interior have been called excess volatiles because their masses cannot be accounted for simply by rock weathering. An estimate of the masses of the various volatiles degassed throughout geologic time is given in Table 3.

Table 3: Estimate of "Excess Volatiles" (units of 10^{20} grams)

Water	16,600
Total carbon as carbon dioxide	910
Sulfur	22
Nitrogen	42
Chlorine	300
Hydrogen	10
Boron, bromine, argon, fluorine, etc.	4

Source: W.W. Rubey (1951).

At an initial crustal temperature of about 600° C, almost all these compounds, including water (H_2O), would be in the atmosphere. The sequence of events that occurred as the crust cooled is difficult to construct. Below 100° C all the H_2O would have condensed, and the acid gases would have reacted with the original igneous crustal minerals to form sediments and an initial ocean. There are at least two possible pathways by which these initial steps could have been accomplished.

One pathway assumes that the 600° C atmosphere contains, together with other compounds, water (as vapour), carbon dioxide, and hydrochloric acid in the ratio of 20:3:1 and would cool to the critical temperature of water. The water vapour therefore would have condensed into an early hot ocean. At this stage, the hydrochloric acid would be dissolved in the ocean (about 1 mole per litre), but most of the carbon dioxide would still be in the atmosphere with about 0.5 mole per litre in the ocean water. This early acid ocean would react vigorously with crustal minerals, dissolving out silica and cations and creating a residue that consisted principally of aluminous clay minerals that would form the sediments of the early ocean basins. This pathway of reaction assumes that reaction rates are slow relative to cooling. A second pathway of reaction, which assumes that cooling is slow, is also possible. In this case, at a temperature of about 400° C most of the water vapour would be removed from the atmosphere by hydration reactions with pyroxenes and olivines. Under these conditions, water vapour would not condense until some unknown temperature was reached, and the Earth might have had at an early stage in its history an atmosphere rich in carbon dioxide and no ocean: the surface would have been much like that of present-day Venus.

The pathways described are two of several possibilities for the early surface environment of the Earth. In either case, after the Earth's surface had cooled to 100° C, it would have taken only a short time geologically for the acid gases to be used up in reactions involving igneous rock minerals. The presence of bacteria and possibly algae in the fossil record of rocks older than 3 billion years attests to the fact that the Earth's surface had cooled to temperatures lower than 100° C by this time and that the neutralization of the original acid gases had taken place. If most of the degassing of primary volatile substances from the Earth's interior occurred early, the chloride released by reaction of hydrochloric acid with rock minerals would be found in the oceans and seas or in evaporite deposits, and the oceans would have a salinity and volume comparable to those that they have today.

Early salinity and volume

This conclusion is based on the assumption that there has been no drastic change in the ratios of volatiles released through geologic time. The overall generalized reaction indicative of the chemistry leading to formation of the early oceans can be written in the form: primary igneous rock minerals + acid volatiles + $H_2O \rightarrow$ sedimentary rocks + oceans + atmosphere. Notice from this equation that if all the acid volatiles and H_2O were released early in the history of the Earth and in the proportions found today, then the total original sedimentary rock mass produced would be equal to that of the present time, and ocean salinity and volume would be near what they are now. If, on the other hand, degassing were linear with time, then the sedimentary rock mass would have accumulated at a linear rate, as would oceanic volume. However, the salinity of the oceans would remain nearly the same if the ratios of volatiles degassed did not change with time. The most likely situation is that presented here—namely, that major degassing occurred early in Earth history, after which minor amounts of volatiles were released episodically or continuously for the remainder of geologic time. The salt content of the oceans based on the

constant proportions of volatiles released would depend primarily on the ratio of sodium chloride (NaCl) locked up in evaporites to that dissolved in the oceans. If all the sodium chloride in evaporites were added to the oceans today, the salinity would be roughly doubled. This value gives a sense of the maximum salinity the oceans could have attained throughout geologic time.

Production of oxygen

One component missing from the early terrestrial surface was free oxygen because it would not have been a constituent released from the cooling crust. As noted earlier, early production of oxygen was by photodissociation of water in the atmosphere as a result of absorption of ultraviolet light. The reaction is $2H_2O + h\nu \rightarrow O_2 + 2H_2$, in which $h\nu$ represents a photon of ultraviolet light. The hydrogen produced would escape into space, and the O_2 would react with the early reduced gases by reactions such as $2H_2S + 3O_2 \rightarrow 2SO_2 + 2H_2O$. Oxygen production by photodissociation gave the early reduced atmosphere a start toward present-day conditions, but it was not until the appearance of photosynthetic organisms approximately 3.3 billion years ago that it was possible for the accumulation of oxygen in the atmosphere to proceed at a rate sufficient to lead to today's oxygenated environment. The photosynthetic reaction leading to oxygen production may be written $6CO_2 + 6H_2O + h\nu \rightarrow C_6H_{12}O_6 + 6O_2$, in which $C_6H_{12}O_6$ represents sugar.

The transition stage. The nature of the rock record from the time of the first sedimentary rocks (about 3.5 billion years ago) to approximately 2 to 1.5 billion years ago suggests that the amount of oxygen in the atmosphere was significantly lower than today and that there were continuous chemical trends in the sedimentary rocks formed and, more subtly, in oceanic composition. The source rocks of sediments during this time were likely to be more basaltic than would later ones; sedimentary detritus was formed by the alteration of these rocks in an oxygen-deficient atmosphere and accumulated primarily under anaerobic marine conditions. The chief difference between reactions involving mineral-ocean equilibriums at this time and at the present time was the role played by ferrous iron. The concentration of dissolved iron in the present-day oceans is low because of the insolubility of oxidized iron oxides. During the period 3.5 to 1.5 billion years ago, oxygen-deficient environments were prevalent; these favoured the formation of minerals containing ferrous iron (reduced state of iron) from the alteration of basaltic rocks. Indeed, the iron carbonate siderite and the iron silicate greenalite, in close association with chert and the iron sulfide pyrite, are characteristic minerals that occur in middle Precambrian iron formations (those about 1.5 to 2.4 billion years old). The chert originally was deposited as amorphous silica; equilibrium between amorphous silica, siderite, and greenalite at 25° C and one atmosphere total pressure requires a carbon dioxide pressure of about $10^{-2.5}$ atmosphere, or 10 times the present-day value.

The oceans at this time can be thought of as the solution resulting from an acid leach of basaltic rocks, and because the neutralization of the volatile acid gases was not restricted primarily to land areas as it is presently, much of this alteration may have occurred by submarine processes. The atmosphere at the time was oxygen-deficient; anaerobic depositional environments with internal carbon dioxide pressures of about $10^{-2.5}$ atmosphere were prevalent, and the atmosphere itself may have had a carbon dioxide pressure near $10^{-2.5}$ atmosphere. If so, the pH of early ocean water was lower than that of modern seawater, the calcium concentration was higher, and the early ocean water was probably saturated with respect to amorphous silica (about 120 parts per million [ppm]).

The pH of early ocean water

To simulate what might have occurred, it is helpful to imagine emptying the Pacific basin, throwing in great masses of broken basaltic material, filling it with hydrochloric acid so that the acid becomes neutralized, and then carbonating the solution by bubbling carbon dioxide through it. Oxygen would not be permitted into the system. The hydrochloric acid would leach the rocks, resulting in the release and precipitation of silica and the production of a chloride ocean containing sodium, potassium, calcium, magnesium, aluminum, iron, and reduced sulfur species in the proportions present in the rocks. As complete neutralization was approached, aluminum could begin to precipitate as hydroxides and then combine with precipitated silica to form cation-deficient aluminosilicates. The aluminosilicates, as the end of the neutralization process was reached, would combine with more silica and with cations to form minerals like chlorite, and ferrous iron would combine with silica and sulfur to make greenalite and pyrite. In the final solution, chlorine would be balanced by sodium and calcium in roughly equal proportions, with subordinate potassium and magnesium; aluminum would be quantitatively removed, and silicon would be at saturation with amorphous silica. If this solution were then carbonated, calcium would be removed as calcium carbonate, and the chlorine balance would be maintained by abstraction of more sodium from the primary rock. The sediments produced in this system would contain chiefly silica, ferrous iron silicates, chloritic minerals, calcium carbonate, calcium magnesium carbonates, and minor pyrite.

If the hydrochloric acid added were in excess of the carbon dioxide, the resultant ocean would have a high content of calcium chloride, but the pH would still be near neutrality. If the carbon dioxide added were in excess of the chlorine, calcium would be precipitated as the carbonate until it reached a level approximately that of the present oceans—namely, a few hundred parts per million.

If this newly created ocean were left undisturbed for a few hundred million years, its waters would evaporate and be transported onto the continents (in the form of precipitation); streams would transport their loads into it. The sediment created in this ocean would be uplifted and incorporated into the continents. Gradually, the influence of the continental debris would be felt, and the pH might shift slightly. Iron would be oxidized out of the ferrous silicates to produce iron oxides, but the water composition would not vary a great deal.

The primary minerals of igneous rocks are all mildly basic compounds. When they react in excess with acids such as hydrochloric acid and carbon dioxide, they produce neutral or mildly alkaline solutions plus a set of altered aluminosilicate and carbonate reaction products. It is highly unlikely that ocean water has changed through time from a solution approximately in equilibrium with these reaction products, which are clay minerals and carbonates.

The modern oceans. The oceans probably achieved their modern characteristics 2 to 1.5 billion years ago. The chemical and mineralogical compositions and the relative proportions of sedimentary rocks of this age differ little from their Paleozoic counterparts (those dating from about 570 to 245 million years ago). The fact that the acid sulfur gases had been neutralized to sulfate by this time is borne out by calcium sulfate deposits of late Precambrian age (roughly 570 million to 1.6 billion years old). Chemically precipitated ferric oxides in late Precambrian sedimentary rocks indicate available free oxygen, whatever its percentage. The chemistry and mineralogy of middle and late Precambrian shales is similar to that of Paleozoic shales. Thus, it appears that continuous cycling of sediments like those of the present time has occurred for 1.5 to 2 billion years and that these sediments have controlled oceanic composition.

Salt content of the oceans

It was once thought that the saltiness of the modern oceans simply represents the storage of salts derived from rock weathering and transported to the oceans by fluvial processes. With increasing knowledge of the age of the Earth, however, it was realized that, at the present-day rate of delivery of salts to the oceans or even at much reduced rates, the total salt content and the mass of individual salts in the oceans could be attained in geologically short-time intervals compared to the Earth's age. The total mass of salt in the oceans can be accounted for at present-day rates of stream delivery in about 12 million years. The mass of dissolved silica in ocean water can be doubled in only 20,000 years by addition of stream-derived silica; to double sodium would take 70 million years. It then became apparent that the oceans were not simply an accumulator of salts, but as water evaporated from the oceans, along with some salt, the introduced salts must

be removed in the form of minerals. Thus, the concept of the oceans as a chemical system changed from that of a simple accumulator to that of a steady-state system in which rates of inflow of materials into the oceans equal rates of outflow. The steady-state concept permits influx to vary with time, but it would be matched by nearly simultaneous and equal variation of efflux. Calculations of rates of addition of elements to the oceanic system and removal from it show that for at least 100 million years the oceanic system has been in a steady state with approximately fixed rates of major element inflow and outflow and, thus, fixed chemical composition. (F.T.M./R.H.By.)

PHYSICAL PROPERTIES OF SEAWATER

Water is a unique substance. Not only is water the most abundant substance at the Earth's surface, but it also has the most naturally occurring physical states of any Earth material or substance (solid, liquid, and gas) and the greatest capacity to do things without being altered significantly. It is essential for sustaining life on Earth and affects the physical environment in a myriad of ways, as evidenced by the sculpting of landscape features by moving water, the maintaining of the Earth's radiation balance by atmospheric water vapour transfer, and the transporting of inorganic and organic materials about the planet's surface by the oceans. The addition of salt to water changes the behaviour of water only slightly.

Salinity distribution. A discussion of salinity, the salt content of the oceans, requires an understanding of two important concepts: (1) the present-day oceans are considered to be in steady state, receiving as much salt as they lose (see above), and (2) the oceans have been mixed over such a long time period that the composition of sea salt is everywhere the same in the open ocean. This uniformity of salt content results in oceans in which the salinity varies little over space or time.

The range of salinity observed in the open ocean is from 33 to 37 grams of salt per kilogram of seawater or parts per thousand (‰). For the most part, the observed departure from a mean value of approximately 35‰ is caused by processes at the Earth's surface that locally add or remove fresh water. Regions of high evaporation have elevated surface salinities, while regions of high precipitation have depressed surface salinities. In nearshore regions close to large freshwater sources, the salinity may be lowered by dilution. This is especially true in areas where the region of the ocean receiving the fresh water is isolated from the open ocean by the geography of the land.

Areas of the Baltic Sea may have salinity values depressed to 10‰ or less. Increased salinity by evaporation is accentuated where isolation of the water occurs. This effect is found in the Red Sea, where the surface salinity rises to 41‰. Coastal lagoon salinities in areas of high evaporation may be much higher. The removal of fresh water by evaporation or the addition of fresh water by precipitation does not affect the constancy of composition of the sea salt in the open sea. A river draining a particular soil type, however, may bring to the oceans only certain salts that will locally alter the salt composition. In areas of high evaporation where the salinity is driven to very high values, precipitation of particular salts may alter the composition too. At high latitudes where sea ice forms seasonally, the salinity of the seawater is elevated during ice formation and reduced when the ice melts.

Effect of submarine volcanism on salinity

At depth in the oceans, salinity may be altered as seawater percolates into fissures associated with deep-ocean ridges and crustal rifts involving volcanism. This water then returns to the ocean as superheated water carrying dissolved salts from the magmatic material within the crust. It may lose much of its dissolved load to precipitates on the seafloor and gradually blend in with the surrounding seawater, sharing its remaining dissolved substances.

Salt concentrations as high as 256‰ have been found in hot but dense pools of brine trapped in depressions at the bottom of the Red Sea. The composition of the salts in these pools is not the same as the sea salt of the open oceans.

The salinities of the open oceans found at the greater depths are quite uniform in both time and space with average values of 34.5 to 35‰. These salinities are determined by surface processes such as those described above when the water, now at depth, was last in contact with the surface.

The average distribution of the surface salinity of the open oceans is depicted in Figure 5. This figure shows the response of the salinity of the ocean surface to the latitudinal variation in exchange of water between the oceans and the atmosphere. It also shows the impact of major ocean currents that displace surface water from one latitudinal zone to another. The northward displacement of subtropical water of higher salinity by the Gulf Stream and the North Atlantic Current is evident.

The intertropical convergence, with its high precipitation centred about 5° N, supports the tropical rain forests of the world and leaves its imprint on the oceans as a latitudinal depression of surface salinity. At approximately 30°–35° N and 30°–35° S, the subtropical zones called the horse latitudes are belts of high evaporation that produce major deserts and grasslands on the continents and cause the surface salinity to rise. At 50°–60° N and 50°–60° S, precipitation again increases. The general trend of mid-ocean salinity with latitude follows the trend of observed evaporation and precipitation shown in Figure 6.

Temperature distribution. Mid-ocean surface temperatures vary with latitude in response to the balance between incoming solar radiation and outgoing long-wave radiation. There is an excess of incoming solar radiation at latitudes less than approximately 45° and an excess of radiation loss at latitudes higher than approximately 45°. Superimposed on this radiation balance are seasonal changes in the intensity of solar radiation and the duration of daylight hours due to the tilt of the Earth's axis to the plane of the ecliptic and the rotation of the planet about this axis. The combined effect of these variables is that average ocean surface temperatures are higher at low latitudes than at high latitudes. Because the Sun, with respect to the Earth, migrates annually between the tropic of Cancer and the tropic of Capricorn, the yearly change in heating of the Earth's surface is small at low latitudes and large at mid- and higher latitudes.

Water has an extremely high heat capacity, and heat is mixed downward during summer surface-heating conditions and upward during winter surface cooling. This heat transfer reduces the actual change in ocean surface temperatures over the annual cycle. In the tropics the ocean surface is warm year-round, varying seasonally about 1° to 2° C. At mid-latitudes the mid-ocean temperatures vary about 8° C over the year. At the polar latitudes the surface temperature remains near the ice point of seawater—about −1.9° C.

Land temperatures have a large annual range at high latitudes because of the low heat capacity of the land surface. Figure 7 shows the average zonal temperature of the open oceans and land, as well as annual temperature ranges. Proximity to land, isolation of water from the open ocean, and processes that control stability of the surface water combine to increase the annual range of nearshore ocean surface temperature.

Effects of the wind on ocean surface temperature

In winter, prevailing winds carry cold air masses off the continents in temperate and subarctic latitudes, cooling the adjacent surface seawater below that of the mid-ocean level. In summer, the opposite effect occurs, as warm continental air masses move out over the adjacent sea. This creates a greater annual range in sea surface temperatures at mid-latitudes on the western sides of the oceans of the Northern Hemisphere but has only a small effect in the Southern Hemisphere as there is little land present. Instead, the oceans of the Southern Hemisphere act to control the air temperature, which in turn influences the land temperatures of the temperate zone and reduces the annual temperature range over the land.

Currents carry water having the characteristics of one latitudinal zone to another zone. The northward displacement of warm water to higher latitudes by the Gulf Stream of the North Atlantic and the Kuroshio (Japan Current) of the North Pacific creates sharp changes in temperature along the current boundaries or thermal fronts, where these northward-moving flows meet colder water flowing

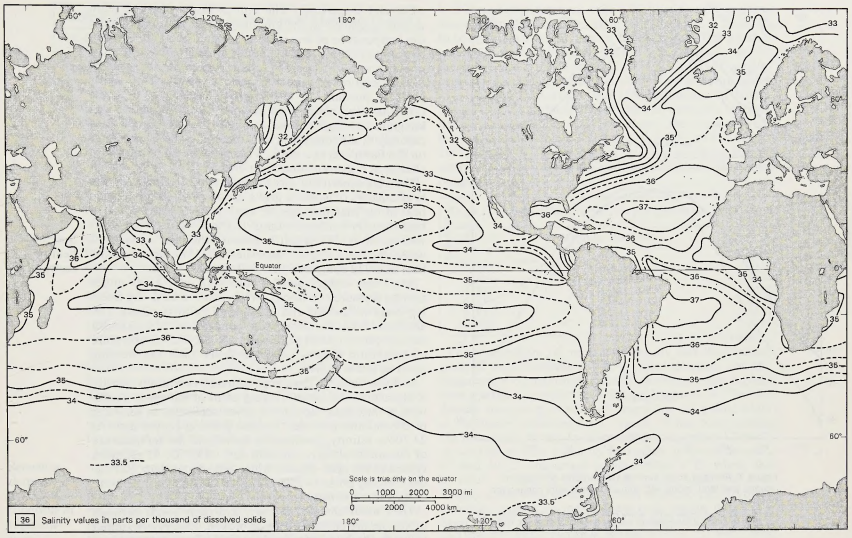


Figure 5: Salinity distribution in surface waters of the world's oceans.

From H.U. Sverdrup, Martin W. Johnson, and Richard H. Fleming, *The Oceans: Their Physics, Chemistry, and General Biology*, © 1942, renewed 1970; Prentice-Hall, Inc., Englewood Cliffs, New Jersey

southward from higher latitudes. Cold water currents flowing from higher to lower latitudes also displace surface isotherms from near constant latitudinal positions (see Figure 8). At low latitudes the trade winds act to move water away from the lee coasts of the landmasses to produce areas of coastal upwelling of water from depth and reduce surface temperatures.

As adapted in *The Encyclopedia of Oceanography*, edited by Rhodes W. Fairbridge, © 1966 by Litton Educational Publishing, Inc., by permission of Van Nostrand Reinhold Company, from A. Defant, *Physical Oceanography* (© 1961); Pergamon Press Ltd., reprinted with permission

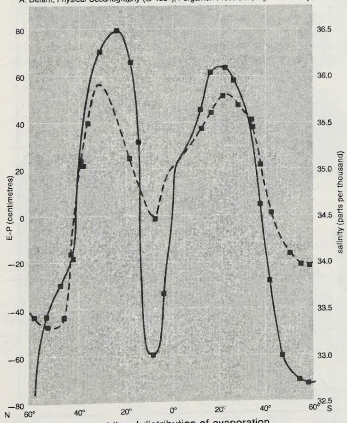


Figure 6: Mean meridional distribution of evaporation precipitation (E-P; solid line) and surface salinity (dashed line) for the entire ocean.

Temperatures in the oceans decrease with increasing depth. There are no seasonal changes at the greater depths. The temperature range extends from 30° C at the sea surface to −1° C at the seabed. Like salinity, the temperature at depth is determined by the conditions that the water encountered when it was last at the surface. In the low latitudes the temperature change from top to bottom in the oceans is large. In high temperate and Arctic regions, the formation of dense water at the surface that sinks to depth produces nearly isothermal conditions with depth.

Areas of the oceans that experience an annual change in surface heating have a shallow wind-mixed layer of elevated temperature in the summer. Below this nearly isothermal layer 10 to 20 metres thick, the temperature decreases rapidly with depth, forming a shallow seasonal thermocline (*i.e.*, layer of sharp vertical temperature change). During winter cooling and increased wind mixing at the ocean surface, convective overturning and mixing erase this shallow thermocline and deepen the isothermal layer. The seasonal thermocline re-forms when summer returns. At greater depths, a weaker nonseasonal thermocline is found separating water from temperate and subpolar sources.

Below this permanent thermocline, temperatures decrease slowly. In the very deep ocean basins, the temperature may be observed to increase slightly with depth. This occurs when the deepest parts of the oceans are filled by water with a single temperature from a common source. This water experiences an adiabatic temperature rise as it sinks. Such a temperature rise does not make the water column unstable because the increased temperature is caused by compression, which increases the density of the water. For example, surface seawater of 2° C sinking to a depth of 10,000 metres increases its temperature by about 1.3° C. When measuring deep-sea temperatures, the

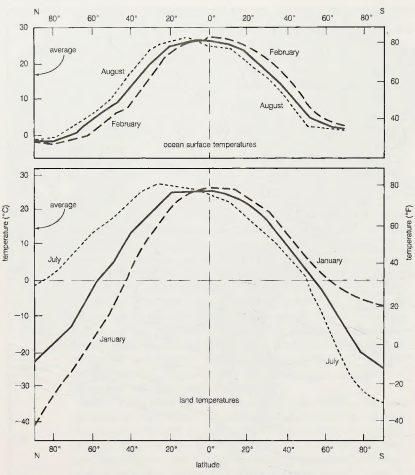


Figure 7: Average zonal surface temperature of the open oceans and land, along with annual temperature ranges for both.

From M. Grant Gross, *Oceanography: A View of the Earth*, 3rd ed. (1982), Prentice-Hall, Inc., Englewood Cliffs, N.J.; (top) after G. Wust, W. Brogmus, and E. Noodt, *Kieler Meeresforschungen*, vol. 10 (1954), (bottom) data from *Smithsonian Physical Tables* (1964)

adiabatic temperature rise, which is a function of salinity, initial temperature, and pressure change, is calculated and subtracted from the observed temperature to obtain the potential temperature. Potential temperatures are used to identify a common type of water and to trace this water back to its source.

Thermal properties. The unit of heat called the gram calorie is defined as the amount of heat required to raise the temperature of one gram of water 1° C. The kilocalorie, or food calorie, is the amount of heat required to raise one kilogram of water 1° C. Heat capacity is the amount of heat required to raise one gram of material 1° C under constant pressure. In the International System of Units (SI), the heat capacity of water is one kilocalorie per kilogram per degree Celsius. Water has the highest heat capacity of all common Earth materials; therefore, water on the Earth acts as a thermal buffer, resisting temperature change as it gains or loses heat energy.

The heat capacity of any material can be divided by the heat capacity of water to give a ratio known as the specific heat of the material. Specific heat is numerically equal to heat capacity but has no units. In other words, it is a ratio without units. When salt is present, the heat capacity of water decreases slightly. Seawater of 35‰ has a specific heat of 0.932 compared to 1.000 for pure water.

Specific heat of seawater

Pure water freezes at 0° C and boils at 100° C under normal pressure conditions. When salt is added, the freezing point is lowered and the boiling point is raised. The addition of salt also lowers the temperature of maximum density below that of pure water (4° C). The temperature of maximum density decreases faster than the freezing point as salt is added.

At 30‰ salinity, the temperature of maximum density is lower than the initial freezing point of saltwater. Therefore, a maximum density is never achieved, as seawater of this salinity is cooled because freezing occurs first. At 24.70‰ salinity, the freezing point and the temperature of maximum density coincide at −1.332° C. At salinities typical of the open oceans, which are greater than 24.7‰, the freezing point is always higher than the temperature of maximum density.

When water changes its state, hydrogen bonds between molecules are either formed or broken. Energy is required to break the hydrogen bonds, which allows water to pass from a solid to a liquid state or from a liquid to a gaseous state. When hydrogen bonds are formed, permitting water to change from a liquid to a solid or from a gas to a liquid, energy is liberated. The heat energy input required to change water from a solid at 0° C to a liquid at 0° C is the latent heat of fusion and is 80 calories per gram of ice.

Adapted from A.C. Duxbury & A.B. Duxbury, *An Introduction to the World's Oceans* (1989), Wm C. Brown Group, Dubuque, Iowa

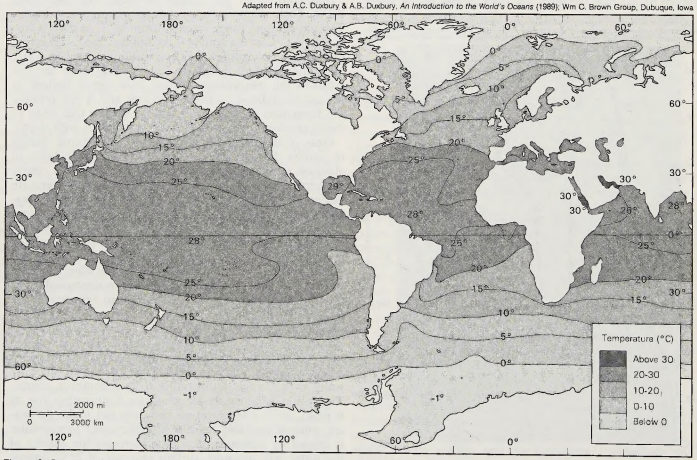


Figure 8: Ocean surface temperatures (in Celsius) during summer in the Northern Hemisphere. The isotherms (lines of equal temperature) in general parallel the equator except in a few areas where ocean currents cause marked displacements.

Table 4: Effect of Salinity on Temperature of Maximum Density

Salinity ‰	0	10	20	30	35
Freezing point	0.00	−0.53	−1.08	−1.63	−1.91
Temperature of maximum density	4.0	1.92	−0.20	−2.40	−3.42

Water's latent heat of fusion is the highest of all common materials. Because of this, heat is released when ice forms and is absorbed during melting, which tends to buffer air temperatures as land and sea ice form and melt seasonally.

When water converts from a liquid to a gas, a quantity of heat energy known as the latent heat of vaporization is required to break the hydrogen bonds. At 100° C, 540 calories per gram of water are needed to convert one gram of liquid water to one gram of water vapour under normal pressure. Water can evaporate at temperatures below the boiling point, and ice can evaporate into a gas without first melting in a process called sublimation. Evaporation below 100° C and sublimation require more energy per gram than 540 calories. At 20° C about 585 calories are required to vaporize one gram of water. When water vapour condenses back to liquid water, the latent heat of vaporization is liberated. The evaporation of water from the surface of the Earth and its condensation in the atmosphere constitute the single most important way that heat from the Earth's surface is transferred to the atmosphere. This process is the source of the power that drives hurricanes and a principal mechanism for cooling the surface of the oceans. The latent heat of vaporization of water is the highest of all common substances.

Seawater density as a function of temperature, salinity, and pressure

Density of seawater and pressure. The density of a material is given in units of mass per unit volume and expressed in kilograms per cubic metre in the SI system of units. In oceanography the density of seawater has been expressed historically in grams per cubic centimetre. The density of seawater is a function of temperature, salinity, and pressure. Because oceanographers require density measurements to be accurate to the fifth decimal place, manipulation of the data requires writing many numbers to record each measurement. Also, the pressure effect can be neglected in many instances by using potential temperature. These two factors led oceanographers to adopt a density unit called sigma-*t* (σ_t). This value is obtained by subtracting 1.0 from the density and multiplying the remainder by 1,000. The σ_t has no units and is an abbreviated density of seawater controlled by salinity and temperature only.

The σ_t of seawater increases with increasing salinity and decreasing temperature. Table 5 demonstrates how density expressed as σ_t changes with both salinity and temperature under conditions of normal atmospheric pressure. Seawater of 35‰ and 5° C in the Table has a density of 1.02770 grams per cubic centimetre (g/cm³).

The relationship between pressure and density is demonstrated by observing the effect of pressure on the density of seawater at 35‰ and 0° C (Table 6). Because a one-metre column of seawater produces a pressure of about one decibar (0.1 atmosphere), the pressure in decibars is approximately equal to the depth in metres. (One decibar is one-tenth of a bar, which in turn is equal to 10^5 newtons per square metre.)

Compressibility of seawater under high pressure

Increasing density values demonstrate the compressibility of seawater under the tremendous pressures present in the deep ocean. If seawater were incompressible, each cubic centimetre of water in the water column would expand, and density values at all depths would be equal in Table 6. If the average pressure over 4,000 metres (the approximate mean depth of the ocean) is calculated, it is found to be approximated by that at 2,000 metres. The average volume change due to pressure for each gram of water in the entire water column is (1/1.02813–1/1.03747) cm³/g, or 0.00876 cm³/g. Because the number of grams of water in a column of seawater 4×10^5 centimetres in length is equal to the number of centimetres times the average density of the water, 1.03747 g/cm³, the expansion of the entire water column is about 4×10^5 cm $\times$ 0.00876 cm³/g $\times$ 1.03747 g/cm³, or an average sea level rise of about 36 metres if the area of the oceans is considered constant.

Table 5: Sigma-*t* (σ_t) Values of Seawater

Salinity ‰	5	10	20	30	35
Temperature (°C)					
0	3.97	8.01	16.07	24.10	28.13
5	4.01	7.97	15.86	23.74	27.70
10	3.67	7.56	15.32	23.08	26.97
15	3.01	6.85	14.50	22.15	25.99
20	2.07	5.86	13.42	20.98	24.78
25	0.87	4.62	12.10	19.60	23.36
30	−0.57	3.15	10.57	18.01	21.75

Table 6: Density Changes with Depth (Seawater 35‰ and 0° C)

depth (m)	pressure (decibars)	density (g/cm³)
0	0	1.02813
1,000	1,000	1.03285
2,000	2,000	1.03747
4,000	4,000	1.04640
6,000	6,000	1.05495
8,000	8,000	1.06315
10,000	10,000	1.07104

The temperature of maximum density and the freezing point of water decrease as salt is added to water, and the temperature of maximum density decreases more rapidly than the freezing point. At salinities less than 24.7‰ the density maximum is reached before the ice point, while at the higher salinities more typical of the open oceans the maximum density is never achieved naturally. Table 5 shows that at 5‰ a density maximum is found between 0° and 10° C. (Its actual position is at 3° C, where the σ_t value is 4.04 for 5‰ salinity.) This ability of low-salinity water and, of course, fresh water to pass through a density maximum makes them both behave differently from marine systems when water is cooled at the surface and density-driven overturn occurs.

During the fall a lake is cooled at its surface, the surface water sinks, and convective overturn proceeds as the density of the surface water increases with the decreasing temperature. By the time the surface water reaches 4° C, the temperature of maximum density for fresh water, the density-driven convective overturn has reached the bottom of the lake, and overturn ceases. Further cooling of the surface produces less dense water, and the lake becomes stably stratified with regard to temperature-controlled density. Only a relatively shallow surface layer is cooled below 4° C. When this surface layer is cooled to the ice point, 0° C, ice is formed as the latent heat of fusion is extracted. In a deep lake the temperature at depth remains at 4° C. In the spring the surface water warms up and the ice melts. A shallow convective overturn resumes until the lake is once more isothermal at 4° C. Continued warming of the surface produces a stable water column.

In seawater in which the salinity exceeds 24.7‰, convective overturn also occurs during the cooling cycle and penetrates to a depth determined by the salinity and temperature-controlled density of the cooled water. Since no density maximum is passed, the thermally driven convective overturn is continuous until the ice point is reached where sea ice forms with the extraction of the latent heat of fusion. Since salt is largely excluded from the ice in most cases, the salinity of the water beneath the ice increases slightly and a convective overturn that is both salt- and temperature-driven continues as sea ice forms.

The continuing overturn requires that a large volume of water be cooled to a new ice point dictated by the salinity increase before additional ice forms. In this manner, very dense seawater that is both cold and of elevated salinity is formed. Such areas as the Weddell Sea in Antarctica produce the densest water of the oceans. This water, known as Antarctic Bottom Water, sinks to the deepest depths of the oceans. The continuing overturn slows the rate at which the sea ice forms, limiting the seasonal thickness of the ice. Other factors that control the thickness of ice are the rate at which heat is conducted through the ice layer and the insulation provided by snow on the ice. Seasonal sea ice seldom exceeds about two metres in thickness. During

Seawater of extremely high density

the warmer season, melting sea ice supplies a freshwater layer to the sea surface and thereby stabilizes the water column (see below *Ice in the sea*).

Surface processes that alter the temperature and salinity of seawater drive the vertical circulation of the oceans. Known as thermohaline circulation, it continually replaces seawater at depth with water from the surface and slowly replaces surface water elsewhere with water rising from deeper depths (see below *Circulation of the ocean currents: Thermohaline circulation*).

Optical properties. Water is transparent to the wavelengths of electromagnetic radiation that fall within the visible spectrum and is opaque to wavelengths above and below this band. However, once in the water, visible light is subject to both refraction and attenuation.

Light rays that enter the water at any angle other than a right angle are refracted (*i.e.*, bent) because the light waves travel at a slower speed in water than they do in air. The amount of refraction, referred to as the refractive index, is affected by both the salinity and temperature of the water. The refractive index increases with increasing salinity and decreasing temperature. This relationship allows the refractive index of a sample of seawater at a constant temperature to be used to determine the salinity of the sample.

Some of the Sun's radiant energy is reflected at the ocean surface and does not enter the ocean. That which penetrates the water's surface is attenuated by absorption and conversion to other forms of energy, such as heat that warms or evaporates water, or is used by plants to fuel photosynthesis. Sunlight that is not absorbed can be scattered by molecules and particulates suspended in the water. Scattered light is deflected into new directional paths and may wander randomly to eventually be either absorbed or directed upward and out of the water. It is this upward scattered light and the light reflected from particles that determine the colour of the oceans, as seen from above.

Water molecules, dissolved salts, organic substances, and suspended particulates combine to cause the intensity of available solar radiation to decrease with depth. Observations of light attenuation in ocean waters indicate that not only does the intensity of solar radiation decrease with depth but also the wavelengths present in the solar spectrum are not attenuated at the same rates. Figure 9 shows the distribution of light intensity with wavelength at the ocean surface and at depth. It is apparent that both short wavelengths (ultraviolet) and long wavelengths (infrared) are absorbed rapidly and are not available for scattering. Only blue-green wavelengths penetrate to any depth, and because the blue-green light is most available for scattering, the oceans appear blue to the human eye. Changes in the colour of the ocean waters are caused either by the colour of the particulates in suspension and dissolved substances or by the changing quality of the solar radiation at the ocean surface as determined by the angle of the Sun and atmospheric conditions. In the clearest ocean waters only about 1 percent of the surface radiation remains at a depth of 150 metres. No sunlight penetrates below 1,000 metres.

Causes of ocean colour changes

There are many ways of measuring light attenuation in the oceans. A common method involves the use of a Secchi disk, a weighted round white disk about 30 centimetres in diameter. The Secchi disk is lowered into the ocean to the depth where it disappears from view; its reflectance equals the intensity of light backscattered from the water. This depth in metres divided into 1.7 yields an attenuation, or extinction, coefficient for available light as averaged over the Secchi disk depth. The light extinction coefficient, x, may then be used in a form of Beer's law, $I_z = I_0 e^{xz}$, to estimate I_z, the intensity of light at depth z from I_0, the intensity of light at the ocean surface. This method gives no indication of the attenuation change with depth or the attenuation of specific wavelengths of light.

A photocell may be lowered into the ocean to measure light intensity at discrete depths and to determine light reduction from the surface value or from the previous depth value. The photocell may sense all available wavelengths or may be equipped with filters that pass only certain

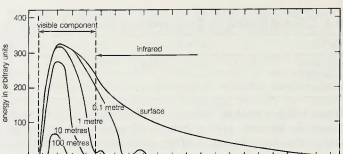


Figure 9: *Variation in the spectral quality of sunlight with depth in the sea.*
The total energy—the area below each curve—decreases with increasing depth, and the colour peak shifts toward the blue side of the system as selective absorption takes place (see text).
From A.C. Duxbury, *The Earth and Its Oceans*, © 1971 A.C. Duxbury, Addison-Wesley Publishing Co. Inc., Reading, Mass., after G. Neumann and W.J. Pierson, *Principles of Physical Oceanography*, © 1966, by permission of the authors

wavelengths of light. Since I_z and I_0 are known, changing light intensity values may be used in Beer's law to determine how the attenuation coefficient changes with depth and quality of light. Measurements of this type are used to determine the level of photosynthesis as a function of radiant energy level with depth and to measure changes in the turbidity of the water caused by particulate distribution with depth.

Different areas of the oceans tend to have different optical properties. Near rivers, silt increases the suspended particle effect. Where nutrients and sunlight are abundant, phytoplankton (unicellular plants) increase the opacity of the water and lend it their colour. Organic substances from excretion and decomposition also have colour and absorb light. Table 7 shows the attenuation of light in different ocean regions with variations in their properties governing scattering and absorption.

Solar radiation received at the ocean surface is constantly changing in time and space. Cloud cover, atmospheric dust, atmospheric gas composition, roughness of the ocean surface, and elevation angle of the Sun combine to change both the quality and quantity of light that enters the ocean. When the Sun's rays are perpendicular to a smooth ocean surface, reflectance is low. When the solar rays are oblique to the ocean surface, reflectance is increased. If the ocean is rough with waves, reflectance is increased when the Sun is at high elevation and decreased when it is at low elevation. Since latitude plays a role in the elevation of the Sun above the horizon, light penetration is always less at the higher latitudes. Cloud cover, density layering, fog, and dust cause refraction and atmospheric scattering of sunlight. When strongly scattered, the Sun's rays are not unidirectional and there are no shadows. Light enters the ocean from all angles under this condition, and the elevation angle of the Sun loses its importance in controlling surface reflectance. Percent of reflectance of direct sunlight related to the Sun's elevation angle is shown in Table 8 for a smooth ocean surface. The solar energy available to penetrate the ocean is 100 percent minus the tabulated reflectance value.

These data indicate that water is a good absorber of solar radiation.

Acoustic properties. Water is an excellent conductor of sound, considerably better than air. The attenuation of sound by absorption and conversion to other energy forms is a function of sound frequency and the properties of water.

Water as an excellent conductor of sound

The attenuation coefficient, x, in Beer's law, as applied to sound, where I_z and I_0 are now sound intensity values, is dependent on the viscosity of water and inversely proportional to the frequency of the sound and the density of the water. High-pitched sounds are absorbed and converted to heat faster than low-pitched sounds. Sound velocity in water is determined by the square root of elasticity divided by the water's density. Because water is only slightly compressible, it has a large value of elasticity and therefore conducts sound rapidly. Since both the elasticity and density of seawater change with temperature, salinity, and pressure, so does the velocity of sound.

Table 7: Loss of Light (Percent) in One Metre of Seawater*

	violet		blue-green			yellow	orange	red
Wavelength (micrometre)	0.30	0.40	0.46	0.50	0.54	0.58	0.64	0.70
Oceanic water, most transparent	16%	4%	2%	3%	5%	9%	29%	42%
Oceanic water, least transparent	57%	16%	11%	10%	13%	19%	36%	55%
Coastal water, average		63%	37%	29%	28%	30%	45%	74%

*According to Jerlov.

In the oceans the speed of sound varies between 1,450 and 1,570 metres per second. It increases about 4.5 metres per second per each degree C increase and 1.3 metres per second per each 1‰ increase in salinity. Increasing pressure also increases the speed of sound at the rate of about 1.7 metres per second for an increase in pressure of 100 metres in depth, which is equal to approximately 10 bars, or 10 atmospheres.

Table 8: Sunlight Reflectance

Sun's elevation angle	90°	50°	40°	30°	20°	10°	5°
Reflectance (%)	3	3	4	6	12	27	42

The greatest changes in temperature and salinity with depth that affect the speed of sound are found near the surface. Changes of sound speed in the horizontal are usually slight except in areas where abrupt boundaries exist between waters of different properties. The effects of salinity and temperature on sound speed are more important than the effect of pressure in the upper layers. Deeper in the ocean, salinity and temperature change less with depth, and pressure becomes the important controlling factor.

In regions of surface dilution, salinity increases with depth near the surface, while in areas of high evaporation salinity decreases with depth. Temperature usually decreases with depth and normally exerts a greater influence on sound speed than does the salinity in the surface layer of the open oceans. In the case of surface dilution, salinity and temperature effects on the speed of sound oppose each other, while in the case of evaporation they reinforce each other, causing the speed of sound to decrease with depth. Beneath the upper oceanic layers the speed of sound increases with depth.

If a sound wave (sonic pulse) travels at a right angle to these layers, as in depth sounding, no refraction occurs; however, the speed changes continuously with depth, and an average sound speed for the entire water column must be used to determine the depth of water. Variations in the speed of sound cause sound waves to refract when they travel obliquely through layers of water that have different properties of salinity and temperature. Sound waves traveling downward and moving obliquely to the water layers will bend upward when the speed of sound increases with depth and downward when the speed decreases with depth. This refraction of the sound is important in the sonar detection of submarines because the actual path of a sound wave must be known to determine a submarine's position relative to the transmitter of the sound. Refraction also produces shadow zones that sound waves do not penetrate because of their curvature (see Figure 10A and 10B).

At depths of approximately 1,000 metres, pressure becomes the important factor: it combines with temperature and salinity to produce a zone of minimum sound speed. This zone has been named the SOFAR (sound fixing and ranging) channel (see Figure 11). If a sound is generated by a point source in the SOFAR zone, it becomes trapped by refraction. Dispersed horizontally rather than in three directions, the sound is able to travel for great distances. Hydrophones lowered to this depth many kilometres from the origin of the sound are able to detect the sound pulse. The difference in arrival time of the pulse at separate listening posts may be used to triangulate the position of the pulse source.

SOFAR channel

Hearing is an important sensory mechanism for marine animals because seawater is more transparent to sound than to light. Animals communicate with each other over long distances and also locate objects by sending directional sound signals that reflect from targets and are received as echoes. Information about the size of a target is gained by varying the frequency of the sound; high-frequency (or short-wavelength) sound waves reflect better from small targets than low-frequency sound waves. The intensity and quality of the returning signal also provide information about the properties of the reflecting target.

(A.C.D.)

Ice in the sea. Formation of sea ice was briefly discussed above (see *Density of seawater and pressure*). Sea ice formation is a thermal physical property of water and plays a role in driving convective overturn in the oceans. It does so by increasing the density of the seawater under the forming ice and thereby helps to drive convective overturn.

There are two types of ice in the seas: sea ice, which is ice formed by the freezing of seawater, and ice that has come from land, such as icebergs and ice islands.

Sea ice. From an initial stage of so-called frazil crystals (floating needles and platelets) and sludge composed of them, sea ice grows to a compact aggregate of crystals of pure ice with pockets of seawater entrapped between them. Because of this composition, the salinity of sea ice is lower than that of the seawater from which it has grown. The initial sea-ice salinity may vary between 2 and 20 parts per thousand; the more rapid the freezing, the saltier the ice, as brine can be trapped in cavities in the forming ice and become isolated from the seawater.

After sea ice has formed, a process of salt removal by drainage of part of the enclosed brine sets in, because the cells in which it is contained are not completely isolated.

From A.C. Duxbury, *The Earth and Its Oceans*, © 1971 A.C. Duxbury, Addison-Wesley Publishing Co. Inc., Reading, Mass., after G. Neumann and W.J. Pierson, *Principles of Physical Oceanography*, © 1966, by permission of the authors.

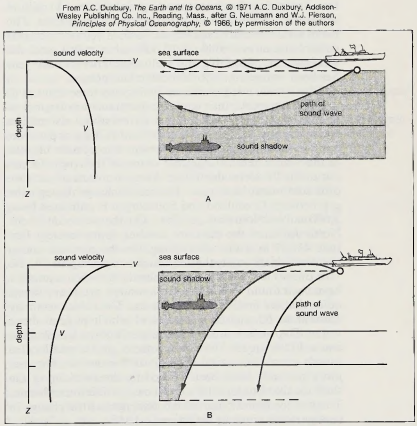


Figure 10: *Refraction of sound resulting from differences in sound speed with depth.*
(A) Sound waves traveling downward and obliquely to the water bend upward when their velocity, *V*, increases with depth, *Z*. (B) They bend downward when the velocity decreases with depth. Refraction of sound waves gives rise to shadow zones (see text).

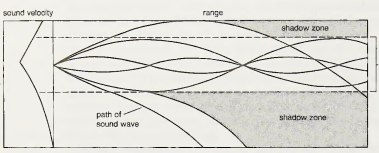


Figure 11: *Zone of minimum sound speed.* Sound waves produced in this zone, known as the SOFAR channel, may become trapped by refraction and travel horizontally for considerable distances.

From A.C. Duxbury, *The Earth and its Oceans*, © 1971 A.C. Duxbury, Addison-Wesley Publishing Co. Inc., Reading, Mass., after L.E. Kinsler and A.R. Frey, *Fundamentals of Acoustics*, 2nd ed., © 1962 John Wiley & Sons, Inc., reprinted by permission of John Wiley & Sons, Inc.

Old ice has very low salinity, on the order of 1 part per thousand or less.

Growth rate of sea ice

The growth rate of sea ice depends on surface temperature, the depth of snow cover, and the heat flux in the underlying water. In the central Arctic, the thickness of an ice cover formed in one growing season is about two metres. If the ice is not broken up or melted each season, it finally reaches an equilibrium thickness of about three to four metres in five to eight years, when the annual ablation (loss by any means) at the top and the bottom equals the annual growth. In the Antarctic, perennial sea ice is found only in the Weddell Sea and a narrow strip around the continent. Most of the Antarctic sea ice is seasonal and reaches a thickness of about 1.5 metres by the end of October.

The high albedo (or reflectivity) of sea ice and its snow cover (80 percent, compared to 5–10 percent for liquid water), the insulation characteristics of ice and snow, and the latent heat of fusion combine to affect the heat budget of the oceans during both freezing and thawing.

The boundaries of the sea ice are highly variable. In the Norwegian and Greenland seas, deviations of 300 kilometres north or south of the average position are not uncommon. The estimated mean areas of sea ice at the end of the summer and at the end of the winter in the Arctic are 9 million square kilometres (3.5 million square miles) and 12 million square kilometres, respectively. In the Antarctic, the corresponding values are 4 million square kilometres and 20 million square kilometres. The mean total volume of sea ice on Earth is 40,000 to 50,000 cubic kilometres (9,600 to 12,000 cubic miles), and the total amount of freezing and melting that occurs each year has been estimated at 30,000 cubic kilometres.

In the Arctic, it is possible to distinguish three regimes of sea ice: the great inner core, the permanent polar cap of sea ice (the Arctic pack), which covers about six million square kilometres; around this the true drift ice or pack ice; and the landfast ice, which is present during nine months of the year, when it fringes the shores of the Arctic Ocean out to the 22-metre depth line. Large amounts of pack ice drift southward each year. The ice discharge through the gap between Greenland and Spitsbergen is estimated to be 3,000 cubic kilometres per year. On the west side of the North Atlantic, the pack ice reaches approximately latitude 45° N in winter and spring. On the east side, along the Norwegian coast, the sea remains open up to 73° N.

Ice islands and icebergs. Ice islands, of which a number have been found drifting in Arctic waters, are heavy sheets of ice that are far thicker than sea ice. Their thickness may amount to 50 metres, 5 metres of which project above water. The surface area of the largest known ice island is about 1,000 square kilometres; others are far smaller. Ice islands consist of a kind of glacierlike snow ice. The majority probably have been formed by the breaking of the shelf ice that borders the north coast of Ellesmere Island. The first ice island reported has undergone little change in configuration since its detection in 1946.

Principal sources of icebergs

Icebergs are formed by the calving (detaching of parts) of glaciers or of inland ice that reaches the sea. The main sources of icebergs in the northern seas are the valley glaciers of Greenland, which produce some 12,000 to 15,000 sizable icebergs annually. Almost as many are calved by the glaciers reaching the sea on the eastern seaboard as by those on the west coast, but the icebergs deriving from the east side do not travel much farther south than Cape Farvel, the southern tip of Greenland. The icebergs of the west coast, on the other hand, after traveling northward and across to the other side of Baffin Bay, are carried far south, along Baffin Island and Labrador, by the Labrador Current. It is estimated that about 1 in every 20 icebergs derived from west Greenland ends up south of Newfoundland (48° N), the greatest numbers arriving there in April, May, and June.

The icebergs of the Antarctic derive from an ice barrier, or shelf ice, a layer of ice that stretches out from the inland ice into the ocean. It rests on the bottom near shore, but farther out to sea it floats on the water. Because of their origin, the Antarctic icebergs are much longer than they are high, occasionally measuring some tens of kilometres in length. For this reason they are called table bergs.

The frequency with which icebergs occur in the Southern Ocean (see above) does not vary much with the season in contrast to the North Atlantic occurrences. Generally speaking, October and November are the months in which they are most numerous in the south because of the release of the bergs from the pack ice in the southern spring. They reach farthest north from November to February. The average northern boundary for icebergs is about 40° S in the Atlantic Ocean, between 40° and 50° S in the Indian Ocean, and about 50° S in the Pacific. At least several thousands of them are adrift every year in the southern seas. (P.Gr./A.C.D.)

Circulation of the ocean waters

GENERAL OBSERVATIONS

The general circulation of the oceans defines the average movement of seawater, which, like the atmosphere, follows a specific pattern. Superimposed on this pattern are oscillations of tides and waves, which are not considered part of the general circulation. There also are meanders and eddies that represent temporal variations of the general circulation. The ocean circulation pattern exchanges water of varying characteristics, such as temperature and salinity, within the interconnected network of oceans and is an important part of the heat and freshwater fluxes of the global climate. Horizontal movements are called currents, which range in magnitude from a few centimetres per second to as much as 4 metres per second. A characteristic surface speed is about 5 to 50 centimetres per second. Currents diminish in intensity with increasing depth. Vertical movements, often referred to as upwelling and downwelling, exhibit much lower speeds, amounting to only a few metres per month. As seawater is nearly incompressible, vertical movements are associated with regions of convergence and divergence in the horizontal flow patterns.

Types of ocean circulation

Ocean circulation derives its energy at the sea surface from two sources that define two circulation types: (1) wind-driven circulation forced by wind stress on the sea surface, inducing a momentum exchange, and (2) thermohaline circulation driven by the variations in water density imposed at the sea surface by exchange of ocean heat and water with the atmosphere, inducing a buoyancy exchange. These two circulation types are not fully independent, since the sea-air buoyancy and momentum exchange are dependent on wind speed. The wind-driven circulation is the more vigorous of the two and is configured as large gyres that dominate an ocean region. The wind-driven circulation is strongest in the surface layer. The thermohaline circulation is more sluggish, with a typical speed of one centimetre per second, but this flow extends to the seafloor and forms circulation patterns that envelop the global ocean.

DISTRIBUTION OF OCEAN CURRENTS

Maps of the general circulation at the sea surface, such as the one shown in Figure 12, are constructed from a vast amount of data obtained from inspecting the residual drift of ships after course direction and speed are accounted for in a process called dead reckoning. This information is

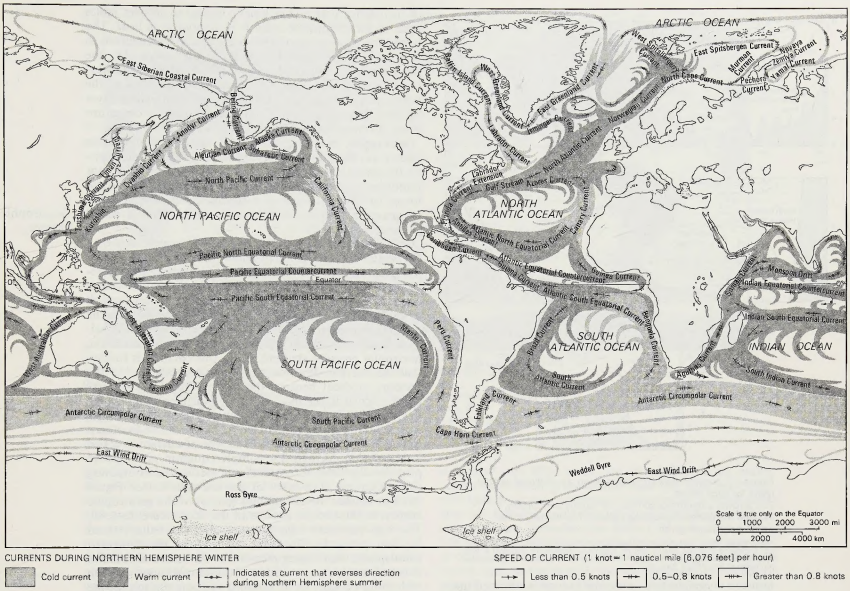


Figure 12: Major ocean-current systems of the world.

amplified by satellite-tracked drifters at sea. The pattern is nearly entirely that of wind-driven circulation.

Deep-ocean circulation (Figure 13) consists mainly of thermohaline circulation. The currents are inferred from the distribution of seawater properties, which trace the spreading of specific water masses. The distribution of density or field of mass is also used to estimate the deep currents. Direct observations of subsurface currents are made by deploying current meters from bottom-anchored moorings and by setting out neutral buoyant instruments whose drift at depth is tracked acoustically.

CAUSES OF OCEAN CURRENTS

The general circulation is governed by the equation of motion, one of Sir Isaac Newton's fundamental laws of mechanics applied to a continuous volume of water. This equation states that the product of mass and current acceleration equals the vector sum of all forces that act on the mass. Besides gravity, the most important forces that cause and affect ocean currents are horizontal pressure-gradient forces, Coriolis forces, and frictional forces. Temporal and inertial terms are generally of secondary importance to the general flow, though they become important for transient features of meanders and eddies.

Pressure gradients. The hydrostatic pressure, p, at any depth below the sea surface is given by the equation $p = gpz$, where g is the acceleration of gravity, ρ is the density of seawater, which increases with depth, and z is the depth below the sea surface. This is called the hydrostatic equation, which is a good approximation for the equation of motion for forces acting along the vertical. Horizontal differences in density (due to variations of temperature and salinity) measured along a specific depth cause the hydrostatic pressure to vary along a horizontal plane or geopotential surface, a surface perpendicular to the direction of the gravity acceleration. Horizontal gradients of pressure, though much smaller than vertical changes in pressure, give rise to ocean currents.

In a homogeneous ocean, which would have a constant potential density, horizontal pressure differences are possible only if the sea surface is tilted. In this case, surfaces of equal pressure, called isobaric surfaces, are tilted in the deeper layers by the same amount as the sea surface. This is referred to as the barotropic field of mass. The unchanged pressure gradient gives rise to a current speed independent of depth. The oceans of the world, however, are not homogeneous. Horizontal variations in temperature and salinity cause the horizontal pressure gradient to vary with depth. This is the baroclinic field of mass, which leads to currents that vary with depth. The horizontal pressure gradient in the ocean is a combination of these two mass fields.

Barotropic and baroclinic fields of mass

The tilt, or topographic relief, of the isobaric surface marking sea surface (defined as $p = 0$) can be constructed from a three-dimensional density distribution (Figure 14) using the hydrostatic equation. Since the absolute value of pressure is not known at any depth in the ocean, the sea surface slope is presented relative to that of a deep isobaric surface; it is assumed that the deep isobaric surface is level. Since the wind-driven circulation attenuates with increasing depth, an associated decrease of isobaric tilt with increasing depth is expected. Representation of the sea surface relief relative to a deep reference surface is a good representation of the absolute shape of the sea surface. The total relief of the sea surface amounts to about two metres, with "hills" in the subtropics and "valleys" in the polar regions. This pressure head drives the surface circulation.

Coriolis effect. The rotation of the Earth about its axis causes moving particles to behave in a way that can only be understood by adding a rotational dependent force. To an observer in space, a moving body would continue to

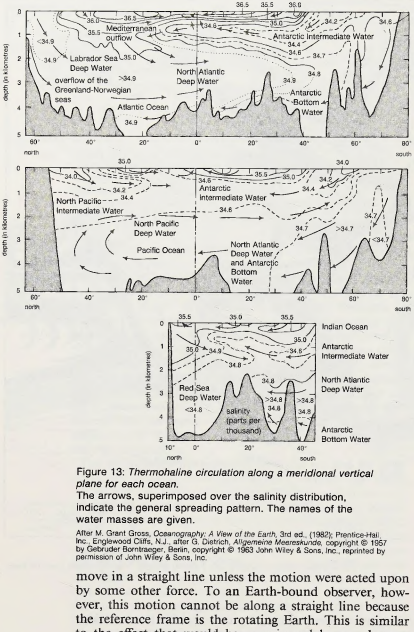


Figure 13: *Thermohaline circulation along a meridional vertical plane for each ocean.*
The arrows, superimposed over the salinity distribution, indicate the general spreading pattern. The names of the water masses are given.

After M. Grant Gross, *Oceanography: A View of the Earth*, 3rd ed., (1982); Prentice-Hall, Inc., Englewood Cliffs, N.J., after G. Dietrich, *Allgemeine Meereskunde*, copyright © 1957 by Gebruder Borntraeger, Berlin, copyright © 1963 John Wiley & Sons, Inc., reprinted by permission of John Wiley & Sons, Inc.

move in a straight line unless the motion were acted upon by some other force. To an Earth-bound observer, however, this motion cannot be along a straight line because the reference frame is the rotating Earth. This is similar to the effect that would be experienced by an observer standing on a large turntable if an object moved over the turntable in a straight line relative to the "outside" world. An apparent deflection of the path of the moving object would be seen. If the turntable rotated counterclockwise, the apparent deflection would be to the right of the direction of the moving object, relative to the observer fixed on the turntable. This remarkable effect is evident in the behaviour of ocean currents. It is called the Coriolis force, named after Gustave-Gaspard Coriolis, a 19th-century French engineer and mathematician. For the Earth, horizontal deflections due to the rotational induced Coriolis force act on particles moving in any horizontal direction. There also are apparent vertical forces, but these are of minor importance to ocean currents. Because the Earth rotates from west to east about its axis, an observer in the Northern Hemisphere would notice a deflection of a moving body toward the right. In the Southern Hemisphere, this deflection would be toward the left. At the equator there would be no apparent horizontal deflection.

It can be shown that the Coriolis force always acts perpendicular to motion. Its horizontal component, C_f, is proportional to the sine of the geographic latitude (θ, given as a positive value for the Northern Hemisphere and a negative value for the Southern Hemisphere) and the speed, c, of the moving body. It is given by $C_f = c\,(2\omega \sin \theta)$, where $\omega = 7.29 \times 10^{-5}$ radian per second is the angular velocity of the Earth's rotation.

Frictional forces. Movement of water through the oceans is slowed by friction, with surrounding fluid moving at a different velocity. A faster-moving fluid layer tends to drag along a slower-moving layer, and a slower-moving layer will tend to reduce the speed of a faster-moving layer. This momentum transfer between the layers is referred to as frictional forces. The momentum transfer is a product of turbulence that moves kinetic energy to smaller scales until at the centimetre scale it is dissipated as heat. The wind blowing over the sea surface transfers momentum to the water. This frictional force at the sea surface (*i.e.*, the wind stress) produces the wind-driven circulation. Currents moving along the ocean floor and the sides of the ocean also are subject to the influence of boundary-layer friction. The motionless ocean floor removes momentum from the circulation of the ocean waters.

Geostrophic currents. For most of the ocean volume away from the boundary layers, which have a characteristic thickness of 100 metres, frictional forces are of minor importance, and the equation of motion for horizontal forces can be expressed as a simple balance of horizontal pressure gradient and Coriolis force. This is called geostrophic balance.

Geostrophic balance

On a nonrotating Earth, water would be accelerated by a horizontal pressure gradient and would flow from high to low pressure. On the rotating Earth, however, the Coriolis force deflects the motion, and the acceleration ceases only when the speed, c, of the current is just fast enough to produce a Coriolis force that can exactly balance the horizontal pressure-gradient force. This geostrophic balance is given as $dp/dn = \rho c 2\omega \sin \theta$, where dp/dn is the horizontal pressure gradient. From this balance, it follows that the current direction must be perpendicular to the pressure gradient because the Coriolis force always acts perpendicular to the motion. In the Northern Hemisphere this direction is such that the high pressure is to the right when looking in current direction, while in the Southern Hemisphere it is to the left. This type of current is called a geostrophic current. The simple equation given above provides the basis for an indirect method of computing ocean currents. The relief of the sea surface (see Figure 14) also defines the streamlines (paths) of the geostrophic current at the surface relative to the deep reference level. The hills represent high pressure, and the valleys stand for low pressure. Clockwise rotation in the Northern Hemisphere with higher pressure in the centre of rotation is called anticyclonic motion. Counterclockwise rotation with lower pressure in its centre is cyclonic motion. In the Southern Hemisphere the sense of rotation is the opposite, because the effect of the Coriolis force has changed its sign of deflection.

Ekman layer. The wind exerts stress on the ocean surface proportional to the square of the wind speed and in the direction of the wind, setting the surface water in motion. This motion extends to a depth of about 100 metres in what is called the Ekman layer, after the Swedish oceanographer V. Walfrid Ekman, who in 1902 deduced these results in a theoretical model constructed to help explain observations of wind drift in the Arctic. Within the oceanic Ekman layer the wind stress is balanced by the Coriolis force and frictional forces. The surface water is directed at an angle of 45° to the wind, to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. With increasing depth in the boundary layer, the current speed is reduced, and the direction rotates farther away from the wind direction following a spiral form, becoming antiparallel to the surface flow at the base of the layer where the speed is $^1/_{23}$ of the surface speed. This so-called Ekman spiral may be the exception rather than the rule, as the specific conditions are not often met, though deflection of a wind-driven surface current at somewhat smaller than 45° is observed when the wind field blows with a steady force and direction for the better part of a day. The average water particle within the Ekman layer moves at an angle of 90° to the wind; this movement is to the right of the wind direction in the Northern Hemisphere and to its left in the Southern Hemisphere. This phenomenon is called Ekman transport, and its effects are widely observed in the oceans.

Since the wind varies from place to place, so does the Ekman transport, forming convergence and divergence zones of surface water. A region of convergence forces surface water downward in a process called downwelling, while a region of divergence draws water from below into the surface Ekman layer in a process known as upwelling. Upwelling and downwelling also occur where the wind blows

Downwelling and upwelling

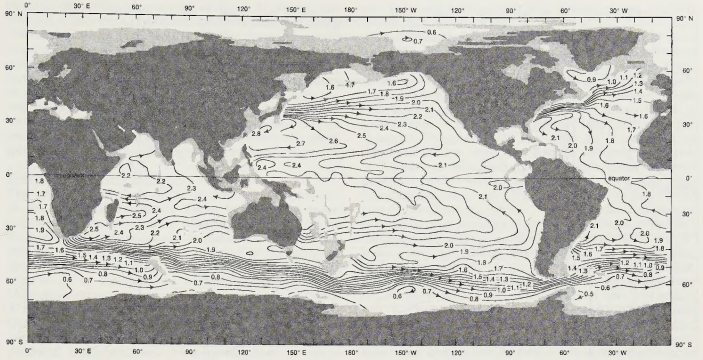


Figure 14: *Sea surface topographic relief relative to the 2,000-decibar (about 2,000-metre depth) isobaric surface.*
The sea surface topography is expressed in energy units (10 metres squared per second squared), with a factor that makes the numbers numerically similar to metres. The geostrophic current flows along lines of equal sea-level elevation in the direction indicated by the arrow heads.
From S. Levitus, *Climatological Atlas of the World Ocean* (1982); National Oceanic and Atmospheric Administration

parallel to a coastline. The principal upwelling regions of the world are along the eastern boundary of the subtropical ocean waters, as, for example, the coastal region of Peru and northwestern Africa. Upwelling in these regions cools the surface water and brings nutrient-rich subsurface water into the sunlit layer of the ocean, resulting in a biologically productive region. Upwelling and high productivity also are found along divergence zones at the equator and around Antarctica. The primary downwelling regions are in the subtropical ocean waters—*e.g.*, the Sargasso Sea in the North Atlantic. Such areas are devoid of nutrients and are poor in marine life.

The vertical movements of ocean waters into or out of the base of the Ekman layer amount to less than one metre per day, but they are important since they extend the wind-driven effects into deeper waters. Within an upwelling region, the water column below the Ekman layer is drawn upward. This process, with conservation of angular momentum on the rotating Earth, induces the water column to drift toward the poles. Conversely, downwelling forces water into the water column below the Ekman layer, inducing drift toward the equator. An additional consequence of upwelling and downwelling for stratified waters is to create a baroclinic field of mass (see above). Surface water is less dense than deeper water. Ekman convergences have the effect of accumulating less dense surface water. This water floats above the surrounding water, forming a hill in sea level and driving an anticyclonic geostrophic current that extends well below the Ekman layer. Divergences do the opposite; they remove the less dense surface water, replacing it with denser, deeper water. This induces a depression in sea level with a cyclonic geostrophic current.

The ocean current pattern produced by the wind-induced Ekman transport is called the Sverdrup transport, after the Norwegian oceanographer H.U. Sverdrup, who formulated the basic theory in 1947. Several years later (1950), the American geophysicist and oceanographer Walter H. Munk and others expanded Sverdrup's work, explaining many of the major features of the wind-driven general circulation by using the mean climatological wind stress distribution at the sea surface as a driving force (Figure 15).

WIND-DRIVEN CIRCULATION

Wind stress induces a circulation pattern that is similar for each ocean. In each case, the wind-driven circulation is divided into large gyres that stretch across the entire ocean: subtropical gyres extend from the equatorial current system to the maximum westerlies in a wind field near 50° latitude, and subpolar gyres extend poleward of the maximum westerlies (see below). The depth penetration of the wind-driven currents depends on the intensity of ocean stratification: for those regions of strong stratification, such as the tropics, the surface currents extend to a depth of less than 1,000 metres. Within the low-stratification polar regions, the wind-driven circulation reaches all the way to the seafloor.

Equatorial currents. At the equator the currents are for the most part directed toward the west, the North Equatorial Current in the Northern Hemisphere and the South Equatorial Current in the Southern Hemisphere. Near the thermal equator, where the warmest surface water is found, there occurs the eastward-flowing Equatorial Counter Current. This current is slightly north of the geographic equator, drawing the northern fringe of the South Equatorial Current to 5° N. The offset to the Northern Hemisphere matches a similar offset in the wind field. The east-to-west wind across the tropical ocean waters induces Ekman transport divergence at the equator, which cools the surface water there.

Equatorial Undercurrent

At the geographic equator a jetlike current is found just below the sea surface, flowing toward the east counter to the surface current. This is called the Equatorial Undercurrent. It attains speeds of more than 1 metre per second at a depth of nearly 100 metres. It is driven by higher sea level in the western margins of the tropical ocean, producing a pressure gradient, which in the absence of a horizontal Coriolis force drives a west-to-east current along the equator. The wind field reverses the flow within the surface layer, inducing the Equatorial Undercurrent.

Equatorial circulation undergoes variations following the irregular periods of roughly three to eight years of the Southern Oscillation (*i.e.*, fluctuations of atmospheric pressure over the tropical Indo-Pacific region). Weakening of the east-to-west wind during a phase of the Southern Oscillation allows warm water in the western margin to slip back to the east by increasing the flow of the Equatorial Counter Current. Surface water temperatures and sea level decrease in the west and increase in the east. This event is called El Niño. The combined El Niño/Southern Oscillation effect has received much attention because it is associated with global-scale climatic variability (see below *Impact of ocean-atmosphere interactions on weather and climate: El Niño/*

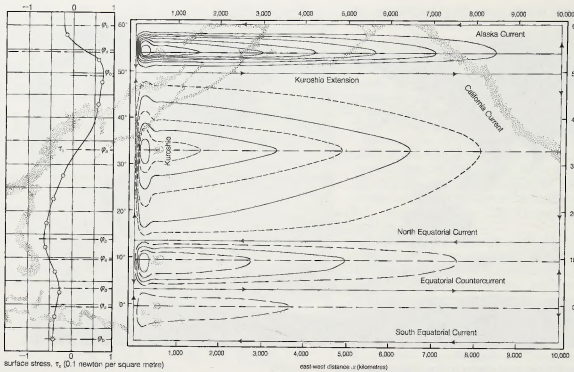


Figure 15: Circulation pattern for the Pacific Ocean north of 10° S calculated from the zonal component of the large-scale wind field shown at the left. The streamlines in map view represent the transport of seawater. Between the adjacent lines flow 10 million cubic metres of water per second.

From W.H. Munk, *Journal of Meteorology* (1950)

Southern Oscillation and climatic change). In the tropical Indian Ocean, the strong seasonal winds of the monsoons induce a similarly strong seasonal circulation pattern.

The subtropical gyres. These are anticyclonic circulation features. The Ekman transport within these gyres forces surface water to sink, giving rise to the subtropical convergence near 20°–30° latitude. The centre of the subtropical gyre is shifted to the west. This westward intensification of ocean currents was explained by the American meteorologist and oceanographer Henry M. Stommel (1948) as resulting from the fact that the horizontal Coriolis force increases with latitude. This causes the polewards-flowing western boundary current to be a jetlike current that attains speeds of two to four metres per second. This current transports the excess heat of the low latitudes to higher latitudes. The flow within the equatorward-flowing interior and eastern boundary of the subtropical gyres is quite different. It is more of a slow drift of cooler water that rarely exceeds 10 centimetres per second. Associated with these currents is coastal upwelling that results from offshore Ekman transport.

The strongest of the western boundary currents is the Gulf Stream in the North Atlantic Ocean. It carries about 30 million cubic metres of ocean water per second through the Straits of Florida and roughly 80 million cubic metres per second as it flows past Cape Hatteras off the coast of North Carolina, U.S. Responding to the large-scale wind field over the North Atlantic, the Gulf Stream separates from the continental margin at Cape Hatteras. After separation, it forms waves or meanders that eventually generate many eddies of warm and cold water. The warm eddies, composed of thermocline water normally found south of the Gulf Stream, are injected into the waters of the continental slope off the coast of the northeastern United States. They drift to the southeast at rates of approximately five to eight centimetres per second, and after a year they rejoin the Gulf Stream north of Cape Hatteras. Cold eddies of slope water are injected into the region south of the Gulf Stream and drift to the southwest. After two years they reenter the Gulf Stream just north of the Antilles Islands. The path that they follow defines a clockwise-flowing recirculation gyre seaward of the Gulf Stream. (For additional details on the Gulf Stream, see below *Impact of ocean-atmosphere interactions on weather and climate: The Gulf Stream and Kuroshio systems.*)

Among the other western boundary currents, the Kuroshio of the North Pacific is perhaps the most like the Gulf Stream, having a similar transport and array of eddies. The Brazil and East Australian currents are relatively weak. The Agulhas Current has a transport close to that of the Gulf Stream. It remains in contact with the margin of Africa around the southern rim of the continent. It then separates from the margin and curls back to the Indian Ocean in what is called the Agulhas Retroflection. Not all the water carried by the Agulhas returns to the east; about 10 to 20 percent is injected into the South Atlantic Ocean as large eddies that slowly migrate across it.

The subpolar gyres. The subpolar gyres are cyclonic circulation features. The Ekman transport within these features forces upwelling and surface water divergence. In the North Atlantic the subpolar gyre consists of the North Atlantic Current at its equatorward side and the Norwegian Current that carries relatively warm water northward along the coast of Norway. The heat released from the Norwegian Current into the atmosphere maintains a moderate climate in northern Europe. Along the east coast of Greenland is the southward-flowing cold East Greenland Current. It loops around the southern tip of Greenland and continues flowing into the Labrador Sea. The southward flow that continues off the coast of Canada is called the Labrador Current. This current separates for the most part from the coast near Newfoundland to complete the subpolar gyre of the North Atlantic. Some of the cold water of the Labrador Current, however, extends farther south.

North Atlantic and Norwegian currents

In the North Pacific the subpolar gyre is composed of the northward-flowing Alaska Current, the Aleutian Current (also known as the Subarctic Current), and the southward-flowing cold Oyashio Current. The North Pacific Current forms the separation between the subpolar and subtropical gyres of the North Pacific.

In the Southern Hemisphere, the subpolar gyres are less defined. Large cyclonic flowing gyres lie poleward of the Antarctic Circumpolar Current and can be considered counterparts to the Northern Hemispheric subpolar gyres. The best-formed is the Weddell Gyre of the South Atlantic sector of the Southern Ocean (see above). The Antarctic coastal current flows toward the west. The northward-flowing current off the east coast of the Antarctic Peninsula carries cold Antarctic coastal water into the circumpolar belt. Another cyclonic gyre occurs north of the Ross Sea.

Antarctic Circumpolar Current. The Southern Ocean links the major oceans by a deep circumpolar belt in the 50°–60° S range. In this belt flows the Antarctic Circumpolar Current from west to east, encircling the globe at high latitudes. It transports 125 million cubic metres of

seawater per second over a path of about 24,000 kilometres and is the most important factor in diminishing the differences between oceans. The Antarctic Circumpolar Current is not a well-defined single-axis current but rather consists of a series of individual filaments separated by frontal zones. It reaches the seafloor and is guided along its course by the irregular bottom topography. Large meanders and eddies develop in the current as it flows. These features induce poleward transfer of heat, which may be significant in balancing the oceanic heat loss to the atmosphere above the Antarctic region farther south.

THERMOHALINE CIRCULATION

The general circulation of the oceans consists primarily of the wind-driven currents. These, however, are superimposed on the much more sluggish circulation driven by horizontal differences in temperature and salinity—namely, the thermohaline circulation. The thermohaline circulation reaches down to the seafloor and is often referred to as the deep, or abyssal, ocean circulation. Measuring seawater temperature and salinity distribution is the chief method of studying the deep-flow patterns. Other properties also are examined; for example, the concentrations of oxygen, carbon-14, and such synthetically produced compounds as chlorofluorocarbons are measured to obtain resident times and spreading rates of deep water.

Abyssal circulation

In some areas of the ocean, generally during the winter season, cooling or net evaporation causes surface water to become dense enough to sink. Convection penetrates to a level where the density of the sinking water matches that of the surrounding water. It then spreads slowly into the rest of the ocean. Other water must replace the surface water that sinks. This sets up the thermohaline circulation. The basic thermohaline circulation is one of sinking of cold water in the polar regions, chiefly in the northern North Atlantic and near Antarctica. These dense water masses spread into the full extent of the ocean and gradually upwell to feed a slow return flow to the sinking regions. A theory for the thermohaline circulation pattern was proposed by Stommel and Arnold Arons in 1960 (Figure 16).

In the Northern Hemisphere, the primary region of deep water formation is the North Atlantic; minor amounts of deep water are formed in the Red Sea and Persian Gulf. A variety of water types contribute to the so-called North Atlantic Deep Water. Each one of them differs, though they share a common attribute of being relatively warm (greater than 2° C) and salty (greater than 34.9 parts per thousand) compared with the other major producer of deep and bottom water, the Southern Ocean (0° C and 34.7 parts per thousand). North Atlantic Deep Water is primarily formed in the Greenland and Norwegian seas, where cooling of the salty water introduced by the Norwegian Current induces sinking. This water spills over the rim of the ridge that stretches from Greenland to Scotland, extending to the seafloor to the south as a convective plume. It then flows southward, pressed against the western edge of the North Atlantic. Additional deep water is formed in the Labrador Sea. This water, somewhat less dense than the overflow water from the Greenland and Norwegian seas, has been observed sinking to a depth of 3,000 metres within convective features referred to as chimneys. Vertical velocities as high as 10 centimetres per second have been observed within these convective features. A third variety of North Atlantic Deep Water is derived from net evaporation within the Mediterranean Sea. This draws surface water into the Mediterranean through the Strait of Gibraltar. The mass of salty water formed within the Mediterranean exits as a deeper stream. It descends to depths of 1,000 to 2,000 metres in the North Atlantic Ocean, forming the uppermost layer of North Atlantic Deep Water. The outflow in the Strait of Gibraltar reaches as high as 2 metres per second, but its total transport amounts to only 5 percent of the total North Atlantic Deep Water formed. The outflow of the Mediterranean plays a significant role in boosting the salinity of North Atlantic Deep Water.

The blend of North Atlantic Deep Water, with a total formation rate of 15 to 20×10^6 cubic metres per second,

From Henry Stommel, *The Gulf Stream: A Physical and Dynamical Description*, originally published by the University of California Press; reprinted by permission of The Regents of the University of California

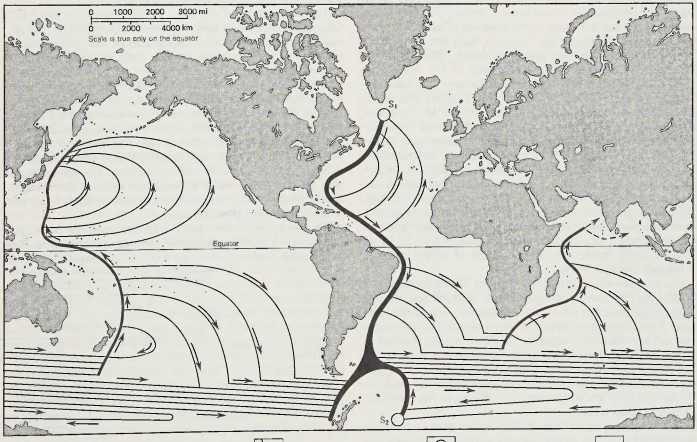


Figure 16: Circulation pattern of deep flow proposed by Henry Stommel and Arnold Arons. According to their theory, surface water sinks down into the deep ocean at two sites: the northern North Atlantic (S_1) and the Weddell Sea (S_2). Primary pathways of meridional motion occur in the western boundaries of the ocean basins. These deep western-boundary currents lose water to the ocean interior, where the deep water slowly upwells and drifts poleward to renew the surface water in the regions at which sinking occurs.

quickly ventilates the Atlantic Ocean, resulting in a residence time of less than 200 years. The deep water spreads away from its source along the western side of the Atlantic Ocean and, on reaching the Antarctic Circumpolar Current, spreads into the Indian and Pacific oceans. The sinking of North Atlantic Deep Water is compensated for by the slow upwelling of deep water, mainly in the Southern Ocean, to replenish the upper stratum of water that has descended as North Atlantic Deep Water. North Atlantic Deep Water exported to the other oceans must be balanced by the inflow of upper-layer water into the Atlantic. Some water returns as cold, low-salinity Pacific water through the Drake Passage in the form of what is known as Antarctic Intermediate Water, and some returns as warm salty thermocline water from the Indian Ocean around the southern rim of Africa.

Remnants of North Atlantic Deep Water mix with Southern Ocean water to spread along the seafloor into the North Pacific Ocean. Here, it upwells to a level of 2,000–3,000 metres and returns to the south lower in salinity and oxygen but higher in nutrient concentrations as North Pacific Deep Water. This North Pacific Deep Water is eventually swept eastward with the Antarctic Circumpolar Current. Modification of deep water in the North Pacific is the direct consequence of vertical mixing, which carries into the deep ocean the low salinity properties of North Pacific Intermediate Water. The latter is formed in the northwestern Pacific Ocean. Because of the immenseness of the North Pacific and the extremely long residence time (more than 500 years) of the water, enormous quantities of North Pacific Deep Water can be produced by vertical mixing.

Considerable volumes of cold water generally of low salinity are formed in the Southern Ocean. Such water masses spread into the interior of the global ocean and to a large extent are responsible for the anomalous cold, low-salinity state of the modern oceans. The circumstances leading to this role for the Southern Ocean are related to the existence of a deep-ocean circumpolar belt around Antarctica that was established some 25 million years ago by the shifting lithospheric plates which make up the Earth's surface (see below *Ocean basins*). This belt establishes the Antarctic Circumpolar Current, which isolates Antarctica from the warm surface waters of the subtropics. The Antarctic Circumpolar Current does not completely sever contact with the lower latitudes. The Southern Ocean does have access to the waters of the north, but through deep- and bottom-water pathways. The basic dynamics of the Antarctic Circumpolar Current lifts dense deep water occurring north of the current to the ocean surface south of it. Once exposed to the cold Antarctic air masses, the upwelling deep water is converted to the cold Antarctic Bottom Water and Antarctic Intermediate Water. The southward and upwelling deep water, which carries heat injected into the deep ocean by processes farther north, is balanced by the northward spread of cooler, fresher, oxygenated water masses of the Southern Ocean. It is estimated that the overturning rate of water south of the Antarctic Circumpolar Current amounts to 35 to 45 million cubic metres per second, most of which becomes Antarctic Bottom Water.

Antarctic Bottom Water

The primary site of Antarctic Bottom Water formation is within the continental margins of the Weddell Sea, though some is produced in other coastal regions, such as the Ross Sea. Also, there is evidence of deep convective overturning farther offshore. Antarctic Bottom Water, formed at a rate of 30 million cubic metres per second, slips below the Antarctic Circumpolar Current and spreads to regions well north of the equator. Slowly upwelling and modified by mixing with less dense water, it returns to the Southern Ocean as deep water.

The remaining upwelling of deep water spreads near the surface to the north, where it forms Antarctic Intermediate Water within the Antarctic Circumpolar Current zone and spreads along the base of the thermoclines farther north. This water mass forms a sheet of low-salinity water that demarcates the lower boundary of the subtropical thermocline. It upwells into the thermocline, partly compensating for the sinking of North Atlantic Deep Water. (A.L.G.)

Waves of the sea

There are many types of ocean waves. Waves differ from each other in size and in terms of the forces that drive them. Waves represent an oscillatory motion of seawater at regular time intervals or periods. Some may be running, or progressive, waves in which the crests propagate, while others are stationary, or standing, waves. Two of the more common types of waves, gravity waves and tides, are considered here. For gravity waves, the stabilizing force—*i.e.*, the force that attempts to restore the crests and troughs of the waves to the average sea level—is the Earth's gravity. The distance between the crests, or wavelength, of gravity waves range from a few centimetres to many kilometres. Tiny waves at the ocean surface with a wavelength of less than 1.7 centimetres are called capillary waves. Their restoring force is the surface tension of seawater. Capillary waves are direct products of the wind stress exerted on the sea surface and tend to feed wind energy into gravity waves, which characteristically have longer wavelengths.

Capillary waves

Tides are essentially gravity waves that have long periods of oscillation. They may be called forced waves, because they have fixed, prescribed periods that are strictly determined by astronomical forces induced by the relative movements of the Moon, Earth, and Sun. Sometimes the term "tidal wave" is used incorrectly to include such phenomena as surges, which are called storm tides, or destructive waves known as tsunamis that are induced by undersea earthquakes. In the following discussion, the use of the words tide and tidal is restricted to tides of astronomical origin and the forces and phenomena associated with them. Figure 17 shows the different types of surface waves and their relative amounts of energy.

SURFACE GRAVITY WAVES

Of the nontidal kinds of running surface waves, three types may be distinguished: wind waves and swell, wind surges, and sea waves of seismic origin (tsunamis).

Wind waves and swell. Wind waves are the wind-generated gravity waves. After the wind has abated or shifted or the waves have migrated away from the wind field, such waves continue to propagate as swell.

The dependence of the sizes of the waves on the wind field is a complicated one. A general impression of this dependence is given by the descriptions of the various states of the sea corresponding to the scale of wind strengths known as the Beaufort scale (Table 9), after the British admiral Sir Francis Beaufort, who drafted it in 1808, using as his yardstick the surface of sail that a fully rigged warship of those days could carry in the various wind forces. In the Table the Beaufort wind force is followed by the name given to such a wind at sea, and the next column provides the range of wind speeds. When considering the descriptions of the sea surface, it must be remembered that the size of the waves depends not only on the strength of the wind but also on its duration and its fetch—*i.e.*, the length of its path over the sea.

The theory of waves starts with the concept of simple waves, those forming a strictly periodic pattern with one wavelength and one wave period and propagating in one

Wave characteristics and motion

From *Waves, Tides and Shallow-Water Processes*, The Open University in association with Pergamon Press, Ltd.

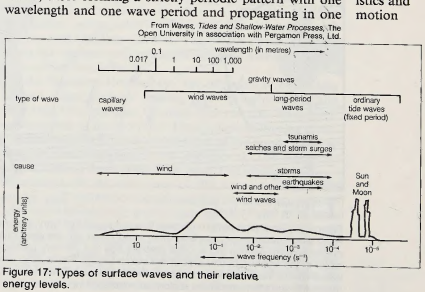


Figure 17: Types of surface waves and their relative energy levels.

direction. Real waves, however, always have a more irregular appearance. They may be described as composite waves, in which a whole spectrum of wavelengths, or periods, is present and which have more or less diverging directions of propagation. In reporting observed wave heights and periods (or lengths) or in forecasting them, one height or one period is mentioned as the height or period, however, and some agreement is needed in order to guarantee uniformity of meaning. The height of simple waves means the elevation difference between the top of a crest and the bottom of a trough. The significant height, a characteristic height of irregular waves, is by convention the average of the highest one-third of the observed wave heights. Period, or wavelength, can be determined from the average of a number of observed time intervals between the passing of successive well-developed wave crests over a certain point, or of observed distances between them.

Wave period and wavelength are coupled by a simple relationship: wavelength equals wave period times wave speed, or $L = TC$, when L is wavelength, T is wave period, and C is wave speed.

The wave speed of surface gravity waves depends on the depth of water and on the wavelength, or period; the speed increases with increasing depth and increasing wavelength, or period. If the water is sufficiently deep, the wave speed is independent of water depth. This relationship of wave speed to wavelength and water depth (d) is given by the equations below. With g being the gravity acceleration (9.8 metres per second squared), $C^2 = gd$, when the wavelength is 20 times greater than the water depth (waves of this kind are called long gravity waves or shallow-water waves); and $C^2 = gL/2\pi$, when the wavelength is less than two times the water depth (such waves are called short waves or deep water waves). For waves with lengths between 2 and 20 times the water depth, the wave speed is governed by a more complicated equation combining these effects:

$$C^2 = \frac{gL}{2\pi} \tanh\left(2\pi \frac{d}{L}\right),$$

where tanh is the hyperbolic tangent.

A few examples are listed below for short waves, giving the period in seconds, the wavelength in metres, and wave speed in metres per second:

period T (sec)	1	2	4	8	16
wavelength L (m)	1.56	6.2	25.0	100	400
wave speed in deep water (m/scc)	1.56	3.1	6.2	12.5	25.0

Wave trains and group velocity

Waves often appear in groups as the result of interference of wave trains of slightly differing wavelengths. A wave group as a whole has a group speed that generally is less than the speed of propagation of the individual waves; the two speeds are equal only for groups composed of long waves. For deepwater waves, the group velocity (V) is half the wave speed (C). In the physical sense, group velocity is the velocity of propagation of wave energy. From the dynamics of the waves, it follows that the wave energy per unit area of the sea surface is proportional to the square of the wave height, except for the very last stage of waves running into shallow water, shortly before they become breakers.

The height of wind waves increases with increasing wind speed and with increasing duration and fetch of the wind (*i.e.*, the distance over which the wind blows). Together with height, the dominant wavelength also increases. Finally, however, the waves reach a state of saturation because they attain the maximum significant height to which the wind can raise them, even if duration and fetch are unlimited. For instance, winds of 5 metres per second, 15 metres per second, and 25 metres per second may raise waves with significant heights up to 0.5 metre, 4.5 metres, and 12.5 metres, respectively, with corresponding wavelengths of 16 metres, 140 metres, and 400 metres, respectively.

After becoming swell, the waves may travel thousands of kilometres over the ocean, particularly if the swell is from the large storms of moderate and high latitudes, whence it easily may travel into the subtropical and equatorial zones, and the swell of the trade winds, which runs into the equatorial calms. In traveling, the swell waves gradually become lower; energy is lost by internal friction and air resistance and by energy dissipation because of some divergence of the directions of propagation (fanning out). With respect to the energy loss, there is a selective damping of the composite waves, the shorter waves of the wave mixture suffering a stronger damping over a given distance than the longer ones. As a consequence, the dominant wavelength of the spectrum shifts toward the greater wavelengths. Therefore, an old swell must always be a long swell.

When waves run into shallow water, their speed of propagation and wavelength decrease, but the period remains the same. Eventually, the group velocity, the velocity of energy propagation, also decreases, and this decrease causes the height to increase. The latter effect may, however, be affected by refraction of the waves, a swerving of the wave crests toward the depth lines and a corresponding deviation of the direction of propagation. Refraction may cause a convergence or divergence of the energy stream and result in a raising or lowering of the waves, especially over nearshore elevations or depressions of the sea bottom.

In the final stage, the shape of the waves changes, and the crests become narrower and steeper until, finally, the waves become breakers (surf). Generally, this occurs where the depth is 1.3 times the wave height.

Wind surges. Running wind surges are long waves caused by a piling up of the water over a large area through the action of a traveling wind or pressure field. Examples include the surge in front of a traveling storm cyclone,

Table 9: The Beaufort Scale

Beaufort number	name of wind	wind speed		description of sea surface
		knots	km/hr	
0	calm	<1	<1	like a mirror
1	light air	1–3	1–5	ripples with the appearance of scales are formed, but without foam crests
2	light breeze	4–6	6–11	small wavelets, still short but more pronounced; crests have a glassy appearance and do not break
3	gentle breeze	7–10	12–19	large wavelets; crests begin to break; foam of glassy appearance, perhaps scattered white horses
4	moderate breeze	11–16	20–28	small waves, becoming longer; fairly frequent white horses
5	fresh breeze	17–21	29–38	moderate waves, taking a more pronounced long form; many white horses are formed (chance of some spray)
6	strong breeze	22–27	39–49	large waves begin to form; the white foam crests are more extensive everywhere (probably some spray)
7	moderate gale	28–33	50–61	sea heaps up and white foam from breaking waves begins to be blown in streaks along the direction of the wind
8	fresh gale	34–40	62–74	moderately high waves of greater length; edges of crests begin to break into spindrift; the foam is blown in well-marked streaks along the direction of the wind
9	strong gale	41–47	75–88	high waves; dense streaks of foam along the direction of the wind; crests of waves begin to topple, tumble, and roll over; spray may affect visibility
10	whole gale	48–55	89–102	very high waves with overhanging crests; the resulting foam, in great patches, is blown in dense white streaks along the direction of the wind; on the whole the surface of the sea takes a white appearance; the tumbling of the sea becomes heavy and shock-like; visibility affected
11	storm	56–65	103–117	exceptionally high waves (small and medium-sized ships might be for a time lost to view behind the waves); the sea is completely covered with long white patches of foam lying along the direction of the wind; everywhere the edges of the wave crests are blown into froth; visibility affected
12–17	hurricane	above 65	above 117	the air is filled with foam and spray; sea completely white with driving spray; visibility very seriously affected

particularly the devastating hurricane surge caused by a tropical cyclone, and the surge occasionally caused by a wind convergence line, such as a traveling front with a sharp wind shift.

Generation of tsunamis

Waves of seismic origin. A tsunami (Japanese: *tsu,* "harbour," and *nami,* "sea") is a very long wave of seismic origin that is caused by a submarine or coastal earthquake, landslide, or volcanic eruption. Such a wave may have a length of hundreds of kilometres and a period on the order of a quarter of an hour. It travels across the ocean at a tremendous speed. (Tsunamis are long waves traveling at the wave speed given by $C^2 = gd$.) To a depth of 4,000 metres, for instance, the corresponding wave speed is about 200 metres per second, or 720 kilometres per hour. In the open ocean the height of tsunamis may be less than one metre, and they pass unnoticed. As they approach a continental shelf, however, their speed is reduced and their height increases dramatically. Tsunamis have caused enormous destruction of life and property, piling up in coastal waters at places thousands of kilometres away from their point of origin, particularly in the Pacific Ocean.

Standing, or seiche, waves. A freestanding wave may arise in an enclosed or nearly enclosed basin as a free swinging or sloshing of the whole water mass. Such a standing wave is also called a seiche, after the name given to the oscillating movements of the water of Lake Geneva, Switz., where this phenomenon first was studied seriously. The period of oscillation is independent of the force that first brought the water mass out of equilibrium (and that is supposed to have ceased thereafter), but depends only on the dimensions of the enclosing basin and on the direction in which the water mass is swinging. Assuming a simple rectangular basin of constant depth and the most simple lengthwise oscillation, the period of oscillation (T) is equal to two times the length of the basin divided by the wave speed computed from the shallow-water formula above. This relationship may be written: $T = L/C$, in which L equals two times the length of the basin and C is the wave speed found from the formula, using the known depth of the basin. Besides this fundamental tone (or response to stimuli), the water mass also may swing according to an overtone, showing one or more nodal lines across the basin.

The water in an open bay or marginal sea also may perform such a free oscillation as a standing wave, the difference being that in an open bay the greatest horizontal displacements are not in the middle of the bay but at the mouth. For the fundamental period of oscillation, the formula given above is used with a wavelength equal to four times the length (from the mouth to the closed end) of the bay. In practice, of course, it is more difficult than that, because the form of a bay or marginal sea is irregular and the depth differs from place to place. The North Sea has a period of lengthwise swinging of about 36 hours. The cause of such free oscillations may be a temporary wind or pressure field, which brought the sea surface out of its horizontal position and which afterward ceased to act more or less abruptly, leaving the water mass out of equilibrium.

Internal waves. Gravity waves also occur on internal "surfaces" within oceans. These surfaces represent strata of rapidly changing water density with increasing depth, and the associated waves are called internal waves. Internal waves manifest themselves by a regular rising and sinking of the water layers around which they centre, whereas the height of the sea surface is hardly affected at all. Because the restoring force, excited by the internal deformation of the water layers of equal density, is much smaller than in the case of surface waves, internal waves are much slower than the latter. Given the same wavelength, the period is much longer (the movements of the water particles being much more sluggish), and the speed of propagation is much smaller; the formulas for the speed of surface waves include the acceleration of gravity, g, but those for internal waves include the gravity factor times the difference between the densities of the upper and the lower water layer divided by their sum.

Cause of internal waves

The cause of internal waves may lie in the action of tidal forces (the period then equaling the tidal period) or in the action of a wind or pressure fluctuation. Sometimes, a ship may cause internal waves (dead water) if there is a shallow, brackish upper layer.

OCEAN TIDES

The tides may be regarded as forced waves, partially running waves and partially standing waves. They are manifested by vertical movements of the sea surface (the height maximum and minimum are called high water [HW] and low water [LW]) and in alternating horizontal movements of the water, the tidal currents. The words ebb and flow are used to designate the falling tide and the rising tide, respectively.

Tide-generating forces. The forces that cause the tides are called the tide-generating forces. A tide-generating force is the resultant force of the attracting force of the Moon or the Sun and the force of inertia (centrifugal force) that results from the orbital movement of the Earth around the common centre of gravity of the Earth-Moon or Earth-Sun system.

Considering the Earth-Moon system, at any time the tide-generating force is directed vertically upward at the two places on the Earth where the Moon is in the vertical (on the same and on the opposite side of the Earth); it is directed vertically downward at all places (forming a circle) where the Moon is in the horizon at that moment. At all other places, the tide-generating force also has a horizontal component. Because this pattern of forces is coupled to the position of the Moon with respect to the Earth and because for any place on the Earth's surface the relative position of the Moon with respect to that place has, on the average, a periodicity of 24 hours 50 minutes, the tide-generating force felt at any place has that same periodicity. When the Moon is in the plane of the equator, the force runs through two identical cycles within this time interval because of the symmetry of the global pattern of forces described above. Consequently, the tidal period is 12 hours 25 minutes in this case; it is the period of the semidiurnal lunar tide. The fact that the Moon is alternately to the north and to the south of the equator causes an inequality of the two successive cycles within the time interval of 24 hours 50 minutes. The effect of this inequality is formally described as the superposition of a partial tide called the diurnal lunar tide, with the period of 24 hours 50 minutes, on the semidiurnal lunar tide.

In the same manner, the Sun causes a semidiurnal solar tide, with a 12-hour period, and a diurnal solar tide, with a 24-hour period. In a complete description of the local variations of the tidal forces, still other partial tides play a role because of further inequalities in the orbital motions of the Moon and the Earth.

The interference of the solar-tidal forces with the lunar-tidal forces (the lunar forces are about 2.2 times as strong) causes the regular variation of the tidal range between spring tide, when it has its maximum, and neap tide, when it has its minimum.

Although the tide-generating forces are very small in comparison with the Earth's force of gravity (the lunar tidal force at its maximum being only 1.14×10^{-7} times the force of gravity), their effects upon the sea are considerable because of their horizontal component. Since the Earth is not surrounded by an uninterrupted envelope of water but rather shows a very irregular alternation of sea and land, the mechanism of the response of the oceans and seas to the tidal forces is extremely complex. A further complication is caused by the deflecting force of the Earth's rotation (the Coriolis force; see above).

In enclosures formed by gulfs and bays, the local tide is generated by interaction with the tides of the adjacent open ocean. Such a tide often takes the form of a running tidal wave that rotates within the confines of the enclosure. In some semi-enclosed seas, such as the Mediterranean, Black, and Baltic seas, a standing wave, or tidal seiche, may be generated by the local tide-raising forces.

Running tidal wave

In these seas, the tidal range of sea level is only on the order of centimetres. In the open ocean, it generally is on the order of tens of centimetres. In bays and adjacent seas, however, the tidal range may be much greater, because the shape of a bay or adjacent sea may favour the

enhancement of the tide inside; in particular, there may be a resonance of the basin concerned with the tide. The largest known tides occur in the Bay of Fundy, where spring tidal ranges up to 15 metres have been measured.

Tidal bores. Tidal bores form on rivers and estuaries near a coast where there is a large tidal range and the incoming tide is confined to a narrow channel. They consist of a surge of water moving swiftly upstream headed by a wave or series of waves. Such bores are quite common. There is a large one, known as the *mascaret* on the Seine, which forms on spring tides and reaches as far upriver as Rouen. There is a well-known bore on the Severn, in England, and another forms on the Petitcodiac River, which empties into the Bay of Fundy in New Brunswick. The classic example is the bore on the Ch'ien-t'ang described by Commander W. Usborne Moore of the British navy in 1888 and 1892. He reported heights of 2.5 to 3.5 metres.

When a tidal bore forms in a river, the direction of flow of the water changes abruptly as the bore passes. Before it arrives, the water may be still or, more usually, a small freshwater current flows outward toward the sea. The tide comes in as a "wall of water" that passes up the river. Behind the bore, the current flows upriver. At the division between the moving water behind the bore and the still water in front, there is a wave, the water surface behind being higher than it is in front. This wave must travel more quickly than the water particles behind it, because, as the advancing water travels upriver, it collects the still water in front and sets it in motion. Upriver, the advancing tide will consist not of salt water from the sea but rather of fresh water that has passed farther down and been collected and returned in front of the incoming tide. It is therefore necessary to distinguish between the velocity of the advancing wave and that of the water particles just behind it. (P.Gr./R.A.R.T./A.L.G.)

Density currents in the oceans

GENERAL OBSERVATIONS

Density currents are currents that are kept in motion by the force of gravity acting on a relatively small density difference caused by variations in salinity, temperature, or sediment concentration. As noted above, salinity and temperature variations produce stratification in oceans. Below the surface layer, which is disturbed by waves and is lighter than the deeper waters because it is warmer or less saline, the oceans are composed of layers of water that have distinctive chemical and physical characteristics, which move more or less independently of each other and which do not lose their individuality by mixing even after they have flowed for hundreds of kilometres from their point of origin.

Stratified flow

An example of this type of density current, or stratified flow, is provided by the water of the Mediterranean Sea as it flows through the Strait of Gibraltar out into the Atlantic. Because the Mediterranean is enclosed in a basin that is relatively small compared with the ocean basins and because it is located in a relatively arid climate, evaporation exceeds the supply of fresh water from rivers. The result is that the Mediterranean contains water that is both warmer and more saline than normal deep-sea water, the temperature ranging from 12.7° to 14.5° C and the salinity from 38.4 to 39.0 parts per thousand. Because of these characteristics, the Mediterranean water is considerably denser than the water in the upper parts of the North Atlantic, which has a salinity of about 36 parts per thousand and a temperature of about 13° C. The density contrast causes the lighter Atlantic water to flow into the Mediterranean in the upper part of the Strait of Gibraltar (down to a depth of about 200 metres) and the denser Mediterranean water to flow out into the Atlantic in the lower part of the strait (from about 200 metres to the top of the sill separating the Mediterranean from the Atlantic at a depth of 320 metres). Because the strait is only about 20 kilometres wide, both inflow and outflow achieve relatively high speeds. Near the surface the inflow may have speeds as high as two metres per second, and the outflow reaches speeds of more than one metre per second at a depth of about 275 metres. One result of the high current speeds in the strait is that there is a considerable amount of mixing, which reduces the salinity of the outflowing Mediterranean water to about 37 parts per thousand. The outflowing water sinks to a depth of about 1,500 metres or more, where it encounters colder, denser Atlantic water. It then spreads out as a layer of more saline water between two Atlantic water masses.

TURBIDITY CURRENTS

Density currents caused by suspended sediment concentrations in the oceans are called turbidity currents. They appear to be relatively short-lived, transient phenomena that occur at great depths. Turbidity currents are thought to be caused by the slumping of sediment that has piled up at the top of the continental slope, particularly at the heads of submarine canyons (see below *Continental margins: Submarine canyons*). Slumping of large masses of sediment creates a dense sediment-water mixture, or slurry, which then flows down the canyon to spread out over the ocean floor and deposit a layer of sand in deep water. Repeated deposition forms submarine fans, which are analogous to the alluvial fans found at the mouths of many river canyons. Sedimentary rocks that are thought to have originated from ancient turbidity currents are called turbidites.

Evidence for turbidity currents

Although large-scale turbidity underflows have never been directly observed in the oceans, there is much evidence supporting their occurrence. This evidence may be briefly summarized: (1) Telegraph cables have been broken in the deep ocean in a sequence that indicates some disturbance at the bottom moving from shallow to deep water at speeds on the order of 20 to 75 kilometres per hour, or 10 to 40 knots. The trigger for this phenomenon is commonly, though not exclusively, an earthquake near the edge of the continental slope. The only disturbance that seems capable of being transmitted downslope at the required speed is a large turbidity current. The best-known example of such a series of cable breaks took place in the North Atlantic following the 1929 earthquake under the Grand Banks of Newfoundland (Figure 18), but other examples have been described from the Magdalena River delta (Colombia), the Congo delta, the Mediterranean Sea north of Orléansville and south of the Straits of Messina, and Kandavu Passage, Fiji. (2) Cores taken from the ocean bottom in the area downslope from cable breaks reveal layers of sand interbedded with normal deep-sea pelagic or hemipelagic oozes (sediments formed in the deep sea by quiet settling of fine particles). In the case of the cable breaks south of the Grand Banks, a large-diameter core taken from the axis of a submarine canyon in the continental slope contained 1 centimetre of gray clay underlain by at least

From H.W. Menard, *Marine Geology of the Pacific* (© 1964); used with permission of McGraw-Hill Book Company

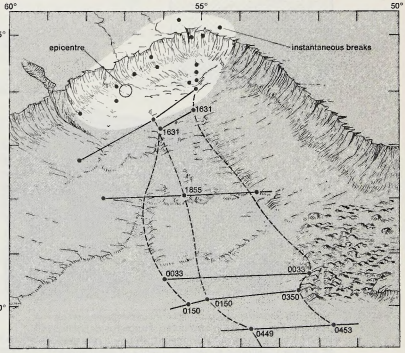


Figure 18: Location of cables (heavy lines) and times of breaks following the Grand Banks earthquake of 1929.

20 centimetres of gray pebble and cobble gravel. Cores farther south showed a graded layer about one metre thick of coarse silt and fine sand. The presence of these gravel and sand layers is consistent with the hypothesis that they were deposited by the turbidity current that broke the cables. (3) Coring has revealed layers of fine-grained sand or coarse silt at many other localities in the abyssal plains of the oceans. These layers are generally moderately well sorted and contain microfossils characteristic of shallow water that are also size-sorted. In some cases the layers are laminated and arranged in a definite sequence (Figure 19). It is clear that the sand forming these layers has been moved down from shallow water, and in many cases the only plausible mechanism appears to be a turbidity current. (4) At the base of many submarine canyons there occur very large submarine fans. Deep-sea channels on the fan surfaces extend for many tens of kilometres and have depths of more than 100 metres and widths of one kilometre or more. Submarine levees are a prominent feature, and these project above the surrounding fan surface to elevations of 50 metres or higher. The gross characteristics of such channels suggest that they were formed by a combination of erosion and deposition by turbidity currents.

Turbidites

(5) Thick deposits of interbedded graded sandstones and fine-grained shales are common in the geologic record. In some cases there is good fossil evidence that the shales were deposited in relatively deep water, perhaps as much as several thousand metres deep. Relatively deepwater deposition is also suggested by the absence of sedimentary structures characteristic of shallow water. The interbedded sandstones, however, contain shallow-water fossils that are sorted by size, have a sharp basal contact with the shale below and a transitional contact with the shale above, and display a characteristic sequence of sedimentary structures (Figure 19). The structures include erosional marks made originally on the mud surface but now preserved as casts on the base of the sandstone bed (sole marks) and internal structures including some or all of the following: massive graded unit, parallel lamination, ripple cross-lamination or convolute lamination, and an upper unit of parallel lamination. This combination of textural and structural features can be explained by deposition from a current that slightly erodes the bottom and then deposits sand that becomes finer grained as the velocity gradually wanes. The properties inferred from these ancient sandstone deposits are consistent with the properties of turbidity currents inferred from laboratory experiments.

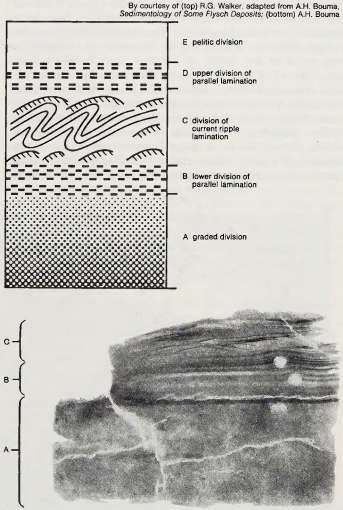


Figure 19: *Beds deposited by turbidity currents—turbidites.* (Top) Ideal sequence of sedimentary structures deposited in a bed by a turbidity current. (Bottom) Radiograph of a turbidite off southern California showing sequence of sedimentary structures. (A) Graded division. (B) Lower division of parallel lamination. (C) Division of current ripple lamination.

In spite of the convincing nature of the evidence, there are still some objections to the turbidity current hypothesis. Most geologists and oceanographers accept that such currents exist and that the currents are important agents of erosion and sediment deposition, in both modern and ancient seas, but researchers believe that the turbidity current hypothesis has been overworked. There is evidence, for example, which suggests that currents flowing parallel to submarine contours exist in many ocean basins. These bottom currents have been observed in a few cases, and velocities as high as 20 to 50 centimetres per second have been recorded. These currents can produce some of the features that previously had been attributed to turbidity current action. Moreover, nearly all features of sands that are produced by turbidity currents can be formed by shallow-water action, such as fluvial processes. Hence the problem of discriminating between deposits formed by turbidity currents and deposits formed by other current types is quite complex and requires a careful assessment of all lines of evidence in each case. Some ancient sandstones have been interpreted as "fluxoturbidites" because the sedimentary structures and other properties suggest a transporting agent intermediate between turbidity currents and large-scale slumping and sliding of sediment.

(G.V.M./Ed.)

Impact of ocean-atmosphere interactions on weather and climate

SEASONAL AND INTERANNUAL OCEAN-ATMOSPHERE INTERACTIONS

General considerations. The notion of a connection between the temperature of the surface layers of the oceans and the circulation of the lowest layer of the atmosphere, the troposphere, is a familiar one. The surface mixed layer of the ocean is a huge reservoir of heat when compared to the overlying atmosphere. The heat capacity of an atmospheric column of unit area cross-section extending from the ocean surface to the outermost layers of the atmosphere is equivalent to the heat capacity of a column of seawater of 2.6-metre depth. The surface layer of the oceans is continuously being stirred by the overlying winds and waves, and thus a surface mixed layer is formed that has vertically uniform properties in temperature and salinity. This mixed layer, which is in direct contact with the atmosphere, has a minimum depth of 20 metres in summer and a maximum depth exceeding 100 metres in late winter in the mid-latitudes. In lower latitudes the seasonal variation in the mixed layer is less marked than at higher latitudes, except in regions such as the Arabian Sea where the onset of the southwestern Indian monsoon may produce large changes in the depth of the mixed layer.

Long-lived temperature anomalies

Temperature anomalies (*i.e.,* deviations from the normal seasonal temperature) in the surface mixed layer have a long residence time compared with those of the overlying turbulent atmosphere. Hence they may persist for a number of consecutive seasons and even for years.

Observational studies to investigate the relationship between anomalies in ocean surface temperature and the tropospheric circulation have been undertaken primarily in the Pacific and Atlantic. They have identified large-scale ocean surface temperature anomalies that have similar spatial scales to monthly and seasonal anomalies in atmospheric circulation. The longevity of the ocean surface temperature anomalies, as compared with the shorter dynamical and thermodynamical "memory" of the atmosphere, has suggested that they may be an important predictor for seasonal and interannual climate anomalies.

Link between ocean surface temperature and climate anomalies. First, it is useful to consider some examples

of the association between anomalies in ocean surface temperature and irregular changes in climate. The Sahel, a region that borders the southern fringe of the Sahara in Africa, experienced a number of devastating droughts during the 1970s and '80s, which can be compared with a much wetter period during the 1950s. Figure 20 shows the difference in ocean surface temperature during the period from July to September between the "driest" and "wettest" rainfall seasons in the Sahel after 1950. Of particular note are the higher-than-normal surface temperatures in the tropical South Atlantic, Indian, and Southeast Pacific oceans and the lower-than-normal temperatures in the North Atlantic and Pacific oceans. This example illustrates that climate anomalies in one region of the world may be linked to ocean surface temperature changes on a global scale. Global atmospheric modeling studies undertaken during the mid-1980s have indicated that the positions of the main rainfall zones in the tropics are sensitive to anomalies in ocean surface temperature similar to that depicted in Figure 20.

Shorter-lived climate anomalies, on time scales of months to one or two years, also have been related to ocean surface temperature anomalies. The equatorial oceans have the largest influence on these climate anomalies because of the evaporation of water. A relatively small change in ocean surface temperature, say, of 1° C, may result in a large change in the evaporation of water into the atmosphere. The increased water vapour in the lower atmosphere is condensed in regions of upward motion known as convergence zones. This process liberates latent heat of condensation, which in turn provides a major fraction of the energy to drive tropical circulation and is one of the mechanisms responsible for the El Niño/Southern Oscillation phenomenon discussed later in this article.

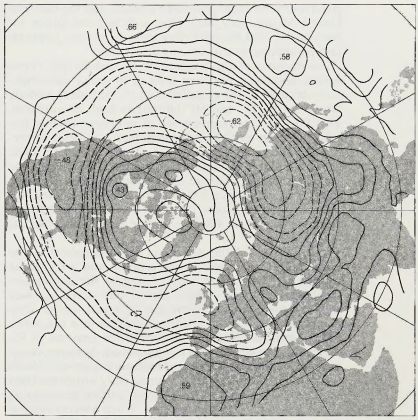


Figure 21: Correlation between sea surface temperature in the eastern equatorial Pacific Ocean (region from the 180th meridian eastward to the coast of South America, from 6° N to 6° S) and the Northern Hemispheric circulation in the middle troposphere during the cold season (November to April) for the period between 1950 and 1979. Contours are at intervals of 0.1; the dashed contours indicate negative correlations.

From J.M. Wallace and Q. Jiang, "On the Observed Structure of the Interannual Variability of the Atmosphere/Ocean Climate System," in *Atmospheric and Oceanic Variability* (1986); Royal Meteorological Society

Sensitivity of the tropical atmosphere to ocean surface temperature

Given the sensitivity of the tropical atmosphere to variations in tropical sea surface temperature, there also has been considerable interest in their influence on extratropical circulation. The sensitivity of the tropospheric circulation to surface temperature in both the tropical Pacific and Atlantic oceans has been shown in theoretical and observational studies alike. Figure 21 demonstrates the correlation between the equatorial ocean surface temperature in the east Pacific (the location of El Niño) and the atmospheric circulation in the middle troposphere during winter. The atmospheric pattern is a characteristic circulation type known as the Pacific-North American (PNA) mode. Such patterns are intrinsic modes of the atmosphere, which may be forced by thermal anomalies in the tropical atmosphere and which in their turn are forced by tropical ocean surface temperature anomalies. As noted earlier, enhanced tropical sea surface temperatures increase evaporation into the atmosphere. In the 1982–83 El Niño event a pattern of circulation anomalies similar to that in Figure 21 occurred throughout the Northern Hemisphere during winter. These modes of the atmosphere, however, account for much less than 50 percent of the variability of the circulation in mid-latitudes, though in certain re-

From C.K. Folland, T.N. Palmer, and D.E. Parker, reprinted by permission from *Nature*, vol. 320, pp. 601–603, copyright © 1986 Macmillan Magazines Ltd.

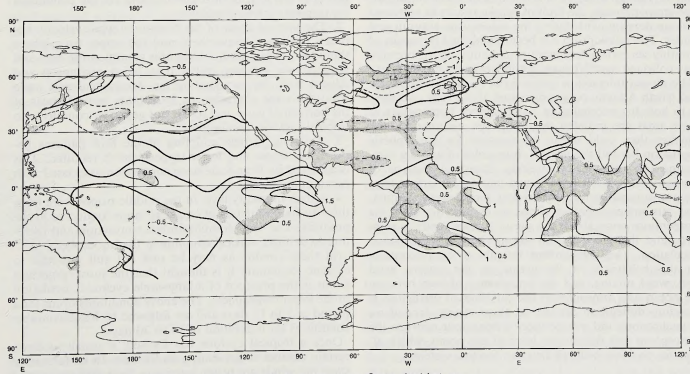


Figure 20: Difference in the average sea surface temperature (for July to September) between the five driest and five wettest rainfall seasons in the Sahel region since 1950. Contours are at intervals of 0.5° C. The shaded areas exhibit significant differences (90 percent); the dashed lines indicate simulated conditions (see text).

gions (northern Japan, southern Canada, and the southern United States), they may have sufficient amplitude for them to be used for predicting seasonal surface temperature perhaps up to two seasons in advance.

The response of the atmosphere to mid-latitude ocean surface anomalies has been difficult to detect unambiguously because of the complexity of the turbulent westerly flow between 20° and 60° latitude in both hemispheres. This flow has many properties of nonlinear chaotic systems and thus exhibits behaviour that is difficult to predict beyond a couple of weeks. The atmosphere alone can exhibit large fluctuations on seasonal and longer time scales without any change in external forcing conditions, such as ocean surface temperature. Notwithstanding this inherent problem, some effects of ocean surface temperature anomalies on the atmosphere have been observed and modeled.

The influence of the oceans on the atmosphere in the mid-latitudes is greatest during autumn and early winter when the ocean mixed layer releases to the atmosphere the large quantities of heat that it has stored up over the previous summer. Anomalies in ocean surface temperature are indicative of either a surplus or a deficiency of heat available to the atmosphere. The response of the atmosphere to ocean surface temperature, however, is not random geographically. The circulation over the North Atlantic and northern Europe during early winter has been found to be sensitive to large ocean surface temperature anomalies south of Newfoundland. When a warm positive anomaly exists in this region, an anomalous surface anticyclone occurs in the central Atlantic at a similar latitude to the temperature anomaly, and an anomalous cyclonic circulation is located over the North Sea, Scandinavia, and central Europe. With colder than normal water south of Newfoundland, the circulation patterns are reversed, producing cyclonic circulation over the central Atlantic and anticyclonic circulation over Europe. The sensitivity of the atmosphere to ocean surface temperature anomalies in this particular region is thought to be related to the position of the overlying storm tracks and jet stream. The region is the most active in the Northern Hemisphere for the growth of storms associated with very large heat fluxes from the surface layer of the ocean.

Another example of a similar type of air-sea interaction event has been documented over the North Pacific Ocean. A statistical seasonal relationship exists between the summer ocean temperature anomaly in the Gulf of Alaska and the atmospheric circulation over the Pacific and North America during the following autumn and winter. The presence of warmer-than-normal ocean surface temperature in the Gulf of Alaska results in increased cyclone development during the subsequent autumn and winter. The relationship has been established by means of monthly sea surface temperature and atmospheric pressure data collected over 30 years in the North Pacific Ocean.

The air-sea interaction events in both the North Pacific and North Atlantic oceans discussed above raise questions as to how the anomalies in ocean surface temperature in these areas are initiated, how they are maintained, and whether they yield useful information for atmospheric prediction beyond the normal time scales of weather forecasting (*i.e.*, one to two weeks). Statistical analysis of previous case studies have shown that ocean surface temperature anomalies initially develop in response to anomalous atmospheric forcing. Once developed, however, the temperature anomaly of the ocean surface tends to reinforce and thereby maintain the anomalous atmospheric circulation. The mechanisms thought to be responsible for this behaviour in the ocean are the surface wind drift, wind mixing, and the interchange of heat between the ocean and atmosphere. The question of prediction is therefore difficult to answer, as these events depend on a synchronous and interconnected behaviour between the atmosphere and the surface layer of the ocean, which allows for positive feedback between the two systems.

Interconnected behaviour between the atmosphere and the oceanic surface layer

FORMATION OF TROPICAL CYCLONES

Tropical cyclones represent still another example of sea-air interactions. These storm systems are known as hurricanes in the North Atlantic and eastern North Pacific and as typhoons in the western North Pacific. The winds of such systems revolve around a centre of low pressure in an anticlockwise direction in the Northern Hemisphere and in a clockwise direction in the Southern Hemisphere. The winds attain velocities in excess of 115 kilometres per hour, or 65 knots, in most cases. Tropical cyclones may last from a few hours to as long as two weeks, the average lifetime being six days.

The oceans provide the source of energy for tropical cyclones both by direct heat transfer from their surface (known as sensible heat) and by the evaporation of water. This water is subsequently condensed within a storm system, thereby releasing latent heat energy. When a tropical cyclone moves over land, this energy is severely depleted and the circulation of the winds is consequently weakened.

Such storms are truly phenomena of the tropical oceans. They originate in two distinct latitude zones, between 4° and 22° S and between 4° and 35° N. They are absent in the equatorial zone between 4° S and 4° N. Most tropical cyclones are spawned on the poleward side of the region known as the intertropical convergence zone (ITCZ).

More than two-thirds of observed tropical cyclones originate in the Northern Hemisphere, and roughly the same proportion occur in the Eastern Hemisphere. The North Pacific has more than one-third of all such storms, while the southeast Pacific and South Atlantic are normally devoid of them. Most northern hemispheric tropical cyclones occur between May and November, with peak periods in August and September. The majority of southern hemispheric cyclones occur between December and April, with peaks in January and February.

Conditions associated with cyclone formation. The formation of tropical cyclones is strongly influenced by the temperature of the underlying ocean or, more specifically, by the thermal energy available in the upper 60 metres of ocean waters. Typically, the underlying ocean should have a temperature in excess of 26° C in this layer. This temperature requirement, however, is only one of five that need to be met for a tropical cyclone to form and develop. The other preconditions relate to the state of the tropical atmosphere between the sea surface and a height of 16 kilometres, the boundary of the tropical troposphere. They can be summarized as follows:

Temperature criterion

1. A deep convergence of air must occur in the troposphere between the surface and a height of seven kilometres that produces a cyclonic circulation in the lower troposphere overlain by an anticyclonic circulation in the upper troposphere. The stronger the inflow, or convergence, of the air, the more favourable are the conditions for tropical cyclone formation.

2. The vertical shear of the horizontal wind velocity between the lower troposphere and the upper troposphere should be at minimum. Under this condition the heat and moisture are retained rather than being exchanged and diluted with the surrounding air. Monsoonal and trade wind flows are characterized by a large vertical shear of the horizontal wind and so are not generally conducive to tropical cyclone development.

3. A strong vertical coupling of the flow patterns between the upper and lower troposphere is required. This is achieved by large-scale deep convection associated with cumulonimbus clouds.

4. A high humidity level in the middle troposphere from three to six kilometres in height is more conducive to the production of deep cumulonimbus convection and therefore to stronger vertical coupling in the troposphere.

All these conditions may be met but still not lead to cyclone formation. It is thought that the most important factor is the presence of a large-scale cyclonic circulation in the lower troposphere. The above conditions occur for a period of 5 to 15 days and are followed by less favourable conditions for a duration of 10 to 20 days.

Once a tropical cyclone has formed, it usually follows certain distinct stages during its lifetime. In its formative stage the winds are below hurricane force and the central pressure is about 1,000 millibars. The formative period is extremely variable in length, ranging from 12 hours to a few days. This stage is followed by a period of inten-

sification, when the central pressure drops rapidly below 1,000 millibars. The winds increase rapidly, and they may achieve hurricane force within a radius of 30 to 50 kilometres of the storm centre. At this stage the cloud and rainfall patterns become well organized into narrow bands that spiral inward toward the centre. In the mature phase the central pressure stops falling and, as a consequence, the winds no longer increase. The region of hurricane force winds, however, expands to occupy a radius of 300 kilometres or more. This expansion is not symmetrical around the storm centre; the strongest winds occur toward the right-hand side of the centre in the direction of the cyclone's path. The period of maturity may last for one to three days. The terminal stage of a tropical cyclone is usually reached when the storm strikes land, followed by a resultant increase in energy dissipation by surface friction and a reduction in its energy supply of moisture. A reduction in moisture input into the storm system may also take place when it moves over a colder segment of the ocean. A tropical cyclone may regenerate in higher latitudes as an extratropical depression, but it loses its identity as a tropical storm in the process. The typical lifetime of a tropical cyclone from its birth to death is about six days.

The paths of tropical cyclones show a wide variation. In both the North Atlantic and the North Pacific, the paths tend to be initially northwestward and then recurve toward the northeast at higher latitudes. It is now known that the tracks of tropical cyclones are largely determined by the large-scale tropospheric flow. This fact opens up the possibility that, with the aid of high-resolution numerical models, accurate predictions of their tracks may become feasible. The development of polar-orbiting and geostationary satellites has made it possible to accurately track cyclones over the remotest areas of the tropical oceans.

Surface cooling

Effects of tropical cyclones on ocean waters. A tropical cyclone can affect the thermal structure and currents in the surface layer of the ocean waters in its path. Cooling of the surface layer occurs in the wake of such a storm. As can be seen in Figure 22, maximum cooling occurs on the right of a hurricane's path in the Northern Hemisphere. In the wake of Hurricane Hilda's passage through the Gulf of Mexico in 1964 at a translational speed of only five knots, the surface waters were cooled by as much as 6° C. Tropical cyclones that have higher translational velocities cause less cooling of the surface. The surface cooling is caused primarily by wind-induced upwelling of cooler water from below the surface layer. The warm surface water is simultaneously transported toward the periphery of the cyclone, where it downwells into the deeper ocean layers (Figure 23). Heat loss across the air-sea interface and the wind-induced mixing of the surface water with those of the cooler subsurface layers make a significant but smaller contribution to surface cooling.

In addition to surface cooling, tropical cyclones may induce large horizontal surge currents and vertical displacements of the thermocline. The surge currents have their largest amplitude at the surface, where they may reach velocities approaching one metre per second. The horizontal currents and the vertical displacement of the thermocline observed in the wake of a tropical cyclone oscillate close to the inertial period. These oscillations remain for a few days after the passage of the storm and spread outward from the rear of the system as an internal wake on the thermocline. The vertical motion may transport nutrients from the deeper layers into the sunlit surface waters, which in turn promotes phytoplankton blooms (*i.e.*, the rapid growth of diatoms and other minute one-celled organisms). The ocean surface temperature normally recovers to its precyclone value within 10 days of a storm's passage.

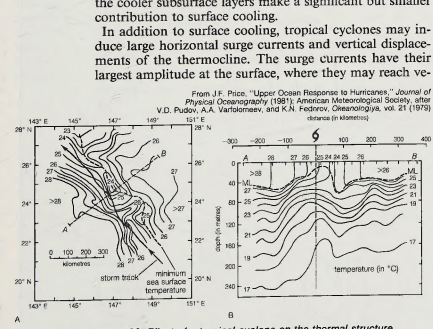


Figure 22: *Effect of a tropical cyclone on the thermal structure of the surface layer.*
(A) The sea surface temperature (in °C) after the passage of the hurricane. (B) Temperature-depth section *AB* across the path of the storm. The base of the surface mixed layer (ML) is indicated by the dashed contour.

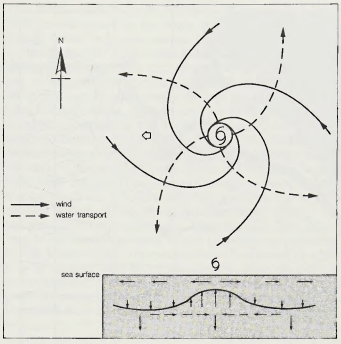


Figure 23: (Top) Surface wind pattern around a symmetrical stationary hurricane in simplified form. (Bottom) Induced water motions.

From G.A. Franceschini and S.Z. El-Sayed, "Effect of Hurricane Inez (1966) on the Hydrography and Productivity of the Western Gulf of Mexico," *Deutsche Hydrographische Zeitschrift* (April 21, 1966)

Influence on atmospheric circulation and rainfall. Tropical cyclones play an important role in the general circulation of the atmosphere, accounting for 2 percent of the global annual rainfall and between 4 and 5 percent of the global rainfall in August and September at the height of the Northern Hemispheric cyclone season. For a local area, the occurrence of a single tropical cyclone can have a major impact on the region's annual rainfall. Furthermore, tropical cyclones contribute approximately 2 percent of the kinetic energy of the general circulation of the atmosphere, some of which is exported from the tropics to higher latitudes.

THE GULF STREAM AND KUROSHIO SYSTEMS

The Gulf Stream. This major current system, as described earlier, is a western boundary current that flows poleward along a boundary separating the warm and more saline waters of the Sargasso Sea to the east from the colder, slightly fresher continental slope waters to the north and west. The warm, saline Sargasso Sea, composed of a water mass known as North Atlantic Central Water, has a temperature that ranges from 8° to 19° C and a salinity between 35.10 and 36.70 parts per thousand. This is one of the two dominant water masses of the North Atlantic Ocean, the other being the North Atlantic Deep Water, which has a temperature of 2.2° to 3.5° C and a salinity between 34.90 and 34.97 parts per thousand, and which occupies the deepest layers of the ocean (generally below 1,000 metres). The North Atlantic Central Water occupies the upper layer of the North Atlantic Ocean between roughly 20° and 40° N. The "lens" of this water is at its lowest depth of 1,000 metres in the northwest Atlantic and becomes progressively shallower to the east and south. To the north it shallows abruptly and outcrops at the surface in winter (Figure 24), and it is at this point that the Gulf Stream is most intense.

Path of the Gulf Stream

The Gulf Stream flows along the rim of the warm North Atlantic Central Water northward from the Florida Straits

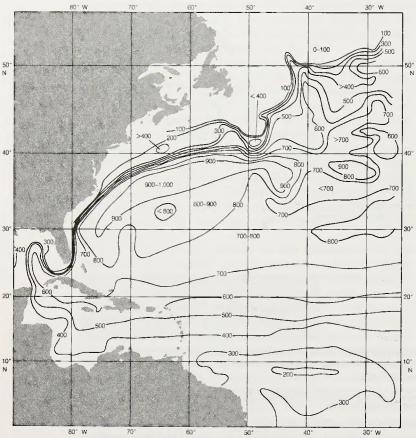


Figure 24: The approximate depth of the North Atlantic Central Water, as indicated by the depth of the 10° C isothermal surface in the western North Atlantic. Depth values are in metres.

From H. Stommel, *The Gulf Stream* (1966); University of California Press and Cambridge University Press, after C.O'D. Iselin, *Papers in Physical Oceanography and Meteorology*, vol. 4, no. 4 (1936)

along the continental slope of North America to Cape Hatteras. There, it leaves the continental slope and turns northeastward as an intense meandering current that extends toward the Grand Banks of Newfoundland. Its maximum velocity is typically between one and two metres per second. At this stage, a part of the current loops back onto itself, flowing south and east. Another part flows eastward toward Spain and Portugal, while the remaining water flows northeastward as the North Atlantic Drift (also called the North Atlantic Current) into the northernmost regions of the North Atlantic Ocean between Scotland and Iceland.

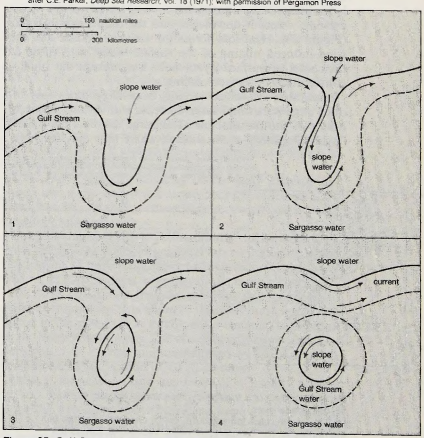
The southward-flowing currents are generally weaker than the Gulf Stream and occur in the eastern lens of the North Atlantic Central Water or the subtropical gyre (see above *Circulation of the ocean waters: Wind-driven circulation: The subtropical gyres*). The circulation to the south on the southern rim of the subtropical gyre is completed by the westward-flowing North Equatorial Current, part of which flows into the Gulf of Mexico; the remaining part flows northward as the Antilles Current. This subtropical gyre of warm North Atlantic Central Water is the hub of the energy that drives the North Atlantic circulation. It is principally forced by the overlying atmospheric circulation, which at these latitudes is dominated by the clockwise circulation of a subtropical anticyclone. This circulation is not steady and fluctuates in particular on its poleward side where extratropical cyclones in the westerlies periodically make incursions into the region. On the western side, hurricanes (during the period from May to November) occasionally disturb the atmospheric circulation. Because of the energy of the subtropical gyre and its associated currents, these short-term fluctuations have little influence on it, however. The gyre obtains most of its energy from the climatological wind distribution over periods of one or two decades. This wind distribution drives a system of surface currents in the uppermost 100 metres of the ocean. Nonetheless, these currents are not simply a reflection of the surface wind circulation as they are influenced by the Coriolis force (see above *Circulation of the ocean waters: Wind-driven circulation: Coriolis effect*). The wind-driven current decays with depth, becoming negligible below 100 metres. The water in this surface layer is transported to the right and perpendicular to the surface wind stress because of the Coriolis force. Hence an eastward-directed wind on the poleward side of the subtropical anticyclone would transport the surface layer of the ocean to the south. On the equatorward side of the anticyclone the trade winds would cause a contrary drift of the surface layer to the north and west. Thus surface waters under the subtropical anticyclone are driven toward the mid-latitudes at about 30° N. These surface waters, which are warmed by solar heating and have a high salinity by virtue of the predominance of evaporation over precipitation at these latitudes, then converge and are forced downward into the deeper ocean.

The effect of the wind on surface currents

Over many decades this process forms a deep lens of warm, saline North Atlantic Central Water. The shape of the lens of water is distorted by other dynamical effects, the principal one being the change in the vertical component of the Coriolis force with latitude known as the beta effect. This effect involves the displacement of the warm water lens toward the west, so that the deepest part of the lens is situated to the north of the island of Bermuda rather than in the central Atlantic Ocean. This warm lens of water plays an important role, establishing as it does a horizontal pressure gradient force in and below the wind-drift current. The sea level over the deepest part of the lens is about one metre higher than outside the lens. The Coriolis force in balance with this horizontal pressure gradient force gives rise to a dynamically induced geostrophic current, which occurs throughout the upper layer of warm water. The strength of this geostrophic current is determined by the horizontal pressure gradient through the slope in sea level. The slope in sea level across the Gulf Stream has been measured by satellite radar altimeter to be one metre over a horizontal distance of 100 kilometres, which is sufficient to cause a surface geostrophic current of one metre per second at 43° N.

Transport of heat and nutrients

The large-scale circulation of the Gulf Stream system is, however, only one aspect of a far more complex and richer structure of circulation. Embedded within the mean flow is a variety of eddy structures that not only put kinetic energy into circulation but also carry heat and other important properties, such as nutrients for biological systems. The best known of these eddies are the Gulf Stream rings, which develop in meanders of the current east of Cape Hatteras. Though the eddies were mentioned as early as



From N. Wells, *The Atmosphere and Ocean: A Physical Introduction* (1986); Taylor & Francis, after C.E. Parker, *Deep Sea Research*, vol. 18 (1971); with permission of Pergamon Press

Figure 25: Gulf Stream ring generation from (1–4) meander formation to separation (see text).

1793 by Jonathan Williams, a grandnephew of Benjamin Franklin, they were not systematically studied until the early 1930s by the oceanographer Phil E. Church. Intensive research programs were finally undertaken during the 1970s. Gulf Stream rings have either warm or cold cores. The warm rings are typically 100 to 300 kilometres in diameter and have a clockwise rotation. They consist of waters from the Gulf Stream and Sargasso Sea and form when the meanders in the Gulf Stream pinch off on its continental slope side (see Figure 25). They move generally westward, flowing at the speed of the slope waters, and are reabsorbed into the Gulf Stream at Cape Hatteras after a typical lifetime of about six months. The cold core rings, composed of a mixture of Gulf Stream and continental slope waters, are formed when the meanders pinch off to the south of the Gulf Stream. They are a little larger than their warm-core counterparts, characteristically having diameters of 200 to 300 kilometres and an anticlockwise rotation. They move generally southwestward into the Sargasso Sea and have lifetimes of one to two years. The cold-core rings are usually more numerous than warm-core rings, typically 10 each year as compared with five warm-core rings annually.

The Kuroshio. This western boundary current is similar to the Gulf Stream in that it produces both warm and cold rings. The warm rings are generally 150 kilometres in diameter and have a lifetime similar to their Gulf Stream counterparts. The cold rings form at preferential sites and in most cases drift southwestward into the Western Pacific Ocean. Occasionally a cold ring has been observed to move northwestward and eventually be reabsorbed into the Kuroshio.

Poleward transfer of heat. A significant characteristic of the large-scale North Atlantic circulation is the poleward transport of heat depicted in Figure 26. Heat is transferred in a northward direction throughout the North Atlantic. This heat is absorbed by the tropical waters of the Pacific and Indian oceans, as well as of the Atlantic, and is then transferred to the high latitudes, where it is finally given up to the atmosphere.

The mechanism for the heat transfer is principally by thermohaline circulation rather than by wind-driven circulation (see above *Circulation of the ocean waters: Thermohaline circulation*). Circulation of the thermohaline type involves a large-scale overturning of the ocean, with warm and saline water in the upper 1,000 metres moving northward and being cooled in the Labrador, Greenland, and Norwegian seas. The density of the water in contact with the atmosphere is increased by surface cooling, and the water subsequently sinks below the surface layer to the lowest depths of the ocean. This water is mixed with the surrounding water masses by a variety of processes to form North Atlantic Deep Water. The water moves slowly southward as the lower limb of the thermohaline circulation. It is this overturning circulation that is responsible for the warm winter climate of northwestern Europe (notably the British Isles and Norway) rather than the horizontal wind-driven circulation discussed above.

Climate moderating effect of thermohaline circulation

From N. Wells, *The Atmosphere and Ocean: A Physical Introduction* (1986): Taylor & Francis, reproduced from H. Stommel, *Proceedings of the National Academy of Science*, vol. 77 (1980), with permission of the National Academy of Science, USA

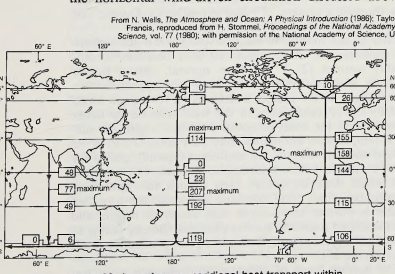


Figure 26: Annual mean meridional heat transport within the oceans. Amount of heat transport is given in units of 10^{13} watts.

The North Atlantic Drift, which is an extension of the Gulf Stream system to the south, provides this northward flow of warm and saline waters into the polar seas. This feature makes the circulation of the North Atlantic Ocean uniquely different from that of the Pacific Ocean, which has a less effective thermohaline circulation. Although there is a northward transfer of heat in the North Pacific, the subtropical wind-driven gyre in the upper ocean is mainly responsible for it. Thus the Kuroshio on the western boundary of the North Pacific gyre is principally driven by the surface wind circulation of the North Pacific.

Studies of the sediment cores obtained from the ocean floor have indicated that the ocean surface temperature was as much as 10° C cooler than today in the northernmost region of the North Atlantic Ocean during the last glacial maximum some 18,000 years ago. This difference in surface temperature would indicate that the warm North Atlantic Drift was much reduced compared to what it is at present, and hence the thermohaline circulation was considerably weaker. In contrast, the Gulf Stream was probably more intense than it is today and exhibited a large shift from its present path to an eastward flow at 40° N. (N.C.W.)

EL NIÑO/SOUTHERN OSCILLATION AND CLIMATIC CHANGE

As was explained earlier, the oceans can moderate the climate of certain regions. Not only do they affect such geographic variations, but they also influence temporal changes in climate. The time scales of climate variability range from a few years to millions of years and include the so-called ice age cycles that repeat every 20,000 to 40,000 years, interrupted by interglacial periods of "optimum" climate, such as the present. The climatic modulations that occur at shorter scales include such periods as the Little Ice Age from the early 16th to the mid-19th centuries, when the global average temperature was approximately 1° C lower than it is today. Several climate fluctuations on the scale of decades have occurred in the 20th century, such as warming from 1910 to 1940, cooling from 1940 to 1970, and the warming trend since 1970.

Although many of the mechanisms of climate change are understood, it is usually difficult to pinpoint the specific causes. Scientists acknowledge that climate can be affected by factors external to the land-ocean-atmosphere climate system, such as variations in solar brightness, the shading effect of aerosols injected into the atmosphere by volcanic activity, or the increased atmospheric concentration of "greenhouse" gases (*e.g.*, carbon dioxide, nitrous oxide, methane, and chlorofluorocarbons) produced by human activities. However, none of these factors explain the periodic variations observed during the 20th century, which may simply be manifestations of the natural variability of climate. The existence of natural variability at many time scales makes the identification of causative factors such as human-induced warming more difficult. Whether change is natural or caused, the oceans play a key role and have a moderating effect on influencing factors.

The El Niño phenomenon. The shortest, or interannual, time scale relates to natural variations that are perceived as years of unusual weather—*e.g.*, excessive heat, drought, or storminess. Such changes are so common in many regions that any given year is about as likely to be considered as exceptional as typical. The best example of the influence of the oceans on interannual climate anomalies is the occurrence of El Niño conditions in the eastern Pacific Ocean at irregular intervals of about 3–10 years. The stronger El Niño episodes of enhanced ocean temperatures (2°–8° C above normal) are typically accompanied by altered weather patterns around the globe, such as droughts in Australia, northeastern Brazil, and the highlands of southern Peru, excessive summer rainfall along the coast of Ecuador and northern Peru, severe winter storminess along the coast of central Chile, and unusual winter weather along the west coast of North America.

Altered weather patterns accompanying stronger El Niño episodes

The effects of El Niño have been documented in Peru since the Spanish conquest in 1525. The Spanish term "la corriente de El Niño" was introduced by fishermen of the Peruvian port of Paita in the 19th century; it refers to a warm, southward ocean current that temporarily dis-

places the normally cool, northward-flowing Humboldt, or Peru, Current. (The name is a pious reference to the Christ child, chosen because of the typical appearance of the countercurrent during the Christmas season.) By the end of the 19th century Peruvian geographers recognized that every few years this countercurrent is more intense than normal, extends farther south, and is associated with torrential rainfall over the otherwise dry northern desert. The abnormal countercurrent also was observed to bring tropical debris, as well as such flora and fauna as bananas and aquatic reptiles, from the coastal region of Ecuador farther north. Increasingly during the 20th century, El Niño has come to connote an exceptional year rather than the original annual event.

As Peruvians began to exploit the guano of marine birds for fertilizer in the early 20th century, they noticed El Niño-related deteriorations in the normally high marine productivity of the coast of Peru as manifested by large reductions in the bird populations that depend on anchovies and sardines for sustenance. The preoccupation with El Niño increased after mid-century, as the Peruvian fishing industry rapidly expanded to exploit the anchovies directly. (Fish meal produced from the anchovies was exported to industrialized nations as a feed supplement for livestock.) By 1971 the Peruvian fishing fleet had become the largest in its history; it had extracted very nearly 13 million metric tons of anchovies in that year alone. Peru was catapulted into first place among fishing nations, and scientists expressed serious concern that fish stocks were being depleted beyond self-sustaining levels, even for the extremely productive marine ecosystem of Peru. The strong El Niño of 1972–73 captured world attention because of the drastic reduction in anchovy catches to a small fraction of prior levels. The anchovy catch did not return to previous levels, and the effects of plummeting fish meal exports reverberated throughout the world commodity markets.

El Niño was only a curiosity to the scientific community in the first half of the 20th century, thought to be geographically limited to the west coast of South America. There was little data, mainly gathered coincidentally from foreign oceanographic cruises, and it was generally believed that El Niño occurred when the normally northward coastal winds off Peru, which cause the upwelling of cool, nutrient-rich water along the coast, decreased, ceased, or reversed in direction. When systematic and extensive oceanographic measurements were made in the Pacific in 1957–58 as part of the International Geophysical Year, it was found that El Niño had occurred during the same period and was also associated with extensive warming over most of the Pacific equatorial zone. Eventually tide-gauge and other measurements made throughout the tropical Pacific showed that the coastal El Niño was but one manifestation of basinwide ocean circulation changes that occur in response to a massive weakening of the westward-blowing trade winds in the western and central equatorial Pacific and not to localized wind anomalies along the Peru coast.

The Southern Oscillation. The wind anomalies are a manifestation of an atmospheric counterpart to the oceanic El Niño. At the turn of the century, the British climatologist Gilbert Walker set out to determine the connections between the Asian monsoon and other climatic fluctuations around the globe in an effort to predict unusual monsoon years that bring drought and famine to the Asian sector. Unaware of any connection to El Niño, he discovered a coherent interannual fluctuation of atmospheric pressure over the tropical Indo-Pacific region, which he termed the Southern Oscillation (SO). During years of reduced rainfall over northern Australia and Indonesia, the pressure in that region (*e.g.*, at what are now Darwin and Jakarta) was anomalously high and wind patterns were altered. Simultaneously, in the eastern South Pacific pressures were unusually low, negatively correlated with those at Darwin and Jakarta. A Southern Oscillation Index (SOI), based on pressure differences between the two regions (east minus west), showed low, negative values at such times, which were termed the "low phase" of the SO. During more normal "high-phase" years, the pressures were low over Indonesia and high in the eastern Pacific, with high, positive values of the SOI. In papers published during the 1920s and '30s, Walker gave statistical evidence for widespread climatic anomalies around the globe being associated with the SO pressure "seesaw."

Interannual fluctuations of atmospheric pressure

In the 1950s, years after Walker's investigations, it was noted that the low-phase years of the SOI corresponded with periods of high ocean temperatures along the Peruvian coast, but no physical connection between the SO and El Niño was recognized until Jacob Bjerknes, in the early 1960s, tried to understand the large geographic scale of the anomalies observed during the 1957–58 El Niño event. Bjerknes, a meteorologist, formulated the first conceptual model of the large-scale ocean-atmosphere interactions that occur during El Niño episodes. His model has been refined through intensive research since the early 1970s.

During a year or two prior to an El Niño event (high-phase years of the SO), the westward trade winds typically blow more intensely along the equator in the equatorial Pacific, causing warm upper-ocean water to accumulate in a thickened surface layer in the western Pacific where sea level rises. Meanwhile, the stronger, upwelling-favourable winds in the eastern Pacific induce colder surface water and lowered sea levels off South America. Toward the end of the year preceding an El Niño, the area of intense tropical storm activity over Indonesia migrates eastward toward the equatorial Pacific west of the international date line (which corresponds in general to the 180th meridian of longitude), bringing episodes of eastward wind reversals to that region of the ocean. These wind bursts excite extremely long ocean waves, known as Kelvin waves (imperceptible to an observer), that propagate eastward toward the coast of South America, where they cause the upper ocean layer of relatively warm water to thicken and sea level to rise.

The tropical storms of the western Pacific also occur in other years, though less frequently, and produce similar Kelvin waves, but an El Niño event does not result and the waves continue poleward along the coast toward Chile and California, detectable only in tide-gauge measurements. Something else occurs prior to an El Niño that is not fully understood: as the Kelvin waves travel eastward along the equator, an anomalous eastward current carries warm western Pacific water farther east, and the warm surface layer deepens in the central equatorial Pacific (east of the international dateline). Additional surface warming takes place as the upwelling-favourable winds bring warmer subsurface water to the surface. (The subsurface water is warmer now, rather than cooler, because the overlying layer of warmer water is now significantly deeper than before.) The anomalous warming creates conditions favourable for the further migration of the tropical storm centre toward the east, giving renewed vigour to eastward winds, more Kelvin waves, and additional warming. Each increment of anomalies in one medium (*e.g.*, the ocean) induces further anomalies in the other (the atmosphere) and vice versa, giving rise to an unstable growth of anomalies through a process of positive feedbacks. During this time, the SO is found in its low phase.

After several months of these unstable ocean-atmosphere interactions, the entire equatorial zone becomes considerably warmer (2°–5° C) than normal, and a sizable volume of warm upper ocean water is transported from the western to the eastern Pacific. As a result, sea levels fall by 10–20 centimetres in the west and rise by larger amounts off the coast of South America, where sea surface temperature anomalies may vary from 2° to 8° C above normal. Anomalous conditions typically persist for 10–14 months before returning to normal. The warming off South America occurs even though the upwelling-favourable winds there continue unabated: the upwelled water is warmer now, rather than cooler as before, and its associated nutrients are less plentiful, thereby failing to sustain the marine ecosystem at its prior productive levels (see Figure 27).

The current focus of oceanographic research is on understanding the circumstances leading to the demise of the El Niño event and the onset of another such event

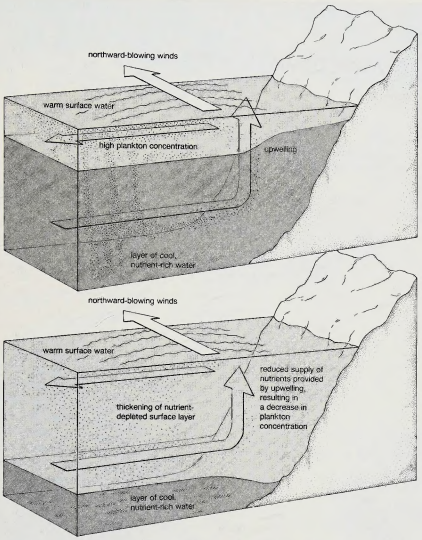


Figure 27: *The upwelling process along the coast of Peru.* A thermocline and nutricline separate the warm, nutrient-deficient upper layer from the cool, enriched layer below. Under normal conditions (top), these interfaces are shallow enough that coastal winds can induce upwelling of the lower-layer nutrients to the surface, where they support an abundant ecosystem. During an El Niño event (bottom), the upper layer thickens so that the upwelled water contains fewer nutrients, thus contributing to a collapse of marine productivity (see text).

After A. Carroll, based on an original drawing by R.T. Barber, copyright National Geographic Society 1984 in T.Y. Canby, *National Geographic* (February, 1984)

Generation of Rossby waves

several years later. The most widely held hypothesis is that a second class of long equatorial ocean waves—Rossby waves with a shallow surface layer—is generated by the El Niño and that they propagate westward to the landmasses of Asia. There, the Rossby waves reflect off the Asian coast eastward along the equator in the form of upwelling Kelvin waves, resulting in a thinning of the upper ocean warm layer and a cooling of the ocean as the winds bring deeper, cooler water to the surface. This process is thought to initiate one to two years of colder-than-average conditions until Rossby waves of a contrary sense (*i.e.*, with a thickened surface layer) are again generated, functioning as a switching mechanism, this time to start another El Niño sequence.

Another goal of scientists is to understand climate change on the scale of centuries or longer and to make projections about the changes that will occur within the next few generations. Yet, determinations of current climatic trends from recent data are made difficult by natural variability at shorter time scales, such as the El Niño phenomenon. Many scientists are attempting to understand the mechanisms of change during an El Niño event from improved global measurements so as to determine how the ocean-atmosphere engine operates at longer time scales. Others are studying prehistoric records preserved in trees, sediments, and fossil corals in an effort to reconstruct past variations, including those like the El Niño. Their aim is to remove such short-term variations so as to be able to make more accurate estimates of long-term trends.

(D.B.En.)

Ocean basins

The first major undersea survey was undertaken during the 1870s, but it was not until the last half of the 20th century that scientists began to learn what lies beneath the ocean surface in any detail. It has been determined that the ocean basins, which hold the vast quantity of water that covers nearly three-quarters of the Earth's surface, have an average depth of almost four kilometres. A number of major features of the basins depart from this average, as, for example, the mountainous ocean ridges, deep-sea trenches, and jagged, linear fracture zones (see Figure 28). Other significant features of the ocean floor include aseismic ridges, abyssal hills, and seamounts and guyots. The basins also contain a variable amount of sedimentary fill that is thinnest on the ocean ridges and usually thickest near the continental margins.

Notable topographic features of the ocean basins

While the ocean basins lie much lower than sea level, the continents stand high—about one kilometre above sea level. The physical explanation for this condition is that the continental crust is light and thick, whereas the oceanic crust is dense and thin. Both the continental and oceanic crust lie over a more uniform layer called the mantle. As an analogy, one can think of a thick piece of styrofoam and a thin piece of wood floating in a tub of water. The styrofoam rises higher out of the water than the wood.

The ocean basins are transient features over geologic time, changing shape and depth while the process of plate tectonics proceeds. The surface layer of the Earth, the lithosphere, consists of a number of rigid plates that are in continual motion. The boundaries between the lithospheric plates form the principal relief features of the ocean basins: the crests of oceanic ridges are spreading centres where two plates move apart from each other at a rate of several centimetres per year. Molten rock material wells up from the underlying mantle into the gap between the diverging plates and solidifies into oceanic crust, thereby creating new ocean floor. At the deep-sea trenches, two plates converge, with one plate sliding down under the other into the mantle where it is melted. Thus, for each segment of new ocean floor created at the ridges, an equal amount of old oceanic crust is destroyed at the trenches, or so-called subduction zones (see below *Deep-sea trenches* and also the article PLATE TECTONICS). It is for this reason that the oldest segment of ocean floor, found in the far western Pacific, is apparently only about 200 million years old, even though the age of the Earth is estimated to be at least 4.6 billion years.

Factors governing seafloor relief and topography

The dominant factors that govern seafloor relief and topography are the thermal properties of the oceanic plates, tensional forces in the plates, volcanic activity, and sedimentation. In brief, the oceanic ridges rise about two kilometres above the seafloor because the plates near these spreading centres are warm and thermally expanded. In contrast, plates in the subduction zones are generally cooler. Tensional forces resulting in plate divergence at the spreading centres also create block-faulted mountains and abyssal hills, which trend parallel to the oceanic ridges. Seamounts and guyots, as well as abyssal hills and most aseismic ridges, are produced by volcanism. Continuing sedimentation throughout the ocean basin serves to blanket and bury many of the faulted mountains and abyssal hills with time. Erosion plays a relatively minor role in shaping the face of the deep seafloor, in contrast to the continents. This is because deep ocean currents are generally slow (they flow at less than 50 centimetres per second) and lack sufficient power.

EXPLORATION OF THE OCEAN BASINS

Mapping the characteristics of the ocean basin has been difficult for several reasons. First, the oceans are not easy to travel over; second, until recent times navigation has been extremely crude, so that individual observations have been only loosely correlated with one another; and, finally, the oceans are opaque to light—*i.e.*, the deep seafloor cannot be seen from the ocean surface. Modern technology has given rise to customized research vessels, satellite and electronic navigation, and sophisticated acoustic instruments that have mitigated some of these problems.

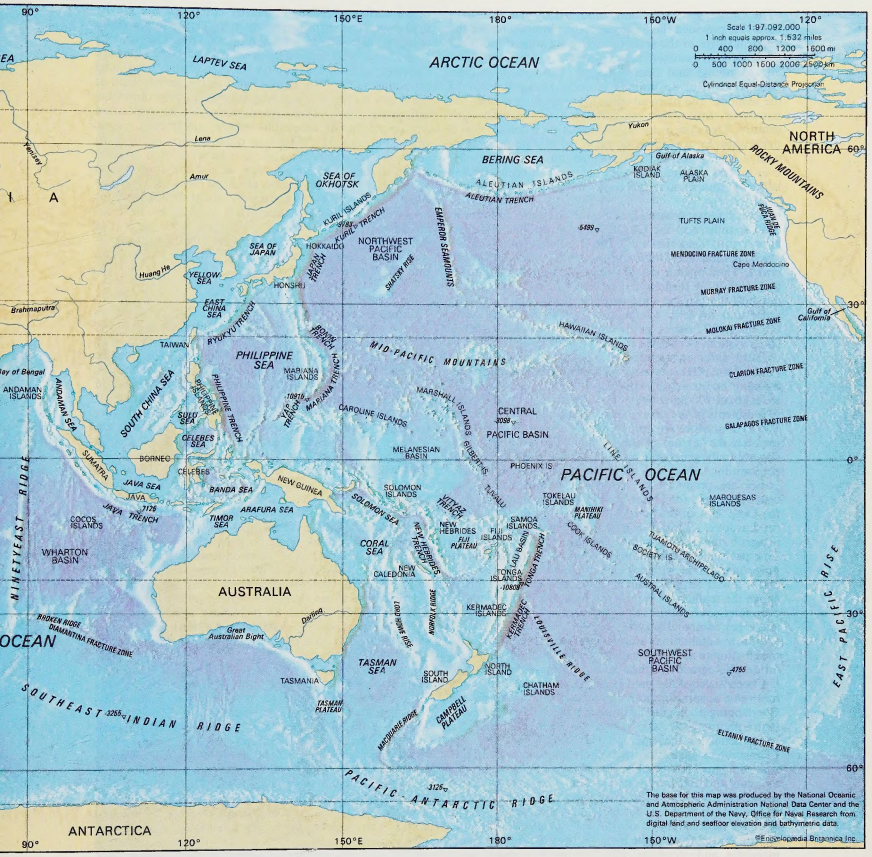


Figure 28: Major features of the ocean basins.

MAP INDEX

Physical features

Agulhas Plateau . 40 00 S 26 00 E
Alaska Plain .... 55 00 N 143 00 W
Aleutian Islands . . 52 55 N 172 28 W
Aleutian Trench . . 51 00 N 170 00 W
Anadaman Islands ..... 12 30 N 92 45 E
Angola Basin .... 15 00 S 3 00 E
Argentine Basin ..... 45 00 S 45 00 W
Ascension, *island* ..... 7 57 S 14 22 W
Atlantic-Indian Ridge ..... 53 00 S 15 00 E
Atlantis Fracture Zone ..... 30 00 N 42 00 W
Austral (Tubuai) Islands ..... 23 00 S 150 00 W
Azores, *islands* . . 38 30 N 28 00 W
Bermuda, *island* . 32 18 N 64 45 W
Bermuda Rise . . . 32 30 N 65 00 W
Bonin Trench . . . 23 00 N 145 00 E
Brazil Basin ..... 15 00 S 25 00 W
British Isles ..... 54 00 N 4 00 W
Broken Ridge, *plateau* ..... 31 30 S 95 00 E
Campbell Plateau ..... 51 00 S 170 00 E
Carlsberg Ridge . 6 00 N 61 00 E
Central Pacific Basin ..... 5 00 N 175 00 W
Chagos-Laccadive Plateau ..... 3 00 N 73 00 E
Charlie-Gibbs Fracture Zone . . 52 45 N 35 30 W
Chile Rise ..... 40 00 S 90 00 W
Clarion Fracture Zone ..... 18 00 N 122 00 W
Cook Islands .... 20 00 S 158 00 W
Diamantina Fracture Zone . . 36 00 S 105 00 E
East Pacific Rise . 20 00 S 115 00 W
Eltanin Fracture Zone ..... 54 00 S 130 00 W
Emperor Seamounts .... 42 00 N 170 00 E
Enderby Plain . . . 60 00 S 40 00 E
Falkland Plateau . 51 00 S 50 00 W
Fiji Plateau ..... 17 00 S 173 00 E
Galapagos Fracture Zone . . 10 00 N 120 00 W
Guinea Rise .... 8 00 S 0 00
Hawaiian Islands . 24 00 N 167 00 W
Iceland, *island* . . . 65 00 N 18 00 W
Japan Trench .... 37 00 N 143 00 E
Java Trench .... 10 30 S 110 00 E
Juan de Fuca Ridge ..... 46 00 N 130 00 W
Kerguelen Islands ..... 49 15 S 69 10 E
Kerguelen Plateau ..... 55 00 S 75 00 E
Kermadec Trench ..... 30 00 S 177 00 W
Kodiac Island . . . 57 20 N 153 22 W
Kuril Trench .... 47 00 N 155 00 E
Lau Basin ..... 20 00 S 177 00 W
Line Islands ..... 0 00 157 00 W
Lord Howe Rise . 32 00 S 162 00 E



The base for this map was produced by the National Oceanic and Atmospheric Administration National Data Center and the U.S. Department of the Navy, Office for Naval Research from digital land and seafloor elevation and bathymetric data.

©Encyclopedia Britannica Inc.

Louisville Ridge . . 31 00 s 172 30 w
Macquarie Ridge . 57 00 s 159 00 e
Madagascar Plateau . . . . . 30 00 s 45 00 e
Manihiki Plateau . . 11 00 s 164 00 w
Mariana Trench . . 15 00 n 147 30 e
Mascarene Plateau . . . . . 10 00 s 60 00 e
Melanesian Basin . . . . . 0 05 n 160 35 e
Mendocino Fracture Zone . . 40 00 n 145 00 w
Mid-Atlantic Ridge . . . . . 0 00 20 00 w
Mid-Indian Ridge . 12 00 s 66 00 e
Mid-Pacific Mountains . . . . . 20 00 n 178 00 e
Molokai Fracture Zone . . . . . 23 00 n 148 00 w
Mozambique Plateau . . . . . 32 00 s 35 00 e
Murray Fracture Zone . . . . . 34 00 n 135 00 w
Nazca Ridge . . . . 22 00 s 82 00 w
New Guinea, *island* . . . . . 5 00 s 140 00 e
New Hebrides, *islands* . . . . . 16 00 s 167 00 e
New Hebrides Trench . . . . . 22 30 n 170 00 e
Ninetyeast Ridge . 4 00 s 90 00 e
Norfolk Ridge . . . 29 00 s 168 00 e
Northwest Pacific Basin . . . . . 40 00 n 155 00 e
Pacific-Antarctic Ridge . . . . . 62 00 s 157 00 w
Peru-Chile Trench . . . . . 20 00 s 73 00 w
Philippine Trench . . . . . 9 00 n 127 00 e
Reykjanes Ridge . 62 00 n 27 00 w
Rio Grande Rise . 31 00 s 35 00 w
Ryukyu Trench . . 24 45 n 128 00 e
Shatsky Rise . . . . 35 00 n 160 00 e
Sohm Plain . . . . . 36 00 n 55 00 w
Solomon Islands . 8 00 s 159 00 e
Somali Basin . . . . 0 00 52 00 e
South Sandwich Trench . . . . . 56 30 s 25 00 w
Southeast Indian Ridge . . . . . 50 00 s 110 00 e
Southwest Indian Ridge . . . . . 43 00 s 40 00 e
Southwest Pacific Basin . . . 40 00 s 150 00 w
Tasman Plateau . 45 00 s 147 00 e
Tonga Trench . . . 20 00 s 173 00 w
Tristan da Cunha, *islands* . 37 15 s 12 30 w
Tuamotu Archipelago . . . . 19 00 s 142 00 w
Tubuai, see Austral Islands
Tufts Plain . . . . . 47 00 n 140 00 w
Verna Fracture Zone . . . . . 11 00 n 44 00 w
Vema Trench . . . 9 00 s 67 20 e
Vityaz Trench . . . 10 00 s 170 00 e
Walvis Ridge . . . . 28 00 s 3 00 e
Weddell Plain . . . 65 00 s 20 00 w
Wharton Basin . . 21 00 s 100 00 e
Yap Trench . . . . . 8 30 n 138 00 e

The Challenger Expedition, mounted by the British in 1872–76, provided the first systematic view of a few of the major features of the seafloor. Scientists aboard the HMS *Challenger* determined ocean depths by means of wire-line soundings and discovered the Mid-Atlantic Ridge. Dredges brought up samples of rocks and sediments off the seafloor. The main advance in mapping, however, did not occur until sonar was developed in the early 20th century. This system for detecting the presence of objects underwater by acoustic echo provided marine researchers with a highly useful tool, since sound can be detected over several thousands of kilometres in the ocean (visible light, by comparison, can only penetrate 100 metres or so of water).

Modern sonar systems include the Seabeam multibeam echo sounder and the GLORIA scanning sonar (see EXPLORATION: *Undersea exploration: Methodology and instrumentation: Exploration of the seafloor and the Earth's crust*). They operate on the principle that the depth (or distance) of the seafloor can be determined by multiplying one-half the elapsed time between a downgoing acoustic pulse and its echo by the speed of sound in seawater (about 1,500 metres per second). Such multifrequency sonar systems permit the use of different pulse frequencies to meet different scientific objectives. Acoustic pulses of 12 kilohertz (kHz), for example, are normally employed to measure ocean depth, while lower frequencies—3.5 kHz to less than 100 hertz (Hz)—are used to map the thickness of sediments in the ocean basins. Very high frequencies of 100 kHz or more are employed in side-scanning sonar to measure the texture of the seafloor. The acoustic pulses are normally generated by piezoelectric transducers. For determining subbottom structure, low-frequency acoustic pulses are produced by explosives, compressed air, or water-jet implosion. Near-bottom sonar systems, such as the Deep Tow of the Scripps Institution of Oceanography (in La Jolla, Calif., U.S.), produce even more detailed images of the seafloor and subbottom structure. The Deep Tow package contains both echo sounders and side-scanning sonars, along with associated geophysical instruments, and is towed behind a ship at slow speed 10 to 100 metres above the seafloor. It yields very precise measurements of even finer-scale features than are resolvable with Seabeam and other comparable systems.

Another notable instrument system is ANGUS, a deep-towed camera sled that can take thousands of high-resolution photographs of the seafloor during a single day. It has been successfully used in the detection of hydrothermal vents at spreading centres (see below *Oceanic ridges*). Overlapping photographic images make it possible to construct photomosaic strips about 10–20 metres wide that reveal details on the order of centimetres.

Three major navigation systems are in use in modern marine geology. These include electromagnetic systems such as loran and Earth-orbiting satellites (see EXPLORATION: *Undersea exploration: Basic elements of undersea exploration: Navigation*). Acoustic transponder arrays of two or more stations placed on the seafloor a few kilometres apart are used to navigate deeply towed instruments, submersibles, and occasionally surface research vessels when detailed mapping is conducted in small areas. These systems measure the distance between the instrument package and the transponder sites and, using simple geometry, compute fixes accurate to a few metres. Although the individual transponders can be used to determine positions relative to the array with great accuracy, the preciseness of the position of the array itself depends on which system is employed to locate it.

Satellite measurements

Such Earth-orbiting satellites as SEASAT and GEOSAT have uncovered some significant topographic features of the ocean basins. SEASAT, launched in 1978, carried a radar altimeter into orbit. This device was used to measure the distance between the satellite path and the surfaces of the ocean and continents to 0.1 metre. The measurements revealed that the shape of the ocean surface is warped by seafloor features: massive seamounts cause the surface to bulge over them owing to gravitational attraction. Similarly, the ocean surface downwarps occur over trenches. Using these satellite measurements of the ocean surface, William F. Haxby computed the gravity field there. The resulting gravity map (Figure 29) provides comprehensive coverage of the ocean surface on a 5′ by 5′ grid (five nautical miles on each side at the equator). Coverage as complete as this is not available from echo soundings made from ships. Because the gravity field at the ocean surface is a highly sensitive indicator of marine topography, this map reveals various previously uncharted features, including seamounts, ridges, and fracture zones, while improving the detail on other known features. In addition, the gravity map shows a linear pattern of gravity anomalies that cut obliquely across the grain of the topography. These anomalies are most pronounced in the Pacific basin; they are apparently about 100 kilometres across and some 1,000 kilometres long. They have an amplitude of approximately 10 milligals (0.001 percent of the Earth's gravity attraction) and are aligned west-northwest—very close to the direction in which the Pacific Plate moves over the mantle below.

OCEANIC CRUST

Structure and composition. The oceanic crust differs from the continental crust in several ways: it is thinner, denser, younger, of different chemical composition, and created in a different plate-tectonic setting. The oceanic crust is formed at spreading centres on the oceanic ridges, whereas continental crust is formed above the subduction zones. The oceanic crust is about six kilometres thick. It is composed of several layers, not including the overlying sediment. The topmost layer, about 500 metres thick, includes lavas of basaltic composition (*i.e.*, rock material consisting largely of plagioclase [feldspar] and pyroxene). The lavas are generally of two types: pillow lavas and sheet flows. Pillow lavas appear to be shaped exactly as the name implies—like large overstuffed pillows about one metre in cross section and one to several metres long. They commonly form small hills tens of metres high at the spreading centres. Sheet flows have the appearance of wrinkled bed sheets. They commonly are thin (only about 10 centimetres thick) and cover a broader area than pillow lavas. There is evidence that sheet flows are erupted at higher temperatures than those of the pillow variety. On the East Pacific Rise at 8° S latitude, a series of sheet flow eruptions (possibly since the mid-1960s) have covered more than 220 square kilometres of seafloor to an average depth of 70 metres.

Below the lava is a layer composed of feeder, or sheeted, dikes that measures more than one kilometre thick. Dikes are fractures that serve as the plumbing system for transporting magmas (molten rock material) to the seafloor to produce lavas. They are about one metre wide, subvertical, and elongate along the trend of the spreading centre where they formed, and they abut one another's sides—hence the term sheeted. These dikes are also of basaltic composition. There are two layers below the dikes totaling about 4.5 kilometres in thickness. Both of these include gabbros, which are essentially basalts with coarser mineral grains. These gabbro layers are thought to represent the magma chambers, or pockets of lava, that ultimately erupt on the seafloor. The upper gabbro layer is isotropic (uniform) in structure. In some places, this layer includes pods of plagiogranite, a differentiated rock richer in silica than gabbro. The lower gabbro layer has a stratified structure and evidently represents the floor or sides of the magma chamber. This layered structure is called cumulate, meaning that the layers (which measure up to several metres thick) result from the sedimentation of minerals out of the liquid magma. The layers in the cumulate gabbro have less silica but are richer in iron and magnesium than the upper portions of the crust. Olivine, an iron-magnesium silicate, is a common mineral in the lower gabbro layer.

The oceanic crust lies atop the Earth's mantle, as does the continental crust. Mantle rock is composed mostly of peridotite, which consists primarily of the mineral olivine with small amounts of pyroxene and amphibole.

Investigations of the oceanic crust. Knowledge of the structure and composition of the oceanic crust comes from several sources. Bottom sampling during early exploration brought up all varieties of the above-mentioned

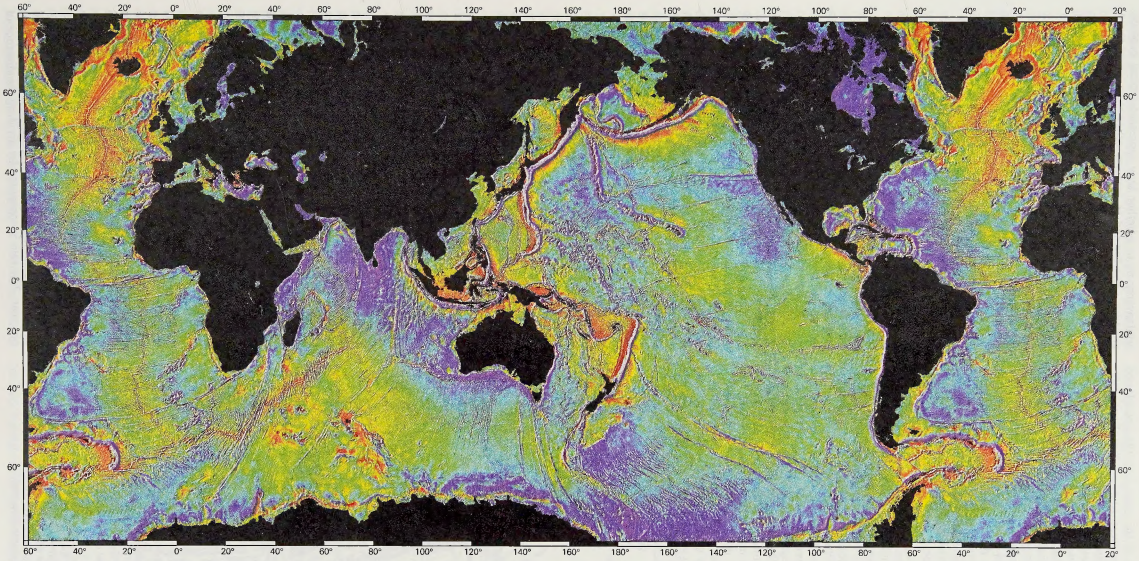


Figure 29: Gravity map of the world's ocean basins, compiled from Seasat satellite data (see text).

D.T. Sandwell from Scripps Institution of Oceanography, W.H.F. Smith from National Oceanic and Atmospheric Administration/National Ocean Service/Office of Ocean & Earth Science/Geoscience Lab

rocks, but the structure of the crust and the abundance of the constituent rocks were unclear. Simultaneously, seismic refraction experiments enabled researchers to determine the layered nature of the oceanic crust. These experiments involve measuring the travel times of seismic waves generated by explosions (*e.g.*, dynamite blasts) set off over distances of several tens of kilometres. The results of early refraction experiments revealed the existence of two layers beneath the sediment cover. More sophisticated experiments and analyses led to dividing these layers into two parts, each with a different seismic wave velocity, which increases with depth. The seismic velocity is a kind of fingerprint that can be attributed to a limited number of rock types. Sampled rock data and seismic results were combined to yield a model for the structure and composition of the crust.

Study of ophiolites

Great strides in understanding the oceanic crust were made by the study of ophiolites. These are slices of the ocean floor that have been thrust above sea level by the action of plate tectonics. In various places in the world, the entire sequence of oceanic crust and upper mantle is exposed. These areas include, among others, Newfoundland and the Pacific Coast Ranges of California, the island of Cyprus in the Mediterranean Sea, and the mountains in Oman on the southeastern tip of the Arabian Peninsula. Ophiolites reveal the structure and composition of the oceanic crust in astonishing detail. Also, the process of crustal formation and hydrothermal circulation, as well as the origin of marine magnetic anomalies (see below), can be studied with comparative clarity. Although it is clear that ophiolites are of marine origin, there is some controversy as to whether they represent typical oceanic crust or crust formed in settings other than an oceanic spreading centre—behind island arcs, for example.

From J.P. Kennett, *Marine Geology* (1982); Prentice-Hall, Inc., Englewood Cliffs, N.J., after R.G. Mason and A.D. Raff, *Geological Society American Bulletin*, vol. 72 (1961); courtesy Geological Society of America

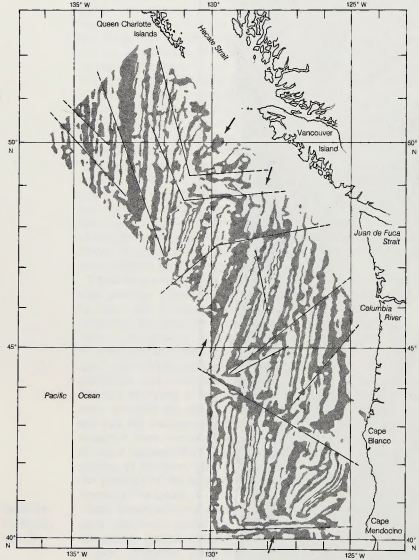


Figure 30: *Magnetic anomalies off the west coast of the United States.*
The gray areas represent positive anomalies formed during periods of normal polarity of the Earth's magnetic field. The white (open) areas mark negative anomalies formed when the field was reversed. The arrows show the central magnetic anomaly over three spreading centres (from north to south): Explorer, Juan de Fuca, and Gorda (see text).

The age of the oceanic crust does not go back farther than about 200 million years. Such crust is being formed today at oceanic spreading centres. Many ophiolites are much older than the oldest oceanic crust, demonstrating continuity of the formation processes over hundreds of millions of years. Methods that may be used to determine the age of the crustal material include direct dating of rock samples by radiometric dating (measuring the relative abundances of a particular radioactive isotope and its daughter isotopes in the samples) or by the analyses of fossil evidence, marine magnetic anomalies, or ocean depth. Of these, magnetic anomalies deserve special attention.

A marine magnetic anomaly is a variation in strength of the Earth's magnetic field caused by magnetism in rocks of the ocean floor. Marine magnetic anomalies typically represent 1 percent of the total geomagnetic field strength. They can be stronger ("positive") or weaker ("negative") than the average total field. Also, the magnetic anomalies occur in long bands that run parallel to spreading centres for hundreds of kilometres (Figure 30) and may reach up to a few tens of kilometres in width.

Marine magnetic anomalies were first discovered off the coast of the western United States in the late 1950s and completely baffled scientists. The anomalies were charted from southern California to northern Washington and out several hundred kilometres. Victor Vacquier, a geophysicist, noticed that these linear anomalies ended at the fracture zones mapped in this area. In addition, he noticed that they had unique shapes, occurred in a predictable sequence across their trends, and could be correlated across the fracture zones. Soon thereafter, linear magnetic anomalies were mapped over the Reykjanes Ridge south of Iceland. They were found to occur on both sides of the ridge crest and parallel to it. Simultaneously, Alan Cox and several other American geophysicists documented evidence that the Earth's magnetic field had reversed in the past: the north magnetic pole had been the south magnetic pole about 700,000 years ago, and there were reasons to believe older reversals existed. Also at this time, Robert S. Dietz and Harry H. Hess were formulating the theory of seafloor spreading—the hypothesis that oceanic crust is created at the crests of the oceanic ridges and consumed in the deep-sea trenches.

It remained for Frederick J. Vine and Drummond H. Matthews of Great Britain and Lawrence W. Morley of Canada to put these observations together in a theory that explained marine magnetic anomalies. The theory rests on three assumptions: (1) that the Earth's magnetic field periodically reverses polarity; (2) that seafloor spreading occurs; and (3) that the oceanic crust is permanently magnetized as it forms and cools at spreading centres. The theory expresses the assumptions—namely, that the oceanic crust records reversals of the Earth's field as it is formed during seafloor spreading. Positive anomalies result when the crust is magnetized in a "normal" polarity parallel to the ambient field of the Earth, and negative anomalies result when the crust is "reversely" magnetized in an opposite sense. As the magnetized crust moves down the flanks of a ridge away from the spreading centre, it remains permanently magnetized and "carries" the magnetic anomalies along with it. (For further details about paleomagnetism and seafloor spreading, see PLATE TECTONICS: *Historical overview: Renewed interest in continental drift.*)

A brilliant leap in understanding was now possible. If the age of the field reversals were known, the age of the ocean crust could be predicted by mapping the corresponding anomaly. By the mid-1960s, Cox and his colleagues had put together a schedule of reversals for the last four or five million years by studying the ages and magnetic polarities of lava flows found on land. Vine and the Canadian geologist J. Tuzo Wilson applied the time scale to marine magnetic anomalies mapped over the Juan de Fuca Ridge, a spreading centre off the northwest United States (Figure 30). They thus dated the crust there and also computed the first seafloor spreading rate of about 30 millimetres per year. The rate is computed by dividing the distance of

Calculation of seafloor spreading rate

an anomaly from the ridge crest by the age of the anomaly twice. Thus the oceanic crust at the Juan de Fuca Ridge is moving at about 15 millimetres per year away from the ridge crest and at about 60 millimetres per year away from the crustal segment on the opposite side of the crest.

During the 1960s and '70s marine magnetic anomalies were mapped over wide areas of the ocean basins. By using estimates of the ages of oceanic crust obtained from core samples by deep-sea drilling, a magnetic anomaly time scale was constructed, and at the same time the spreading history for the ocean basins covering the last 200 million years or so was proposed.

It is thought that the most important contributor to marine magnetic anomalies is the layer of lavas in the upper oceanic crust. A secondary contribution originates in the upper layer of gabbros. The dike layer is essentially demagnetized by the action of hydrothermal waters at the spreading centres. The dominant mechanism of permanent magnetization is the thermoremanent magnetization (or TRM) of iron-titanium oxide minerals. These minerals lock in a TRM as they cool below 200° to 300° C in the presence of the Earth's magnetic field. Although several processes are capable of altering the TRM, including reheating and oxidation at the seafloor, it is remarkably robust, as is evidenced by magnetic anomalies as old as 165 million years in the far western equatorial Pacific.

OCEANIC RIDGES

The largest features of the ocean basin are the oceanic ridges. In the past these features were referred to as mid-ocean ridges, but, as will be seen, the largest oceanic ridge, the East Pacific Rise, is far from a mid-ocean location, and the nomenclature is thus inaccurate. Oceanic ridges are not to be confused with aseismic ridges, which have an entirely different origin (see below).

Principal characteristics. Oceanic ridges are linear mountain chains comprising the largest features on Earth. They are found in every ocean basin and appear to girdle the Earth. The ridges rise from depths near 5 kilometres to an essentially uniform depth of about 2.6 kilometres and are roughly symmetrical in cross section. They can be thousands of kilometres wide. In places, the crests of the ridges are offset across transform faults, or fracture zones, which can be followed down the flanks of the ridges. (Transform faults are those along which lateral movement occurs.) The flanks are marked by sets of mountains and hills that are elongate and parallel to the ridge trend.

New oceanic crust (and part of the upper mantle, which, together with the crust, makes up the lithosphere) is formed at seafloor spreading centres at the crests of the oceanic ridges. Because of this, certain unique geologic features are found there. Fresh basaltic lavas are exposed on the seafloor at the ridge crests. These lavas are progressively buried by sediments as the seafloor spreads away from the site. The flow of heat out of the crust is many times greater at the crests than elsewhere in the world. Earthquakes are common along the crests and in the transform faults that join the offset ridge segments. Analysis of earthquakes occurring at the ridge crests indicates that the oceanic crust is under tension there. A high-amplitude magnetic anomaly is centred over the crests because fresh lavas at the crests are being magnetized in the direction of the present geomagnetic field.

The depths over the oceanic ridges are rather precisely correlated with the age of the ocean crust; specifically, it has been demonstrated that the ocean depth is proportional to the square root of crustal age (Figure 31). The theory explaining this relationship holds that the increase in depth with age is due to the thermal contraction of the oceanic crust and upper mantle as they are carried away from the seafloor spreading centre in an oceanic plate. Because such a plate is ultimately about 100 kilometres thick, contraction of only a few percent predicts the entire relief of an oceanic ridge. It then follows that the width of a ridge can be defined as twice the distance from the crest to the point where the plate has cooled to a steady thermal state. Most of the cooling takes place within 70 or 80 million years, by which time the ocean depth is about 5 to 5.5 kilometres. Because this cooling is a function of age, slow-spreading ridges, such as the Mid-Atlantic Ridge, are narrower than faster-spreading ridges, like the East Pacific Rise (see below). Further, a correlation has been found between global spreading rates and the transgression and regression of ocean waters onto the continents. During the Early Cretaceous period about 100 million years ago, when global spreading rates were uniformly high, oceanic ridges occupied comparatively more of the ocean basins, causing the ocean waters to transgress (spill over) onto the continents, leaving marine sediments in areas now well away from coastlines.

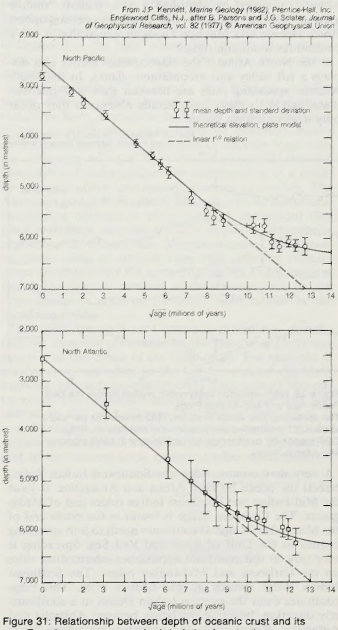


Figure 31: Relationship between depth of oceanic crust and its age. Error bars are one standard deviation (see text).

Global seafloor spreading rates

Besides ridge width, other features appear to be a function of spreading rate. Global spreading rates range from 10 millimetres per year (mm/yr total rate) or less up to 160 mm/yr. Oceanic ridges can be classified as slow (up to 50 mm/yr), intermediate (up to 90 mm/yr), and fast (up to 160 mm/yr). Slow-spreading ridges are characterized by a rift valley at the crest. Such a valley is fault-controlled. It is typically 1.4 kilometres deep and 20 to 40 kilometres wide. Faster-spreading ridges lack rift valleys. At intermediate rates, the crest regions are broad highs with occasional fault-bounded valleys no deeper than 200 metres. At fast rates, an axial high is present at the crest. The slow-spreading rifted ridges have rough faulted topography on their flanks, while the faster-spreading ridges have much smoother flanks (Figure 32).

Distribution of major ridges and spreading centres. Oceanic spreading centres are found in all the ocean basins. In the Arctic Ocean a slow-rate spreading centre is located near the eastern side in the Eurasian basin. It can be followed south, offset by transform faults, to Iceland. Iceland has been created by a hot spot (see below) located directly below an oceanic spreading centre. The

ridge leading south from Iceland is named the Reykjanes Ridge, and, although it spreads at 20 mm/yr or less, it lacks a rift valley. This is thought to be the result of the influence of the hot spot.

The Mid-Atlantic Ridge extends from south of Iceland to the extreme South Atlantic Ocean near 60° S latitude. It bisects the Atlantic Ocean basin, which led to the earlier designation of mid-ocean ridge for features of this type. The Mid-Atlantic Ridge became known in a rudimentary fashion during the 19th century. In 1855 Matthew Fontaine Maury of the U.S. Navy prepared a chart of the Atlantic in which he identified it as a shallow "middle ground." During the 1950s the American oceanographers Bruce Heezen and Maurice Ewing proposed that it was a continuous mountain range.

In the North Atlantic the ridge spreads slowly and displays a rift valley and mountainous flanks. In the South Atlantic spreading rates are between slow and intermediate, and rift valleys are generally absent, as they occur only near transform faults.

After K.C. Macdonald, *The Geology of North America*, vol. N (1989); Geological Society of America

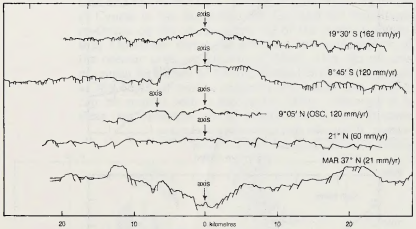


Figure 32: *High-resolution bathymetric profiles across the East Pacific Rise and Mid-Atlantic Ridge.* The spreading rate varies from fast (159 millimetres per year) to slow (21 millimetres per year) on the Mid-Atlantic Ridge. OSC stands for overlapping spreading centre: MAR signifies Mid-Atlantic Ridge.

A very slow oceanic ridge, the Southwest Indian Ridge, bisects the ocean between Africa and Antarctica. It joins the Mid-Indian and Southeast Indian ridges east of Madagascar. The Carlsberg Ridge is found at the north end of the Mid-Indian Ridge. It continues north to join spreading centres in the Gulf of Aden and Red Sea. Spreading is very slow at this point but approaches intermediate rates on the Carlsberg and Mid-Indian ridges. The Southeast Indian Ridge spreads at intermediate rates. This ridge continues from the western Indian Ocean in a southeasterly direction, bisecting the ocean between Australia and Antarctica. Rifted crests and rugged mountainous flanks are characteristic of the Southwest Indian Ridge. The Mid-Indian Ridge has fewer features of this kind, and the Southeast Indian Ridge has generally smoother topography. The latter also displays distinct asymmetric seafloor spreading south of Australia. Analysis of magnetic anomalies shows that rates on opposite sides of the spreading centre have been unequal at many times over the past 50 or 60 million years.

The Pacific-Antarctic Ridge can be followed from a point midway between New Zealand and Antarctica northeast to where it joins the East Pacific Rise off the margin of South America. The former spreads at intermediate to fast rates.

East Pacific Rise

The East Pacific Rise extends from this site northward to the Gulf of California, where it joins the transform zone of the Pacific-North American plate boundary. Offshore from Chile and Peru, the East Pacific Rise is currently spreading at fast rates of 159 mm/yr or more. Rates decrease to about 60 mm/yr at the mouth of the Gulf of California. The crest of the ridge displays a low topographic rise along its length rather than a rift valley. The East Pacific Rise was first detected during the Challenger Expedition of the 1870s. It was described in its gross form during the 1950s and '60s by oceanographers, including Heezen, Ewing, and Henry W. Menard. During the 1980s, Kenneth C. Macdonald, Paul J. Fox, and Peter F. Lonsdale discovered that the main spreading centre appears to be interrupted and offset a few kilometres to one side at various places along the crest of the East Pacific Rise. However, the ends of the offset spreading centres overlap each other by several kilometres. These were identified as a new type of geologic feature of oceanic spreading centres and designated overlapping spreading centres. Such centres are thought to result from interruptions of the magma supply to the crest along its length and define a fundamental segmentation of the ridge on a scale of tens to hundreds of kilometres.

Many smaller spreading centres branch off the major ones or are found behind island arcs. In the western Pacific, spreading centres occur on the Fiji Plateau between the New Hebrides and Fiji Islands and in the Woodlark Basin between New Guinea and the Solomon Islands. A series of spreading centres and transform faults lie between the East Pacific Rise and South America near 40° to 50° S latitude. The Scotia Sea between South America and the Antarctic Peninsula contains a spreading centre. The Galápagos spreading centre trends east-west between the East Pacific Rise and South America near the equator. Three short spreading centres are found a few hundred kilometres off the shore of the Pacific Northwest. These are the Gorda Ridges off northern California, the Juan de Fuca Ridge off Oregon and Washington, and the Explorer Ridge off Vancouver Island. In a careful study of the seafloor spreading history of the Galápagos and the Juan de Fuca spreading centres, the American geophysicist Richard N. Hey developed the idea of the propagating rift. In this phenomenon, one branch of a spreading centre ending in a transform fault lengthens at the expense of the spreading centre across the fault. The rift and fault propagate at one to five times the spreading rate and create chevron patterns in magnetic anomalies and the grain of the seafloor topography resembling the wake of a boat.

Spreading centre zones and associated phenomena. From the 1970s highly detailed studies of spreading centres using deeply towed instruments, photography, and manned submersibles have resulted in new revelations about the processes of seafloor spreading. The most profound discoveries have been of deep-sea hydrothermal vents (see below) and previously unknown biological communities.

Spreading centres are divided into several geologic zones. The neovolcanic zone is at the very axis. It is 1 to 2 kilometres wide and is the site of recent and active volcanism and of the hydrothermal vents. It is marked by chains of small volcanoes or volcanic ridges. Adjacent to the neovolcanic zone is one marked by fissures in the seafloor. This may be 1 to 2 kilometres wide. Beyond this point occurs a zone of active faulting. Here, fissures develop into normal faults with vertical offsets. This zone may be 10 or more kilometres wide. At slow spreading rates the faults have offsets of hundreds of metres, creating rift valleys and rift mountains. At faster rates the vertical offsets are 50 metres or less. A deep rift valley is not formed because the vertical uplifts are cancelled out by faults that downdrop uplifted blocks. This results in linear, fault-bounded abyssal hills and valleys trending parallel to the spreading centre.

Warm springs emanating from the seafloor in the neovolcanic zone were first found on the Galápagos spreading centre. These waters were measured to have temperatures about 20° C above the ambient temperature. In 1979 hydrothermal vents with temperatures near 350° C were discovered on the East Pacific Rise off Mexico (Figure 33). Since then, similar vents have been found on the spreading centres off the Pacific Northwest coast of the United States, on the south end of the northern Mid-Atlantic Ridge, and at many locations on the East Pacific Rise.

Seafloor vents

Hydrothermal vents are localized discharges of heated seawater. They result from cold seawater percolating down into the hot oceanic crust through the zone of fissures and returning to the seafloor in a pipelike flow at the axis of the neovolcanic zone. The heated waters often carry sulfide minerals of zinc, iron, and copper leached from the

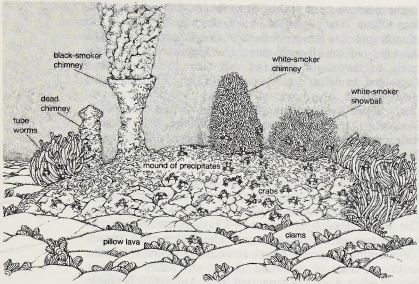


Figure 33: *Hydrothermal field at a spreading centre.* The black-smoker chimney is one or two metres high and is composed of sulfide minerals precipitated from hot water of approximately 350° C. The white-smoker chimney is precipitated from cooler water near 300° C and consists chiefly of silica.

From K.C. Macdonald and B.P. Luyendyk, "The Crest of the East Pacific Rise," *Scientific American* (May 1981); copyright © 1981 Scientific American, Inc., all rights reserved

crust. Outflow of these heated waters probably accounts for 20 percent of the Earth's heat loss. Exotic biological communities exist around the hydrothermal vents. These ecosystems are totally independent of energy from the Sun. They are not dependent on photosynthesis but rather on chemosynthesis by sulfur-fixing bacteria. The sulfide minerals precipitated in the neovolcanic zone can accumulate in substantial amounts and are sometimes buried by lava flows at a later time. Such deposits are mined as commercial ores in ophiolites on Cyprus and in Oman.

Magma chambers have been detected beneath the crest of the East Pacific Rise by seismic experiments. (The principle underlying the experiments is that partially molten or molten rock slows the travel of seismic waves and also strongly reflects them.) The depth to the top of the chambers is about two kilometres below the seafloor. The width is more difficult to ascertain, but is probably one to four kilometres. Their thickness seems to be about two to six kilometres based on studies of ophiolites (Figure 34). The chambers have been mapped along the trend of the crest between 9° and 13° N latitude. The top is relatively continuous, but is apparently interrupted by offsets of transform faults and overlapping spreading centres.

FRACTURE ZONES AND TRANSFORM FAULTS

Fracture zones. As was noted above, oceanic ridges (and their associated spreading centres) are offset along their trend by fracture zones. These are ridges and valleys on the order of tens of kilometres wide that cut across the crests of the ridges at approximately right angles and offset their trend (Figure 35). Typically, a regional depth offset is present across a fracture zone, owing to the juxtaposition of crust of different ages (and, therefore, depth) across it. In the Atlantic, on the slow spreading Mid-Atlantic Ridge, fracture zones are numerous and occur every 55 kilometres on average along the trend of the ridge. They offset the crest between 5 and 40 kilometres. Some of the larger fracture zones in the North Atlantic are the Gibbs at 52° N, the Atlantis at 30° N, and the Vema at 11° N. These and others can be followed across both flanks of the ridge for some 3,000 kilometres. The Vema Fracture Zone offsets the Mid-Atlantic Ridge 320 kilometres to the left. It is marked by a sediment-filled valley more than 5 kilometres deep and 10 to 20 kilometres wide and is flanked by mountains 3,500 metres high. Basalts, gabbros, and serpentinized peridotites (*i.e.*, those peridotites that have been altered in varying degrees to serpentine) of the oceanic crust and mantle have been recovered from the mountain flanks.

Fracture zones occur less frequently on the East Pacific Rise, but they offset the ridge by a greater amount. More than a dozen can be found between 20° N and 30° S. Typical offsets are roughly 100 kilometres. Several fracture zones more than 3,000 kilometres long are found off the shore of western North America. These include the Mendocino, Murray, Molokai, and Clarion fracture zones. They are not associated with a ridge crest. Rather, they occur on the west flank of the defunct Pacific-Farallon oceanic ridge. The Farallon Plate has all but disappeared down a subduction zone that extended along the entire coast of California and Baja California until about 25 to 30 million years ago. Subduction now occurs north of the Mendocino Fracture Zone. These fracture zones off western North America were among the first mapped. Menard has traced them almost 10,000 kilometres westward across the Pacific. The continental margin of northern California is displaced to the right where the Mendocino Fracture Zone and its transform portion, the Gorda Escarpment, intersect it.

Transform faults. The portion of a fracture zone between different offset spreading centres constitutes a transform fault. Transform faults also connect spreading centres to subduction zones (deep-sea trenches). Faults of this kind are the only segments of fracture zones that are seismically active. J. Tuzo Wilson recognized this and other features and explained the phenomenon as a transfer of motion from one spreading centre to another. The American geologist W. Jason Morgan, one of the several outstanding pioneers in plate tectonics, recognized that transform faults are zones where opposing lithospheric plates slip past one another. Morgan proposed that opposing plates along an oceanic ridge crest offset by fracture zones are divided by the spreading centres and transform faults. The inactive portions of the fracture zone on the ridge flanks are scars on the ocean floor created in the transform faults.

This theory made a very dramatic prediction: namely, that the direction of motion on the transform faults was opposite to the offsets of the ridge crests. For example, if a ridge crest was offset to the left by a transform fault, implying leftward movement on a fault joining the offset crests, the movement across the transform fault was instead to the right (Figure 35). This is clear when it is realized that the plate boundaries are confined to the spreading centres and transform faults, not to the inactive part of the fracture zone. Seismic studies of earthquakes

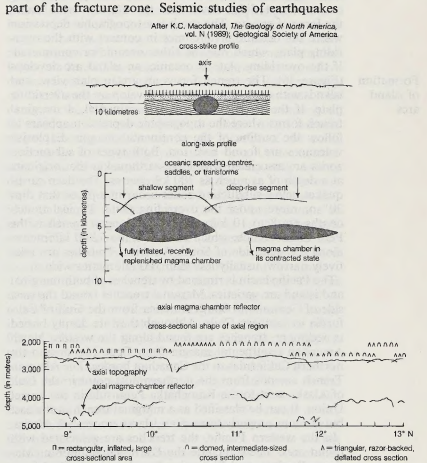


Figure 34: Cross-trend and along-trend views of an axial magma chamber on the East Pacific Rise. The two upper diagrams are idealized; the one at the bottom is made from seismic reflection surveys (see text).

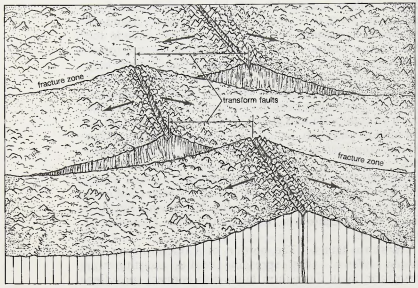


Figure 35: Oceanic ridges offset by transform faults and fracture zones. The arrows show the direction of movement across the transform faults.

From W.H. Menard, *Scientific American* (September 1969); copyright © 1969 Scientific American, Inc., all rights reserved

from transform faults soon revealed that the motion was opposite, as predicted.

Not everywhere in the ocean basins are plate motions exactly parallel to transform faults. In places where a component of opening motion occurs across the transform, volcanic activity results, and the fracture zone is termed a leaky transform fault. South of New Zealand, between it and the Pacific-Antarctic Ridge, a component of shortening is occurring across a transform called the Macquarie Ridge. Here, subduction may be taking place at a slow rate.

DEEP-SEA TRENCHES

Types. Although the term trench has been applied to many deep, long linear troughs in the ocean floor, the most common and accurate usage relates it to subduction zones. According to plate tectonic theory, subduction zones are locations where a lithospheric plate bearing oceanic crust slides down into the upper mantle under the force of gravity. The result is a topographic depression where the oceanic plate comes in contact with the overriding plate, which may be either oceanic or continental. If the overriding plate is oceanic, an island arc develops (Figure 36). The trench forms an arc in plan view, and islands with explosive volcanoes develop on the overriding plate. If the overriding plate is continental, a marginal trench forms where the topographic depression appears to follow the outline of the continental margin. Explosive volcanoes are found here too. Both types of subduction zones are associated with large earthquakes that originate at a depth of as much as 700 kilometres. The deep earthquakes below subduction zones occur in a plane that dips 30° or more under the overriding plate. Typical trench depths are 8 to 10 kilometres. The longest trench is the Peru-Chile Trench, which extends some 5,900 kilometres along the west side of South America. Trenches are relatively narrow, usually less than 100 kilometres wide.

Formation of island arcs

The Pacific basin is rimmed by trenches of both marginal and island arc varieties. Marginal trenches bound the west side of Central and South America from the Gulf of California to southern Chile. Although they are deeply buried in sediment, trenches are found along the western North American continental margin from Cape Mendocino (in northern California) to the Canadian border. The Aleutian Trench extends from the northernmost point in the Gulf of Alaska west to the Kamchatka Peninsula in the Soviet Union. It can be classified as a marginal trench in the east but is more properly termed an island arc west of Alaska.

In the western Pacific, the trenches are associated with island arcs. These include the Kuril, Japan, Bonin, Mariana, Ryukyu, and Philippine trenches that extend from Kamchatka to near the equator. A complex pattern of island arcs is found in Indonesia. The major island arc here is the Java Trench extending from northern Australia to the northwestern end of Sumatra in the northeast Indian Ocean. The region of New Guinea and the Solomon Islands includes the New Britain and Solomon trenches, the latter of which joins the New Hebrides Trench directly to the south. East of this area the Tonga and Kermadec trenches extend south from the Fiji Islands to New Zealand.

Two island arcs occur in the Atlantic Ocean. The South Sandwich Trench is located west of the Mid-Atlantic Ridge between South America and Antarctica. The Puerto Rico Trench joins the Lesser Antilles Island arc in the eastern Caribbean.

Some seafloor features bear the name trench and are deep linear troughs but are not subduction zones. The Vema Trench on the Mid-Indian Ridge is a fracture zone. The Vityaz Trench northwest of Fiji is an aseismic (inactive) feature of unknown origin. The Diamantina trench (Diamantina Fracture Zone) extends westward from the southwest coast of Australia. It is a rift valley that was formed when Australia separated from Antarctica between 60 and 50 million years ago.

The deepest water on Earth (11,034 metres) is located in the southern end of the Mariana Trench near Guam. A few trenches are partially filled with sediments derived from the bordering continents. The Aleutian Trench is effectively buried east of Kodiak Island in the Gulf of Alaska. Here, the ocean floor is smooth and flat. To the west farther from the sediment supply on Alaska, the trench reaches depths of more than seven kilometres. The Lesser Antilles trench in the eastern Caribbean also is buried by sediments originating from South America.

Structure. Oceanward of trenches the seafloor is usually bulged upward in an outer ridge or rise of up to 1,000 metres relief. This condition is thought to be the elastic response of the oceanic plate bending down into a subduction zone. The landward or island-arc slope of the trench is often interrupted by a submarine ridge, which sometimes breaks the ocean surface, as in the case of the Java Trench. Such a ridge is constructed from deformed sediments scraped off the top of the descending oceanic plate and is termed an accretionary prism. A line of explosive volcanoes, extruding (erupting) a lava that forms the volcanic rock andesite, is found on the overriding plate usually 100 kilometres or so from the trench. In marginal trenches these volcanoes form mountain chains, such as the Cascades in the Pacific Northwest or the great volcanoes of the Andes. In island arcs they form active volcanic island chains, such as the Mariana Islands.

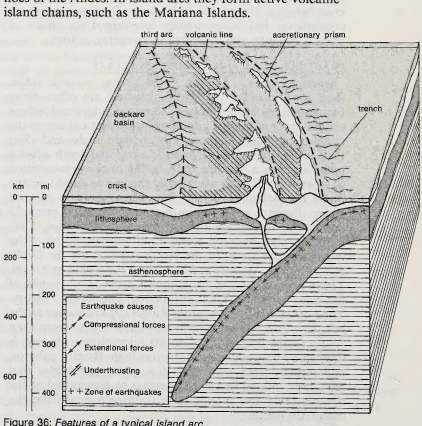


Figure 36: *Features of a typical island arc.* The deep-sea trench, volcanic line, accretionary prism, backarc basin, and third arc generally associated with island arcs are shown here.

Behind the volcanic line of island arcs are sometimes found young, narrow ocean basins. These basins are bounded on the opposite side by submarine ridges. Such interarc, or backarc, basins are sites of seafloor spreading directly caused by the dynamics of subduction. They originate at the volcanic line, so that the outer bounding submarine ridge, or third arc, represents an older portion of the volcanic line that has spread away. These backarc basins bear many of the features characteristic of oceanic spreading centres. Well-studied examples of these features are found in the Lau Basin of the Tonga arc and also west of the Mariana Islands. The Sea of Japan originated from backarc spreading behind the Japanese arc that began some 30 million years ago. At least two backarc basins have opened behind the Mariana arc, creating seafloor in two phases from about 30 to 17 million years ago in the western Parece Vela Basin and from 5 million years ago in the Mariana Trough next to the islands.

ASEISMIC RIDGES

In some oceans the basin floors are crossed by long, linear and mountainous aseismic ridges. The term aseismic distinguishes these ridges from oceanic spreading centres because the former lack earthquakes. Most aseismic ridges are constructed by volcanism from a hot spot and are composed of coalescing volcanoes of various sizes. A hot spot is a magma-generating centre fixed in the Earth's deep mantle and leaves a trail of volcanic outpourings on the seafloor as an oceanic plate travels over it. This form of volcanism is not associated with the volcanism at spreading centres and is distinct from it chemically in that the magma extruded onto the surface has a higher alkali composition. (For additional information on hot spots, see VOLCANISM: *Volcanism and tectonic activity: Intraplate volcanism.*)

Formation by hot spot volcanism

The Hawaiian-Emperor chain is the best displayed aseismic ridge. Earthquakes do occur here, but only at the end of the ridge where volcanism is current—in this case, on the island of Hawaii (commonly known as the Big Island) to the southeast end of the island chain. Taking into account the relief of the island of Hawaii above the seafloor, it is the largest volcanic edifice on Earth. The Hawaiian-Emperor chain stretches from the Big Island to the intersection of the Kuril and Aleutian trenches in the northwest Pacific. There are roughly 18 volcanoes or seamounts (see below) per 1,000 kilometres along the Hawaiian segment and 13 per 1,000 kilometres on the Emperor portion beyond the bend. The Hawaiian Islands are a part of the chain—the young part—that rises above sea level. The Hawaiian-Emperor chain has two main trends: (1) from the Hawaiian Islands west to the Kammu and Yūryaku seamounts (near 32° N, 168° W), the trend of the Hawaiian portion is just west of northwest; and (2) from this point to the Aleutian Trench, the trend of the Emperor segment is north-northwest. The hot spot interpretation infers that this change in trend is due to a change in the direction of Pacific Plate motion, from north-northwest prior to 38 million years ago (the age of the ridge at the change in trend) to west of northwest until the present day. Radiometric dating of rocks from the ridge indicates that it is 70 million years old at its extreme north end.

Other prominent aseismic ridges include the Ninetyeast Ridge and the Chagos-Laccadive Plateau in the Indian Ocean and the Walvis Ridge and Rio Grande Rise in the South Atlantic. The Ninetyeast Ridge is thought to have originated from hot spot volcanic activity now located at the Kerguelen Islands near Antarctica. These islands lie atop the Kerguelen Plateau, which also originated from volcanism at this hot spot. The Ninetyeast Ridge stretches parallel to 90° E longitude in a long, linear chain of seamounts and volcanic ridges from the Andaman Islands in the Bay of Bengal more than 4,500 kilometres to the south where it intersects Broken Ridge at 30° S latitude. Broken Ridge is an aseismic ridge and was once part of the Kerguelen Plateau. It was split away from the plateau as Australia separated from Antarctica.

Core samples of the seafloor along the Ninetyeast Ridge have been retrieved through deep-sea drilling. Analyses of the samples show that the ridge is slightly less than 30 million years old in the south and about 80 million years old in the north. Additionally, sediments on the ridge indicate that parts of it were above sea level while it was being built near a spreading centre. The ridge then subsided as it rode north on the Indian Plate.

The Walvis Ridge and Rio Grande Rise originated from hot spot volcanism now occurring at the islands of Tristan da Cunha 300 kilometres east of the crest of the Mid-Atlantic Ridge. The Walvis Ridge trends northeast from this location to the African margin. The Rio Grande Rise trends roughly southeast from the South American margin toward the Mid-Atlantic Ridge. Both the Walvis Ridge and Rio Grande Rise began forming from the same hot spot near the spreading centre as the South Atlantic was in its initial opening stages 100 to 80 million years ago. The spreading centre shifted west of the hot spot about 80 million years ago, ending construction of the Rio Grande Rise but continuing to build the Walvis Ridge. Volcanic activity has since diminished, resulting in the younger part of the latter ridge being smaller. The findings of ocean drilling on the Rio Grande Rise show that it was once a volcanic island some two kilometres high.

SEAMOUNTS, GUYOTS, AND ABYSSAL HILLS

Submarine volcanoes

Seamounts are submarine volcanoes with more than 1,000 metres of relief. Aseismic ridges are built by chains of overlapping seamounts. A seamount is akin to a subaerial shield volcano in that it also has gently sloping sides (5° to 15°) and is constructed by nonexplosive eruptions of alkaline basalt lavas that are thought to originate from depths of roughly 150 kilometres. About 2,000 seamounts are known; they are most common in the Pacific and on fast-spreading ridges. Like the Hawaiian-Emperor chain, the lines of seamounts and islands trending northwest-southeast in the central and south Pacific (Marshall Islands, Line Islands, Tuamotu Archipelago, and Cook and Austral Islands) may be due to hot spot volcanism. Isolated seamounts also occur, and many of these are located in the far western Pacific. Another group of smaller seamounts is found in the northeastern Pacific.

Flat-topped seamounts are called guyots. They are particularly abundant in the western Pacific and along the Emperor seamount chain. Bottom samples and drill cores of shallow-water sediments and fossils capping guyots have been retrieved. The presence of such geologic materials suggest that guyots are seamounts that have had their peaks planed off by wave action and have since subsided below sea level. The western Pacific guyots are capped by drowned coral atolls and reefs. These reefs are generally of Late Cretaceous age (about 95 million years old). The cause of the subsidence is attributed to the sinking of the seafloor as it moves down the flanks of an oceanic ridge. However, the reason for the demise of the coral reefs on the Cretaceous guyots is less clear. Under normal conditions, coral growth can easily keep up with sinking due to seafloor spreading. The Cretaceous guyots may have resulted from the northward drift of seamounts and reefs on the Pacific Plate away from the tropical zone of favourable growth. Another hypothesis is that the reefs were killed by unusually anoxic (oxygen-depleted) conditions that developed suddenly, a situation possibly related to intense seafloor volcanism in the Pacific at this time.

Abyssal hills are low-relief (less than 1 kilometre) features usually 1 to 10 kilometres wide and elongate parallel to spreading centres or to marine magnetic anomalies located in the vicinity of the latter. The tops of the hills are often flat, in which case they have steep sides. Gently sloping sides, however, are equally common. Abyssal hills are extremely numerous, so much so that Menard declared them "the most widespread physiographic forms of the face of the earth." Abyssal hills are most common in the Pacific basin, where they cover 80 to 85 percent of the seafloor. Because they cover the entire flanks and crests of the oceanic ridges, such hills are thought to form during crustal accretion at spreading centres. They are commonly associated with intermediate- and fast-spreading ridges. On slow-spreading ridges, such as the Mid-Atlantic, the topographic features are much larger and have steeper sides. Bottom-sampling and seismic reflection studies re-

veal that abyssal hills are relief features on the top of the oceanic crust; they are not constructed from ocean-bottom sediments. In areas such as the abyssal plains (see below), abyssal hills are buried by sediments.

Apparently the hills are constructed by two processes: volcanism and block faulting. The relative contribution of each may depend on the spreading rate. At slower rates, faulting of the oceanic crust is a dominant factor in forming the relief, and the relief of the hills is greater as the rate is slower. At the crest of a spreading centre, volcanism in the neovolcanic zone initiates the construction of volcanic hills. The zone of active faulting is where they form or are modified by block faulting. The existence of discrete and separate volcanic hills indicates that volcanism at a spreading centre is episodic.

DEEP-SEA SEDIMENTS

The ocean basin floor is everywhere covered by sediments of different types and origins. The only exception are the crests of the spreading centres where new ocean floor has not existed long enough to accumulate a sediment cover. Sediment thickness in the oceans averages about 450 metres. The sediment cover in the Pacific basin ranges from 300 to 600 metres thick, and that in the Atlantic is about 1,000 metres. Generally, the thickness of sediment on the oceanic crust increases with the age of the crust. Oceanic crust adjacent to the continents can be deeply buried by several kilometres of sediment. Deep-sea sediments can reveal much about the last 200 million years of Earth history, including seafloor spreading, the history of ocean life, the behaviour of the Earth's magnetic field, and the changes in the ocean currents and climate.

The study of ocean sediments has been accomplished by several means. Bottom samplers, such as dredges and cores up to 30 metres long, have been lowered from ships by wire to retrieve samples of the upper sediment layers. Deep-sea drilling has retrieved core samples from the entire sediment layer in several hundred locations in the ocean basins. The seismic reflection method has been used to map the thickness of sediments in many parts of the oceans. Besides thickness, seismic reflection data can often reveal sediment type and the processes of sedimentation. (For more information on the equipment and techniques used by investigators to study deep-sea sediments, see EXPLORATION: *Undersea exploration.*)

Sediment types. Deep-sea sediments can be classified as terrigenous, originating from land; biogenic, consisting largely of the skeletal debris of microorganisms; and authigenic, formed in place on the seafloor (Figure 37). Pelagic sediments, either terrigenous or biogenic, are those that are deposited very slowly in the open ocean either by settling through the volume of oceanic water or by precipitation. The sinking rates of pelagic sediment grains are extremely slow because they ordinarily are no larger than several micrometres. However, fine particles are normally bundled into fecal pellets by zooplankton, which allows sinking at a rate of 40 to 400 metres per day.

Terrigenous sediments are transported to the oceans by rivers and wind. The sediments that reach the continental shelf are often stored in submarine canyons on the continental slope. Turbidity currents carry these sediments down into the deep sea (see above *Density currents in the oceans: Turbidity currents*). These currents create sedimentary deposits called turbidites, which are layers up to several metres thick composed of sediment particles that grade upward from coarser to finer sizes. The turbidites build sedimentary deep-sea fans adjacent to the base of the continental slope. Turbidites also are found below the major river deltas of the world where they build features called abyssal cones. The largest of these is the Ganges Fan (also called the Ganges Cone or Bengal Cone) in the Bay of Bengal east of the Indian subcontinent. It measures 3,000 kilometres long (north-south) by 1,000 kilometres wide (east-west) and is up to 12 kilometres thick. The Bengal Cone is forming from rock material eroded from the Himalayas and transported by the Ganges and Brahmaputra rivers.

Abyssal plains

Abyssal plains are formed by the accumulation of turbidites beyond the limits of deep-sea fans and abyssal cones in locations where there is a very large sediment supply. In contrast to fans and cones, abyssal plains are flat and featureless. They are prominent near both margins of the Atlantic and in the northeast Pacific. Tectonic and climate controls have influenced the formation of abyssal plains. The last major glaciation near the end of the Pleistocene epoch about 10,000 years ago greatly increased erosion and sediment supply to the deep sea, but deep-sea trenches interrupted the flow of turbidity currents to the ocean floor. Off the Pacific Northwest coast of the United States, however, the trenches were filled by turbidites, and subsequent turbidity currents passed beyond them to form the Alaska and Tufts abyssal plains.

Brown clays are a variety of pelagic sediment, mostly of terrigenous origin, which are composed largely of four different clay minerals: chlorite, illite, kaolinite, and montmorillonite. By definition, clays have less than 30 percent biogenic components. Quartz, volcanic ash, and micrometeorites are common as minor constituents. Brown clays are widespread in the deeper areas of the oceans below four kilometres. They dominate the floor of the

From J.P. Kennett, *Marine Geology* (1982; Prentice-Hall, Englewood Cliffs, N.J., after W.H. Berger in *The Geology of Continental Margins*, C.A. Burk and C.L. Drake (eds.), (1974); Springer-Verlag, New York

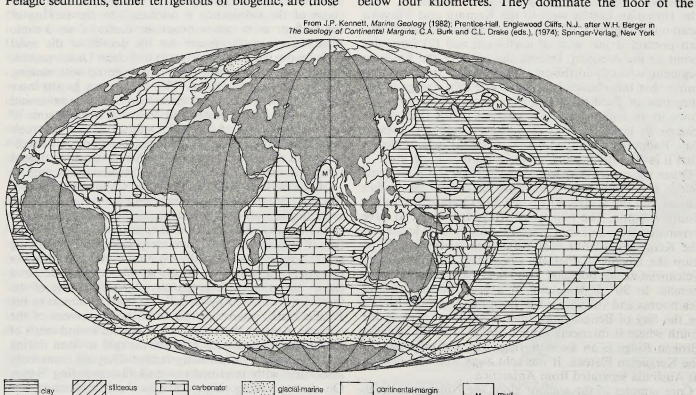


Figure 37: Global map of the distribution of sediment types on the ocean floor.

central North Pacific. Clays accumulate very slowly, averaging about one millimetre per 1,000 years. The type of clay found in a given area is a function of the source region on land and the climate. For example, chlorite is dominant in polar regions and kaolinite in the tropics. Clays are introduced into the oceans by river transport, although kaolinite is also carried by the wind from the arid regions of Africa and Australia. Montmorillonite is an alteration product of volcanic material and can form from either wind-blown volcanic ash or basaltic glass on the seafloor.

Sediments composed mostly or entirely of volcanic ash are commonly found adjacent to the island arcs and marginal trenches. These are normally deposited as turbidites. Volcanic ash that has been ejected higher than five kilometres during an eruption can be carried by wind and settle out through the atmosphere and oceans as pelagic sediment. The ocean floor encircling Antarctica is covered by glacial marine sediments. These sediments are carried by icebergs from the continent as far north as the Antarctic Convergence at 45° to 55° latitude.

Biogenic oozes are pelagic sediments that have more than 30 percent skeletal material. They can be either carbonate (or calcareous) ooze or siliceous ooze. The skeletal material in carbonate oozes is calcium carbonate usually in the form of the mineral calcite but sometimes aragonite. The most common contributors to the skeletal debris are such microorganisms as foraminiferans and coccoliths, microscopic carbonate plates that coat certain species of marine algae and protozoa. Siliceous oozes are composed of opal (amorphous, hydrated silica) that forms the skeleton of various microorganisms, including diatoms, radiolarians, siliceous sponges, and silicoflagellates. The distribution of biogenic oozes depends mainly on the supply of skeletal material, dissolution of the skeletons, and dilution by other sediment types, such as turbidites or clays.

Primary productivity in the ocean surface waters controls supply to a large extent. Productivity is high at the equator and in zones of coastal upwelling and also where oceanic divergences occur near Antarctica. Productivity is lowest in the central areas of the oceans (the gyres) in both hemispheres. Siliceous oozes are more reliable indicators of high productivity than carbonate oozes. This is because silica dissolves quickly in surface waters and carbonate dissolves in deep water; hence, high surface productivity is required to supply siliceous skeletons to the ocean floor. Carbonate oozes dominate the deep Atlantic seafloor, while siliceous oozes are most common in the Pacific; the floor of the Indian Ocean is covered by a combination of the two (see Figure 37).

Carbonate oozes cover about half of the world's seafloor. They are present chiefly above a depth of 4,500 metres; below that they dissolve quickly. This depth is named the Calcite Compensation Depth (or CCD). It represents the level at which the rate of carbonate accumulation equals the rate of carbonate dissolution. In the Atlantic basin the CCD is 500 metres deeper than in the Pacific basin, reflecting both a high rate of supply and low rate of dissolution in comparison to the Pacific. The input of carbonate to the ocean is through rivers and deep-sea hydrothermal vents. Variation in input, productivity, and dissolution rates in the geologic past have caused the CCD to vary over 2,000 metres. The CCD intersects the flanks of the world's oceanic ridges, and as a result these are mostly blanketed by carbonate oozes.

Calcite Compensation Depth

Siliceous oozes predominate in two places in the oceans: around Antarctica and a few degrees of latitude north and south of the equator. At high latitudes the oozes include mostly the shells of diatoms. South of the Antarctic Convergence diatom oozes dominate the seafloor sediment cover and mix with glacial marine sediments closer to the continent. Seventy-five percent of all the oceans' silica supply is being deposited in the area surrounding Antarctica. Radiolarian oozes are more common near the equator in the Pacific. Here, both siliceous oozes and calcareous oozes occur, but carbonate deposition dominates the region immediately near the equator. Siliceous oozes bracket the carbonate belt and blend with pelagic clays farther north and south. Because siliceous skeletons dissolve so quickly in seawater, only the more robust skeletal remains are found in the siliceous oozes. Thus, fossils of this kind are not completely representative of the organisms living in the waters above.

The most significant authigenic sediments in the ocean basins today are metal-rich sediments and manganese nodules. Metal-rich sediments include those enriched by iron, manganese, copper, chromium, and lead. These sediments are common at spreading centres, indicating that processes at the centres are responsible for their formation—specifically, hydrothermal circulation is the controlling factor. Deep-sea drill cores have revealed the presence of metal-rich sediments on top of ancient oceanic crust away from ridge crests. It can be inferred from this that the processes controlling their formation existed in the past, but with variations. Which type of enriched sediment is deposited depends on the degree of mixing between the hydrothermal water deep in the crust at a spreading centre and the cold seawater percolating down into the crust. Little mixing produces sulfides, liberal mixing yields manganese-rich crustal material, and intermediate conditions give rise to sediments enriched in iron and manganese.

Manganese nodules are pebbles or stones about the size of walnuts that are built of onionlike layers of manganese and iron oxides. Minor constituents include copper, nickel, and cobalt, making the nodules a potential ore of these valuable elements. Mining of manganese nodules has been the subject of study and experimentation since the 1950s (see below *Economic aspects of the oceans: Sources of minerals and other raw materials*). The nodules grow very slowly, about one to four millimetres per million years. They are found in areas of slow sedimentation, usually five millimetres per thousand years or less. The North and South Pacific hold the greatest concentration of manganese nodules; in some places, the nodules cover 90 percent of the surface of the ocean floor. Coverages this high also are found in the southernmost South Atlantic. The Indian Ocean floor is largely devoid of manganese nodules. Because seawater is supersaturated in manganese, the direct precipitation of the element onto an available surface is the most likely mode of nodule formation.

Manganese nodules

Two significant mysteries surround manganese nodules. Drilling and coring in the sediment column has shown that nodules are vastly more abundant at the seafloor than below it and that the rate of growth of nodules is 10 times slower than the lowest known sedimentation rates. If such is the case, the nodules should be quickly buried and should be common in the sediment below the seafloor. Current theories for explaining these observations propose that bottom currents keep areas of nodule growth free of sediment deposition and that burrowing organisms nudge and roll the nodules in the process of feeding, thereby keeping them at the surface of the seafloor. Observations in the deep sea support both explanations.

Sedimentation patterns. The patterns of sedimentation in the ocean basins have not been static over geologic time. The existing basins, no more than 200 million years old, contain a highly variable sedimentary record. The major factor behind the variations is plate movements and related changes in climate and ocean water circulation. Since about 200 million years ago, a single vast ocean basin has given way to five or six smaller ones. The Pacific Ocean basin has shrunk, while the North and South Atlantic basins have been created. The climate has changed from warm and mild to cool, stormy, and glacial. Plate movements have altered the course of surface and deep ocean currents and changed the patterns of upwelling, productivity, and biogenic sedimentation. Seaways have opened and closed. The Strait of Gibraltar, for example, was closed off about 6 million years ago, allowing the entire Mediterranean Sea to evaporate and leave thick salt deposits on its floor. Changes in seafloor spreading rates and glaciations have caused sea level to rise and fall, greatly altering the deep-sea sedimentation pattern of both terrigenous and biogenic sediments. The CCD has fluctuated more than 2,000 metres in response to changes in carbonate supply and the corrosive nature of ocean bottom waters. Bottom currents have changed, becoming erosive or nondepositional in some regions to produce

unconformities and redistributing enormous volumes of sediment to other locations. The Pacific Plate has been steadily moving northward, so that biogenic sediments of the equatorial regions are found in drill cores taken in the barren North Pacific.

EVOLUTION OF THE OCEAN BASINS THROUGH PLATE MOVEMENTS

Through most of geologic time, probably extending back 2 billion years, the ocean basins have both grown and been consumed as plate tectonics continued on Earth. The latest phase of ocean basin growth began just less than 200 million years ago with the breakup of the supercontinent Pangaea, the enormous landmass composed of nearly all the present-day continents. Since that time, the major developments have included a shrinking of the Pacific basin at the expense of the growing Atlantic and Arctic basins, the opening of the Tethys seaway circling the globe in tropical latitudes and its subsequent closing, and the opening of the Southern Ocean (see above *General considerations*) as the southern continents moved north away from Antarctica.

As was noted earlier, the oldest known oceanic crust (estimated to be about 200 million years old) is located in the far western equatorial Pacific, east of the Mariana Island arc. The Pacific ocean floor at this site was generated during seafloor spreading from a pattern of ridges and plates that had existed for some unknown period of time. At least five different seafloor spreading centres were involved (Figure 38). In the Indian Ocean the oldest segment of seafloor was formed about 165 to 145 million years ago by the rifting away of Africa and South America from Gondwana, a supercontinent consisting largely of the present-day continents of the Southern Hemisphere. At this time, Africa was joined to South America, Eurasia, and North America. Today, this old seafloor is found along the east coast of Africa from the Somali Basin to the east coast of South Africa and adjacent to Queen Maud Land and Enderby Land in East Antarctica.

Close to 180 million years ago (but before 165 million years ago), North America and Eurasia, which together made up most of the large northern continent of Laurasia, began drifting away from Africa and South America, creating the first seafloor in the central region of the North Atlantic and opening the Gulf of Mexico. The Tethys seaway also opened during this rifting phase as Europe pulled away from Africa. Shortly after this time continental fragments, including possibly Tibet, Myanmar (Burma), and Malaya, rifted away from the northwest coast of Australia and moved northward, thereby creating the oldest seafloor in the Timor Sea. During this period spreading continued in the Pacific basin with the growth of the Pacific Plate and the consumption by subduction of its bordering plates, including the Izanagi, Farallon, and Phoenix. The Pacific Plate moved northward during this phase and continues to do so today.

Opening of the Caribbean and the Gulf of Mexico

India and Madagascar, as a unit, rifted away from Australia and Antarctica prior to 130 million years ago and began drifting northward, creating seafloor adjacent to Western Australia and East Antarctica (Figure 39). Possibly simultaneously or shortly after this rifting began, South America started to separate from Africa, initiating the formation of seafloor in the South Atlantic Ocean.

Between 90 and 80 million years ago, Madagascar and India separated, and the spreading ridges in the Indian Ocean were reorganized. India began drifting northward directly toward Asia. During this same period Europe, joined to Greenland, began drifting away from North America, which resulted in the emergence of the seafloor in the Labrador Sea and the northernmost Atlantic Ocean. This spreading phase affected the passages in the Tethys seaway between Europe (Iberia) and northwest Africa, intermittently opening and closing it. In the southwest Pacific, New Zealand, along with the Lord Howe Rise and the Norfolk Ridge, rifted away from Australia and Antarctica between 80 and 60 million years ago, opening the Tasman Sea.

About 60 million years ago a new rift and oceanic ridge formed between Greenland and Europe, separating them and initiating the formation of oceanic crust in the Norwegian Sea and the Eurasian basin in the eastern Arctic Ocean (Figure 40). The Amerasian basin in the western Arctic Ocean had formed during an earlier spreading phase from about 130 to 110 million years ago. Between 60 and 50 million years ago, significant events occurred in the Indian Ocean and southwest Pacific. Australia began drifting northward, away from East Antarctica, creating seafloor there. The northward movement of Australia resulted in the emergence of several subduction zones and island arcs in the southwest and equatorial Pacific. The Indian sub-

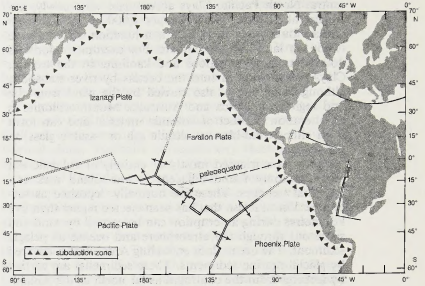


Figure 38: Possible configuration of spreading centres in the Pacific Ocean about 110 million years ago (see text).

From J.P. Kennett, *Marine Geology* (1982); Prentice-Hall, Inc., Englewood Cliffs, N.J., after R.L. Larson and W.C. Pitman III, *Geological Society America Bulletin*, vol. 83 (1972); Geological Society of America

From C.McA. Powell, S.R. Roots, and J.J. Veevers, *Tectonophysics*, vol. 155 (1988); Elsevier Science Publishers P.V.

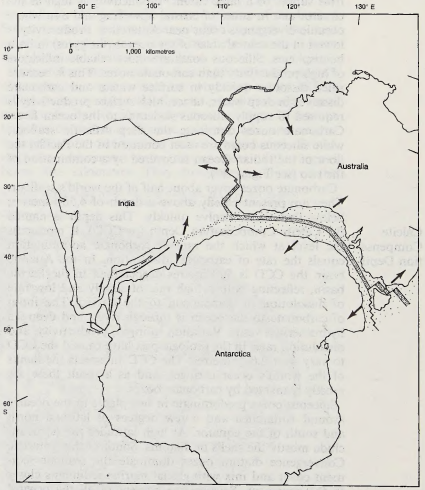


Figure 39: *Early phase of seafloor spreading in the eastern Indian Ocean.*
The rift between Australia and Antarctica developed into seafloor spreading between 60 and 50 million years ago. The arrows show the relative motion between these segments of Gondwana. The heavy stipple between Australia and Antarctica represents the amount of continental extension, and the light stipple indicates the continental extension over the whole area.

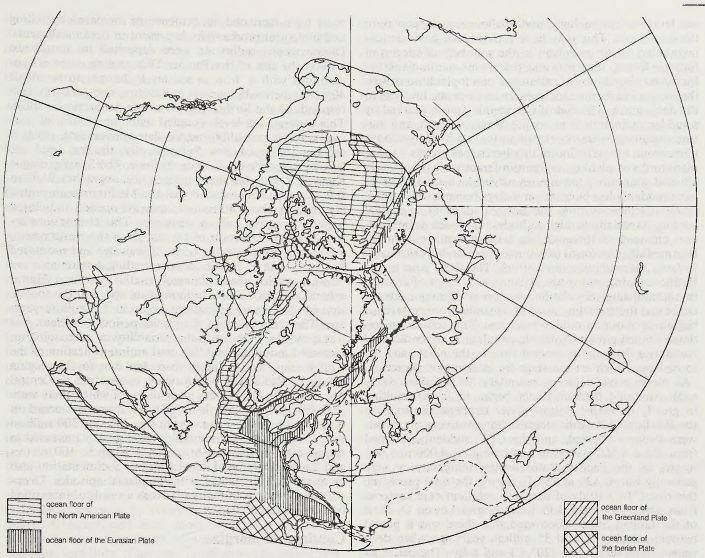


Figure 40: Reconstruction of the North Atlantic and Arctic basins for 56 million years ago. Seafloor spreading began at this time in the Norwegian Sea and Eurasian (eastern Arctic) basin areas.

From D.B. Rowley and A.L. Lottes, *Tectonophysics*, vol. 155 (1988); Elsevier Science Publishers P.V.

continent first touched against the Asian continent about 53 million years ago, developing structures that preceded the main Himalayan orogeny (mountain-building event), which began in earnest some 40 million years ago.

Less than 30 million years ago, seafloor spreading ceased in the Labrador Sea. Along the west coast of North America, the Pacific Plate and the North American Plate converged along what is now California shortly after 30 million years ago. This resulted in the cessation of a long history of subduction in the area and the gradual conversion of this continental margin to a transform fault zone. Continued closure between Africa and Europe, which began about 100 million years ago, caused the isolation of the Mediterranean Sea, so that by 6 million years ago it had completely evaporated.

The present-day Mediterranean seafloor was formed during a complex sequence of rifting between small plates in this region, beginning with the separation of North America and Europe from Africa about 200 million years ago. In the eastern Mediterranean, the seafloor is no older than about 100 million years. West of Italy it was created during subsequent spreading between 30 and 20 million years ago.

The Caribbean Sea and the Gulf of Mexico formed as a result of the relative movement between North America and South America. The seafloor of the Gulf of Mexico began forming some 160 to 150 million years ago. A proto- or ancient Caribbean seafloor also was formed during this period but was later subducted. The present Caribbean seafloor consists of a captured piece of the Farallon Plate (from the Pacific basin) and is estimated to be for the most part of Cretaceous age (*i.e.*, about 120 to 85 million years old).

The seafloor in the western portion of the Philippine Sea developed between 60 and 35 million years ago. In the east, it was formed by backarc spreading from 30 million years ago (see above). The origin of the older crust is not completely clear. It either was created by spreading in the Pacific basin and subsequent capture by the formation of the Bonin and Mariana arcs, or it resulted from backarc spreading behind trenches to the south.

PALEOCEANOGRAPHY

Through knowledge of the ocean sedimentary record, the history of plate motions, glacial changes, and established relations between present sedimentation patterns and environmental factors, it is possible to reconstruct an oceanographic history for approximately the past 200 million years. This is the emerging field of paleoceanography.

Prior to the breakup of Pangaea, one enormous ocean, Panthalassa, existed on Earth. Currents in this ocean would have been simple and slow, and the Earth's climate was, in all likelihood, warmer than today. The Tethys seaway formed as Pangaea broke into Gondwana and Laurasia (see above). In the narrow ocean basins of the central North Atlantic, restricted ocean circulation favoured deposition of evaporites (halite, gypsum, anhydrite, and other less abundant salts). Evaporites also were deposited some 100 million years ago in the equatorial regions of the South Atlantic during the early opening of this ocean. Panthalassa

Sequences of organic-rich, black shales were deposited during the early phases of spreading in the North and South Atlantic. These sediments indicate anoxic conditions in the deep ocean waters. The oceans must have been well stratified into dense layers to prevent the overturning and mixing required to replace depleted oxygen. Black shales also were deposited in the older areas of the eastern Indian Ocean.

During the time interval between 200 and 65 million years ago, but especially from 100 to 65 million years ago, microplankton abundance and diversity increased enormously in the oceans. This resulted in increased deposition of biogenic sediments in the ocean basin. During Cretaceous time (from 144 to 66.4 million years ago),

sea level was often high, and shallow seas lapped onto the continents. This may have provided an environment favourable to the explosion in the numbers of species of foraminiferans, diatoms, and calcareous nannoplankton. Increased abundance of calcareous nannoplankton shifted the locus of carbonate sedimentation from shallow seas to the deep ocean. The end of Cretaceous time is marked by a sudden extinction of many life-forms on Earth, and marine organisms were no exception (see GEOCHRONOLOGY: *Cretaceous Period*). Coccolithophores (calcareous nannoplankton) and planktonic foraminiferans were particularly affected, and only a few species survived. Ocean sediments were suddenly less biogenic, and clays became widespread.

After Cretaceous time the Earth underwent a gradual cooling, especially at high latitudes. Deep-sea sedimentation changed as thermohaline bottom water circulation became fully developed (see above *Circulation of the ocean waters: Thermohaline circulation*). The CCD rose in the Pacific and dropped in the Atlantic as a result of changes in thermohaline circulation. An event of major significance was the spreading away of Australia from Antarctica beginning about 53 million years ago. This separation initiated limited circum-Antarctic circulation, which isolated Antarctica from the warmer oceans to the north, and led to cooling, which set the stage for later major glaciation.

Formation of Antarctic Bottom Water and its effects

At the Eocene-Oligocene boundary (36.6 million years ago), Antarctic Bottom Water began to form, resulting in greatly decreased bottom-water temperatures in both the Pacific and Atlantic oceans. Bottom-living organisms were strongly affected, and the CCD suddenly dropped from about 3,500 metres to approximately 4,000 to 5,000 metres in the Pacific. Bottom-water temperatures were generally warm, 12° to 15° C, during the time preceding this event. In a study of deep-sea sediment core material from near Antarctica, J.P. Kennett and Lowell D. Stott of the United States discovered that there was a period between roughly 50 and 35 million years ago when deep waters were very warm (20° C) and salty. The origin of these ocean waters was most likely in the low latitudes and resulted from high evaporation rates there.

The modern oceans are distinguished by very cold bottom water. The gradual changes toward this condition began 10 million years after the origination of Antarctic Bottom Water. Particularly significant among these changes was the closing of the Tethys seaway as Australia and several microcontinents moved north into the Indonesian region. Also, Australia moved far enough north that circum-Antarctic surface circulation became fully established.

The modern ocean circulation patterns and basin shapes were mostly in place by the beginning of Miocene time (nearly 24 million years ago). An exception was an ocean connection between the Pacific and Caribbean Sea in Central America that persisted until about 3 million years ago. Major and probably permanent ice sheets on Antarctica formed during Miocene time, and glacial sediments began to dominate the seafloor surrounding the continent shortly thereafter. Siliceous oozes also became widespread around Antarctica. Siliceous sedimentation increased in this area at the expense of siliceous sedimentation in equatorial regions. Ocean circulation became more vigorous, global climate became cooler, and sedimentation rates in the ocean basins increased. Planktonic microorganisms were segregated into latitudinal belts. Bottom-water flow north through the Drake Passage between South America and Antarctica began in Miocene time, resulting in erosion and nondeposition of sediments in the southwest Atlantic and southeast Pacific oceans. Also during Miocene time rifting between Greenland and Europe had progressed to a point where a connection was established between the North Atlantic and the Norwegian Sea. This resulted in the formation of North Atlantic Deep Water (see above *Circulation of the ocean waters: Thermohaline circulation*), which began flowing south along the continental rise of North America at this time. Sediments redistributed and deposited by this deep current are called contourites and have been extensively studied by Bruce Heezen, Charles D. Hollister, and Brian E. Tucholke, among others.

Sudden global cooling set in near the end of the Miocene some 6 million years ago. The strength of ocean circulation must have increased, as evidence of increased upwelling and biological productivity is present in ocean sediments. Diatomaceous sediments were deposited in abundance around the rim of the Pacific. This cooling event is synchronous with a drop in sea level, thought to be about 40 or 50 metres by various authorities, and probably corresponds to the further growth of the Antarctic ice sheet. This lowered sea level, coupled with the closure of narrow seaways probably due to plate movements, isolated the Mediterranean Sea. Subsequently, the sea dried up, leaving evaporite deposits on its floor. The Swiss geologist Kenneth J. Hsü and the American oceanographer William B.F. Ryan have concluded that the Mediterranean probably dried up about 40 times as seaways opened and closed between 6 and 5 million years ago. This evaporation removed about 6 percent of the salt from the world ocean, which raised the freezing point of seawater and promoted further growth of the sea ice surrounding Antarctica.

Enormous ice sheets emerged in the Northern Hemisphere between 3 and 2 million years ago, and the succession of Quaternary glaciations began at 1.6 million years ago. The exact cause of the glacial period is unclear, but it is most likely related to the variability in solar isolation, increased mountain building, and an intensification of the Gulf Stream at 3 million years ago due to the closing off of the Pacific-Caribbean ocean connection in Central America. The Quaternary glaciations, of which there were probably 30 episodes, left the most dramatic record in ocean sediments of any event in the previous 200 million years. Terrigenous sedimentation rates greatly increased in response to fluctuations in sea level of up to 100 metres and a more extreme climate. Biogenic sedimentation also increased and fluctuated with the glacial episodes. Deep-sea erosion began in many places as a result of intensified bottom-water circulation. (B.P.L.)

Continental margins

GENERAL CONSIDERATIONS

Physiographic divisions

Continental margins are the submarine edges of the continental crust, which is, as noted above, relatively light and isostatically high-floating in comparison with the adjacent oceanic crust. Figure 41 is a block diagram that schematically shows the physiographic divisions of a continental margin. These divisions are the continental shelf, the continental slope, and the continental rise. Characteristics of the various continental margins are shaped by a number of factors. Chief among these are tectonics, fluctuations of sea level, the size of the rivers that empty onto a margin as determined by the amount of sediment they carry, and the energy conditions or strength of the ocean waves and currents along the margin.

Margin types. Continental margins on the leading edges of tectonic plates, like those around the rim of the Pacific Ocean, are usually narrow and have steep continental slopes and either poorly developed continental rises or none at all. The continental slope is often steep and falls away directly into a deep-sea trench. In many cases, the leading-edge margins are backed by mountain ranges. Continental margins on the trailing side of tectonic plates, like those around the Atlantic Ocean, are broad, with gen-

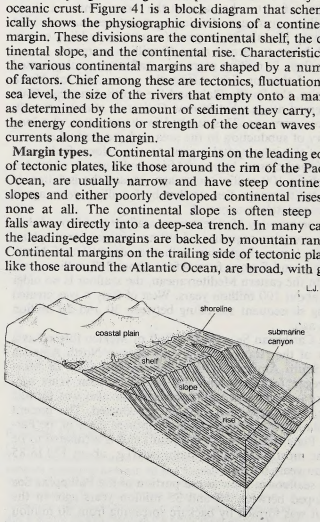


Figure 41: Physiographic divisions of the continental margin.

tle continental slopes and well-developed continental rises. The adjacent land area is commonly a broad coastal plain (see Figure 41) that, depending on the state of sea level, may become submerged from time to time and hence part of the continental margin.

Since continental margins are the shallowest parts of the world's oceans, they are most affected by changes in sea level. Worldwide changes in sea level, called eustatic sea-level changes, have occurred throughout geologic history. The most common causes of such sea-level changes are global climatic fluctuations that lead to major glacial advances and retreats—*i.e.*, ice ages and interglacial periods. Other causes that are not as well understood may include major mountain-building events and isostatic changes in crustal plates. When continental glaciers advance, as they did several times during the Pleistocene Epoch (which extended from about 1,600,000 to 10,000 years ago), water that would normally be in the oceans is locked up as ice on land, resulting in a drop in sea level. As the glaciers retreat, more water is fed to the ocean basins and the sea level rises. Fluctuations from highstand to lowstand have totaled 250 metres or more during Cenozoic time (roughly the last 66.4 million years), with concomitant fluctuations in exposure and flooding of the continental margins. (During a highstand the sea level is above the edge of the continental shelf, while during a lowstand it is below the shelf edge.)

Rivers bring a variety of sediments to the coast. These are classified by their mineralogy and by particle size and include sand, silt, and clay. Sand to sedimentologists is a grain of any composition from 63 to 2,000 micrometres (0.0063 to 0.2 centimetres) in its largest diameter. Silt is 4 to 62 micrometres, and clay is any particle less than 4 micrometres. Most of the detrital minerals brought to the continental margins by rivers in sand and silt sizes are quartz, feldspars, and mica; those of clay size are a suite of clay minerals that most commonly include smectite, kaolinite, and illite. (Clay can, in other words, refer either to particle size or to a group of minerals.) These, then, are the mineral constituents, which, together with calcium carbonates produced in the oceans by biogenic activity as shells and the hard parts of plants and animals, go to make up the sedimentary packages that are deposited on and constitute a fundamental part of continental margins.

A constant battle is being waged between the rivers that bring sediments eroded from the land to the sea and the waves and currents of the receiving body of water. This dynamic struggle goes on year after year, century after century, sometimes for millions of years. Take, for example, the north coast of the Gulf of Mexico into which the Mississippi River flows. The continental margin at this site is subject to relatively low wave and current energy, and so the river has filled up most of the adjacent continental shelf with a delta and dumps roughly 256 million tons of sediment each year directly at the top of the continental slope. By contrast, the Columbia River in the Pacific Northwest of the United States carries 131 million tons to the coast, where the sediments are attacked by the large waves and currents normal for that margin. As a result, sediments are widely dispersed, and the shelf is not filled with a large subaerial delta. Figure 42 shows the discharge of sediments to the continental margins of the world.

Effects of human intervention on sedimentation

The effects of this battle are easily seen where human activities have interfered with the transport of sediments to the sea by major rivers. For example, the Nile River delta is retreating rapidly, widening the submerged portion of the continental margin, because the Aswan High Dam has trapped much of the sediment normally fed to the delta front. The lower Mississippi River has been artificially maintained in a channel by high man-made levees. These have stopped the floods that fed much of the western delta margin. Because of this, coupled with a slow rise in sea level and the effects of canals dug in the delta wetlands, the coast has begun to retreat significantly.

When rivers carrying sediment from the interiors of continents reach the sea, several things happen. Velocity in the river jet decreases rapidly, and the sand particles drop out to be picked up by the waves and currents along the coast, where they feed beaches or barrier island systems (see CONTINENTAL LANDFORMS: *Structured landforms: Landforms produced by coastal processes: Landforms of depositional coasts*). If the river has a large enough discharge, the finer-than-sand-sized materials may be carried for kilometres onto the margin in a fresh- or brackish-water plume. The surf system then acts as a wave filter, trapping the sand in the coastal zone but allowing the finer materials to be carried out onto the margin. When estuaries are the receiving bodies of water on the coastal boundaries of continental margins, as in the case of the east coast of North America, virtually all the sediments brought down by the rivers are trapped within the confines of the estuaries.

In addition to the two primary types of continental mar-

Adapted from J.D. Milliman and R.H. Meade, "World-wide Delivery of River Sediment to the Oceans," *Journal of Geology*, 91(1), p. 16; copyright © 1983 by The University of Chicago Press, by permission of the publisher

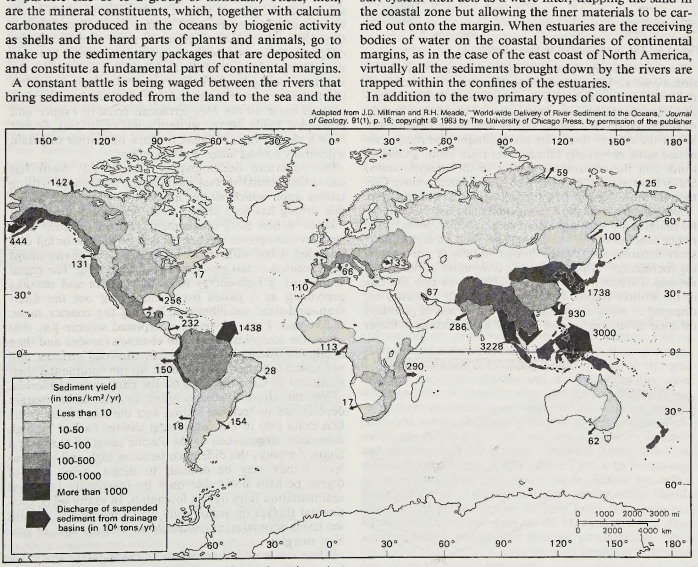


Figure 42: Discharge of sediments to the world's continental margins.

gin, there also are special types that do not readily fit either category. One of the most intensely studied margins of the world is the Borderland, the continental margin of southern California and northern Baja California. It consists of a series of offshore basins and ridges, some of which are exposed as islands. Figure 43 is a physiographic diagram of the Borderland. This system of basins and ridges formed as the result of faulting associated with the movement of the Pacific Plate past the North American Plate. It remains tectonically active today and is related to the San Andreas Fault system of California. A second special type is the marginal plateau. The Blake Plateau off the east coast of Florida is a good example. Such a plateau constitutes a portion of a continental margin that has many of the features of a normal system but is found at much greater depth—1,000 metres in the case of the Blake Plateau.

Marginal plateau

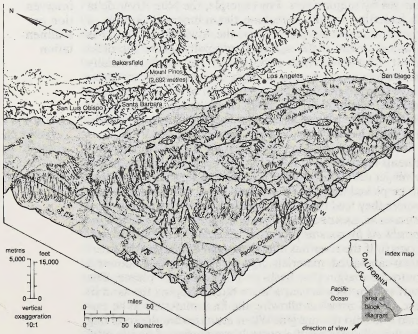


Figure 43: The Borderland of southern California.

Continental margins can be either constructional or erosional over varying periods of geologic time, depending on the combination of factors discussed above. When deposition exceeds erosion, the margin grows seaward, a process of progradation that builds out as well as up (Figure 44). When the erosive forces are predominant, the margin remains static or actually retreats over time. Some geologists think that the continental margin of the eastern United States has retreated as much as 5–30 kilometres since Cretaceous time—*i.e.,* during the last 66 million years or so.

Economic importance of continental margins. Continental margins are very significant economically. Most of the major fisheries of the world are located on them. Of these, sport fisheries and related tourist industries are becoming increasingly important to the economies of developed nations. Paradoxically, continental margins also are one of the world's biggest dump sites. All kinds of wastes are disposed of along the margins, and the effects of pollution on their environment and ecology have become a major global concern (see also below *Economic aspects of the oceans: Waste disposal and other related actions*).

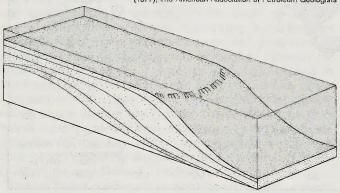


Figure 44: Prograded slope.

Continental margins are the only parts of the world's oceans to be effectively exploited for mineral resources. Millions of tons of sand are mined by dredges each year off the U.S. coasts alone for beach renourishment projects. From time to time placer deposits also have been worked. Examples include tin off Indonesia, gold off Alaska, and diamonds off Namibia. By far and away the largest mineral resources to be exploited from continental margins are oil and natural gas. Exploration of the continental margins by major oil companies has intensified and is expected to continue for the foreseeable future because the margins are the most likely sites of giant undiscovered petroleum deposits. Continental margins are made of thick accumulations of sedimentary rock, the type of rock in which oil and gas generally occur. In fact, most of the sedimentary rocks exposed on the continents were originally deposited on continental margins; thus, even the hydrocarbon deposits found on land were formed for the most part on ancient continental margins.

CONTINENTAL SHELF

Continental shelves make up about 8 percent of the entire oceanic area. They extend from the shore to the shelf break, the point at which the angle of inclination steepens sharply, marking the boundary between the continental shelf and the continental slope. Built entirely on continental crust, the shelves are terraces gently inclined seaward at an average slope of about 0.1°, or about 2 metres per kilometre (see above Figure 41). Most shelves have a gently rolling topography called ridge and swale. Their geology is often similar to that of the adjacent exposed portion of the continent. The width of continental shelves at the present sea-level stand varies from a few kilometres to more than 400 kilometres. Throughout geologic time, the width of continental shelves has varied greatly with the rise and fall of eustatic sea level. Figure 45 is a representation of Florida and its continental shelf 18,000 years ago and at the present time. The depth of the shelf break varies widely from about 70 metres in the Beaufort and Chukchi seas in the Arctic region to more than 400 metres off parts of Norway, with the average at about 145 metres. Abnormal depth of the shelf break is most often the result of loading by glacial ice, either at present or within the last few tens of thousands of years. A few margins, such as those off the Mediterranean coast of France and Porcupine Bank, off the western coast of Ireland, do not have a sharply defined break in slope but rather maintain a generally convex shape to the seafloor.

Ridge and swale

The American oceanographer Donald J.P. Swift has called continental shelves palimpsests, parchment writing tablets upon which stories are written after each previous writing has been erased. Each new stand of sea level "writes" a new story of sedimentation on the shelf after the previous episode has been erased by the rise or fall that preceded it but with traces of the previous environment of deposition or last erosional event remaining. The eraser is the surf, a high-energy force that erodes and reworks everything as it passes over, winnowing out the finer-than-sand-sized sediment and leaving the coarser material behind. Figure 46 is an interpreted seismic line that shows the complicated array of channels (eroded and then filled), old deltaic deposits, ancient erosional surfaces, and winnowed sand bodies that make up the continental shelf southwest of Cape San Blas on the panhandle of Florida.

How the above processes affect any particular margin depends on its tectonic setting and the size of the rivers that drain into it. On continental shelves backed by high mountain ranges, such as the Pacific coast of North and South America, the difference between high and low sea-level stands may be difficult to detect, being one of degree perhaps noticeable only by marginally increased sedimentation rates during lowstands. In many ways, continental shelves on tectonically active margins at present sea levels approximate lowstands on trailing-edge, or passive, margins.

When sea level is lowered on a trailing-edge shelf that has no adjacent high mountains, such as the Atlantic coast

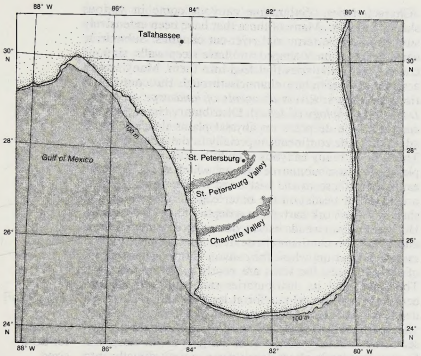


Figure 45: A representation of Florida's continental shelf some 18,000 years ago.

of North America, rivers are rejuvenated—*i.e.,* their base level is lowered and they begin to erode their beds, carrying sediment from the continent across the former continental shelf that is now exposed and depositing it at the new coast. When sea level falls below the shelf break, the coast lies on the continental slope. As sea level rises again on tectonically stable or sinking shelves, medium- and small-sized river mouths drown and estuaries form, trapping the sediment within them and starving the shelves. In these cases, sediment for the shelf is primarily produced by erosion of the coastline as the surf zone advances landward with rising sea level. Fine-grained material is winnowed out, to be either deposited back in the estuaries or carried in steps by advective processes across the shelf to the deeper water beyond. As a result, continental shelf surfaces on trailing-edge margins into which no large rivers flow are veneered with a sand sheet lying over a complex of older deposits, some of which peek through the surface as outcrops—vestiges of an earlier story written on the palimpsest. Large rivers that drain a large, high continent, such as the Mississippi, are able to keep pace with rising sea level and deliver enough sediment to keep an estuary from forming, and, at a high stillstand like that of the present, even fill their entire shelf area. (For a description of modern deposits of this type, see RIVERS: *Rivers as agents of landscape evolution: Deltas.*)

For many years after World War II, the period when many of the world's continental shelves were first described in detail, it was thought that the sand deposits on continental shelves were "relict," deposits left stranded by a higher sea level from the higher-energy regime of the surf zone that passed over them perhaps as much as a few thousand years before. Geophysical investigations of the shelf area since the mid-1970s have revealed the presence of many types of sand waves and ripple marks in seafloor sediments that show submerged continental shelf sediments to be constantly undergoing reworking and erosion. Figure 47 shows a side-scan sonograph of a large-scale sand-wave field with the associated bathymetric record. As scientific understanding of the physical processes that affect continental shelves has increased, it has been found that currents set up by large winter storms, monsoons, hurricanes, and typhoons are reworking the bottom by winnowing out the fine-grained materials and carrying them either back into the estuaries or beyond the shelf break, where they are lost from the system.

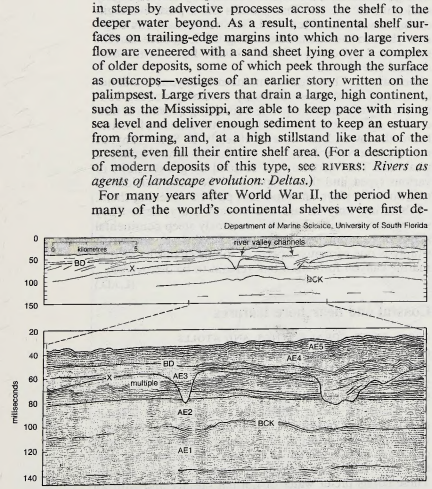


Figure 46: Seismic line interpretation revealing features of the continental shelf southwest of Cape San Blas in Florida. These features, which include the filled river valley channels, old deltaic deposits, and ancient erosional surfaces shown in the top drawing, can be discerned from the seismic record section from this shelf segment, bottom.

Factors affecting shelf sediment type

In short, the kind of sediment that covers the surface of a continental shelf is determined by the interplay among the tectonic setting, the size of the rivers that empty into it (size based on how much sediment they carry), and the wave energy that affects it, just as is the case with continental margins in general. Shelves such as that of western Florida that have been cut off from clastic input (*i.e.,* sediments composed chiefly of quartz and clay minerals derived from erosion of the continent) may be covered with carbonate sediments. In some cases, as in the islands of The Bahamas, the carbonate shelf, called a bank, is cut off from a continental source by deep water. Continental shelves with rivers that carry sediments from the continents to the shelf and beyond only at lowstands of sea level and those that drain mountainous areas on high-energy coasts are dominated by quartz sands, and shelves with rivers that drain large continental areas and carry enough sediment to keep abreast of sea-level rise or dominate ambient wave-energy conditions will accumulate muddy sediment deposits out across their surfaces.

CONTINENTAL SLOPE AND RISE

About 8.5 percent of the ocean floor is covered by the continental slope-rise system. This system is an expression of the edge of the continental crustal block. Beyond the shelf-slope break, the continental crust thins quickly, and the rise lies partly on continental crust and partly on the oceanic crust of the deep sea. Inclination of the continental slope averages about 4°, although it can approach vertical on carbonate margins, on faulted margins, or on leading-edge, tectonically active margins such as that off the above-mentioned Borderland. Steep slopes usually have either a very poorly developed continental rise or none at all and are called escarpments. Partway down the continental slope there is often a second inflection point where the gradient drops to 1:100 to 1:700. This lower gradient area defines the continental rise, which may extend seaward several hundred kilometres to the deep seafloor at an average depth of about 4,000 metres. Although the slope-rise was formerly often divided into separate provinces, the trend among sedimentologists is to consider it as part of one continuum. The sedimentary processes that form the slope and rise, however, are quite distinct.

Formation of continental slopes. The continental slopes are temporary depositional sites for sediments. During lowstands of sea level, rivers may dump their sedimentary burden directly on them. Sediments build up until the mass becomes unstable and sloughs off to the lower slope and the continental rise. During highstands of sea level, these processes slow down as the coastline retreats landward across the continental shelf, and more of the sediments delivered to the coast are trapped in estuaries and lagoons. Still the process continues, albeit slowly, as sediments are brought across the shelf break by winnowing of the shelf surface and by advection. Slopes are sometimes scoured by such major ocean currents as the Florida Current that work to erode their surfaces. Off active major deposition centres, such as the Mississippi delta, slope sequences may accumulate through progradation, while the active slope front is continuously shedding sediments downslope by gravity processes.

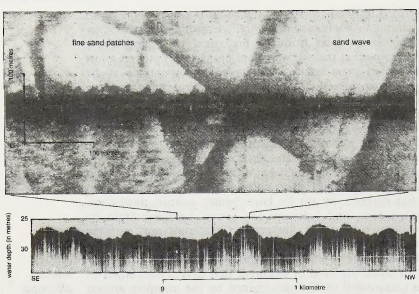


Figure 47: (Top) Side-scan sonograph of a sand wave field and (bottom) corresponding bathymetric record.
From K.T. Logue, "Bedforms and Surface Sediments of the Northwest Florida Inner Continental Shelf" (1980), University of South Florida

Formation of continental rises. Continental rises are major depositional regimes. They consist of thick sequences of continental material that build up as the result of three sedimentary processes.

Sedimentary processes responsible for rise development

The first such process is a downslope movement of sediments by mass wasting, a set of gravity-deposition events, including submarine landslides, slumps, debris flows, and high-velocity, sediment-laden density flows known as turbidity currents (see above). Several phenomena may initiate gravity events. In tectonically active areas earthquakes are important triggering mechanisms. Even in the Atlantic they play a significant role. One of the few documented major gravity events took place on the Grand Banks of Newfoundland in 1929, when an earthquake triggered a gravity flow that possibly attained velocities of more than 90 kilometres per hour and was traced for hundreds of kilometres as it successively broke transatlantic cables. Other triggering events may be oversteepening of deposits on the sharply inclined portions of the continental slope, breaking internal waves that have been shown to affect the upper slope, and storm waves and storm-induced currents.

A second process that may be equally important, although its overall significance is subject to considerable scientific debate, is deposition from bottom currents that flow parallel to the slope of the continental rise—namely, contour currents. Resulting sediment accumulations are called contourites. The major points of contention concerning the efficacy of contour currents are (1) whether or not they are strong enough—they flow at a speed of about 20 centimetres per second—to build the huge thicknesses of sediment that make up the rises, and (2) how the sediments get into the contour currents in the first place. It is probable that most of the mass of rise material is originally brought downslope by gravity events and then redistributed by contour currents.

Vertical settling through the water column of both clastic and biogenic particles is the third contributor of slope and rise sediments. These pelagic sediments are composed of clay minerals and fine-grained particles (chiefly quartz, mica, and carbonate) swept off the continental shelf, windblown dust, organic detritus, and the tests of planktonic plants and animals. Chief among the last group are the tests of foraminiferans, pteropods, and coccolithophores that are composed of calcium carbonate and those of diatoms and radiolarians that are made of silicon dioxide.

SUBMARINE CANYONS

Originating on the continental shelf and cutting into the continental slope-rise system are some of the most spectacular physiographic features on the Earth—the submarine canyons. They have been the object of investigation for many years. Much of the definitive work was done by Francis P. Shepard of the United States, widely regarded as the father of modern marine geology, and Robert F. Dill, one of his students.

Characteristics. Submarine canyons come in various shapes and sizes. Many of those that have been extensively studied resemble terrestrial river-cut canyons. Submarine canyons are often V-shaped and have steep walls, sinuous courses, and tributaries that feed into them. Many end in a series of distributary channels through the continental rise (see RIVERS: *Rivers as agents of landscape evolution: Deltas: Morphology of deltas*). Distributary channels often emerge in the deep sea on abyssal plains that are found adjacent to the continents, especially in the Atlantic basin. The scale of many canyons is truly herculean. For example, the Hudson submarine canyon, which serves as a type canyon for those of the east coast of the United States, originates (or heads) on the outer edge of the continental shelf off New York harbour at a depth of about 90 metres. Hudson Canyon wends its way down the continental slope and enters the continental rise at a depth of about 2,100 metres. Higher up, where the canyon floor reaches a depth of 1,800 metres, the walls are nearly 1,220 metres high. The valley and its distributaries can be traced seaward across the continental rise for at least 300 kilometres to a depth of more than 4,500 metres.

Conduits for sand-sized sediments

Submarine canyons act as conduits to bring sand-sized sediments from the continental margins to the deep sea. During lowstands of sea level, rivers empty directly into the heads of many Atlantic canyons. Sand and mud are carried down these systems, many times bypassing the slope-rise system to be carried directly out onto the abyssal plains on the deep seafloor. Cores taken in abyssal plains in the Atlantic Ocean off North Carolina show distinctive shelf molluscan fragments in gravity-flow deposits that have traveled hundreds of kilometres and that can be traced back through the Hatteras submarine canyon system. During highstands of sea level, the submarine canyons off the east coast of North America are many tens of kilometres from the coastline, and the downslope movement through them is slowed down dramatically or perhaps even ceases. On the west coast of the continent where active tectonism results in a narrow margin backed by high mountains, a situation exists today that is roughly analogous to the one that prevails on a passive margin during lowstands of sea level. Submarine canyons originate close to the coast, intersecting the longshore current system and siphoning off sand to basin floors in the Borderland (see above).

Formation. The origin of submarine canyons is still not completely understood. Probably several processes contribute to their formation. The underlying question is how erosion occurs, often in solid rock, thousands of metres below the lowest known stand of sea level. Some canyons may be located on fault lines or other structural zones of weakness. They certainly are scoured by gravity flows of various types, and the collapsing of the outer portions of canyon walls plays an important part in widening them.

Submarine canyons are not present everywhere. They are rare on margins that have extremely steep continental slopes or escarpments. The reasons for this are unclear. It simply is one of many aspects of the mechanisms of submarine canyon formation that remain to be discovered.

(L.J.D.)

Coastal and nearshore features

CORAL REEFS, CORAL ISLANDS, AND ATOLLS

Coral reefs are masses of carbonate of lime built up from the seafloor by the accumulation of the skeletons of a profusion of animals and algae; eventually they rise to the surface of the water.

Reef-building corals, chiefly the stony corals or Scleractinia, grow best in shallow, sunlit water, between the low-water mark and a depth of 11 metres, but they can still construct reefs in water as deep as 40 metres, and they may have a sparse existence between 40 and 55 metres. These corals prefer water of normal salinity and with an annual maximum temperature above 22° C but below 28° C. Their reef-building activities, however, may be carried on in waters whose minimum temperature in winter is not less than 15° C.

A second group of corals in present-day seas grows in

thickets and coppices that develop banks rather than reefs on the outer, deeper, colder, and darker parts of continental shelves and platforms. These organisms flourish in water with a winter minimum temperature ranging between about 4° and 15° C at depths of about 60 to 200 metres. In any one thicket there are commonly only two genera of delicately branching corals involved. Such coral banks are known along the eastern Atlantic shelf edge (or continental slope) from Norway to the Cape Verde islands and again off the Niger River delta and in the west Atlantic around the Gulf of Mexico, The Bahamas, and the Orinoco River delta. Off New Zealand such banks have been recognized on the Campbell Plateau and the Chatham Rise; they also occur in the northwest Pacific near Japan.

The third coral assemblage of the modern seas is associated with even colder or deeper seas; it consists of small, solitary corals of relatively few genera, known from the abyssal floors of the oceans and from the shelves around Antarctica, Patagonia, and the Falkland Islands in waters 2° to 6° C in temperature.

Fossil reefs

Buried fossil reefs on ancient continental shelves are targets for petroleum exploration. The porosity of reefs and the characteristic curvature of nonporous enclosing sediments cause them to be prospective reservoirs for oil and gas. The rich oil fields of Alberta, for example, are associated with Devonian reefs (about 408 to 360 million years old). Fossil reefs recently have become targets for metal prospecting because some corals contain small percentages of metals, such as zinc and copper, selectively incorporated from seawater. A living coral reef may also have economic potential in that it constitutes a major tourist attraction.

Reef accumulation. *Tropical water conditions.* Water conditions favourable to the growth of reefs exist in tropical or near-tropical surface waters. Regional differences may result from the presence or absence of upwelling currents of colder waters or from the varying relation of precipitation to evaporation.

Tropical seas are well lit, the hours of daylight varying with the latitude. Light intensity and radiant energy also vary with the depth. Thus, at latitude 32°44′ N (the Madeira Islands) the "day" in March has a length of 11 hours at a depth of 20 metres, 5 hours at 30 metres, and only about a quarter of an hour at 40 metres. Nearer the pole these figures decrease further. Light intensity has a profound effect on the growth of the individual reef-coral skeleton because of the symbiont zooxanthellae of reef corals (see below *Biological factors*). The number of species present on a reef also may be related to light intensity and radiant energy.

Turbidity may be high in lagoons, where shallow water lies over a silt-covered seafloor and where storms and windy periods cause considerable disturbance of the bottom silt. The average transparency may be low (about 12 metres), and light penetration is reduced.

Inside the Great Barrier Reef, on the shallow continental shelf of Queensland, the oxygen content of the water is high, exceeding 90 percent saturation most of the time; in deeper water, during the calm periods of the rainy season, the saturation may fall to about 80 percent. Plant nutrients such as phosphate and nitrate show no seasonal change in quantity; both are present in very small quantities throughout the year. Constant mixing of the shallow sea prevents any stratification of the nutrients. As a result, growth of phytoplankton is possible and almost uniform throughout the year, providing a constant supply of food for the zooplankton, which in turn form the chief food supply of the corals. Some nutrients enter the lagoonal waters with the oceanic water that flows through the reef openings, but the dissolved phosphates in the lagoons are probably derived chiefly from bacterial decomposition of the organic matter on the sea bottom, as well as from detritus swept in from the reef surfaces. This environmental pattern is typical of many atoll lagoons.

Geochemistry of reefs. Minute quantities of metallic elements are present in solution in seawater and also occur in marine invertebrate skeletons, though not in the same proportions as in the surrounding water. Magnesium and strontium are the most frequently occurring trace elements in reef skeletons and are measured in parts per thousand, but barium, manganese, and iron are also present and can be measured in parts per million. In Pacific corals, 2.17 parts per million of uranium have been found, and in Florida coral, 2.36–2.95 parts per million. Strontium is concentrated in aragonitic skeletons, and magnesium in calcitic skeletons; coral aragonite has a higher strontium content than (some) molluscan aragonite; the magnesium content in the calcite of coralline algae is high; that of bar-

Trace elements in reef skeletons

By courtesy of the Geological Society of America

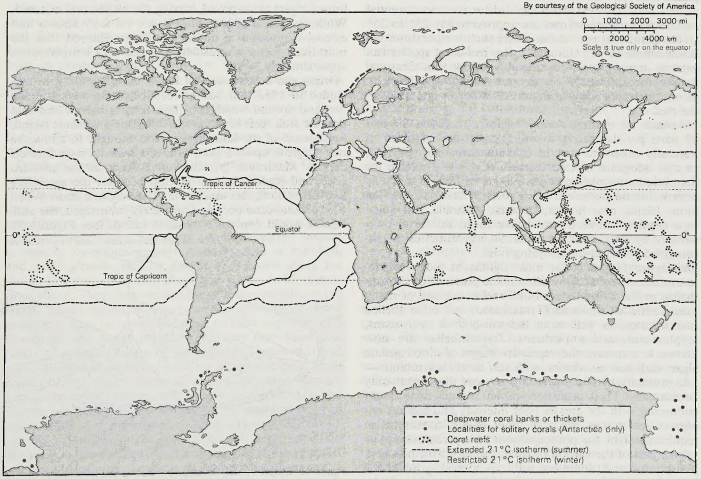


Figure 48: Distribution of coral reefs and their relation to sea-surface temperature.

nacle shells is low (11.5 parts per thousand). By identifying these trace elements and their degree of assimilation in different organisms, sediments formed predominantly of coral-skeletal detritus can be distinguished from sediment derived chiefly from mollusks or coralline algae.

Atomic-absorption spectrophotometry has shown that ultratraces of metals are present in the aragonite skeleton of the hydrozoan coral *Millepora* from a reef flat on the Coral Sea Plateau off Queensland. These are, in parts per billion: lead (100), copper (71), cadmium (23), cobalt (17), nickel (1,480), iron (507), and zinc (507).

Another aspect of reef geochemistry is the carbon and oxygen isotopic composition of coral skeletons and shells. Determination of the number of carbon isotopes present provides a method of assessing the age of a sample, and determinations of oxygen isotopes present are useful in indicating water-temperature changes that occurred during the period of growth of the reef.

Winds, currents, temperature, and salinity. Winds and currents are important in shaping individual reefs and in determining the orientation, shape, and position of the coral sand cays, or "low islands," that develop on reefs. Currents are primarily those generated by the prevailing winds, but, in areas where the tidal range is great, tidal effects may become paramount.

Cays may be round, oval (or boat-shaped), or irregular in outline. They originate when sediment is lifted from the reef surface and carried leeward by waves or tidal currents and then deposited where the water velocity is reduced abruptly. Thus, they commonly form on the more protected leeward end of the reef. Wind action at low tide on these deposits may build dunes above the high-water mark. Beach rock may form by carbonate cementation of grains in deposits lying between tide levels. It then acts as a stabilizing factor. Storm waves may drive forward coral fragments derived from "stag-horn" corals growing on the windward slopes of the reef, forming shingle banks; successive, superposed banks may thus be formed. The shingle on the banks may become cemented and thus add considerable stability to the cay, as does the growth of vegetation. Hurricanes, however, may carve back the shorelines of even stabilized cays. Huge, isolated boulders of coral or coral limestone are fairly common along reef margins. Some may be remnants of a once-emergent reef platform; others are hurricane or storm jetsam.

Restrictions on reef formation

Coral reefs are best developed where the mean annual surface-water temperatures are approximately 23° to 25° C. No significant reefs occur where such temperatures fall below about 18° C, although a few reef-coral species can exist in temperatures considerably below this. Seasonal temperature differences on any one reef are usually slight, as are differences due to depths of water or situation on the reef.

Seawater of normal oceanic salinity (between 30 and 40 parts per thousand), to which corals are restricted, is normally supersaturated in calcium carbonate ($CaCO_3$), so that adequate ionized calcium (Ca^{2+}) is available for the skeleton-forming process. Floods of fresh water may destroy life on inshore fringing reefs. A luxuriant reef on Stone Island, near Bowen, Queens., Australia, was killed to a depth of three metres below mean tide level by a week of cyclonic rains, in which 90.7 centimetres of rain coincided with full-moon spring tides.

Biological factors. The most significant biological determinant of reef accumulation is the presence of zooxanthellae in the living tissues of all reef corals and of many massive-shelled mollusks (Tridacnidae) and other shelled invertebrates, as well as in the soft-bodied hydrozoans, scyphozoans, and anthozoans. Zooxanthellae are now known to represent the vegetative stages of dinoflagellate algae, and their association with reef corals is symbiotic—*i.e.*, mutually helpful. In temperate seas they occur only occasionally. Their profusion in reef animals is no doubt connected with the greater light intensity and radiant energy of reef waters, for, like other plants, zooxanthellae require sunlight for photosynthesis. They remove at the source part of the carbon dioxide, together with nitrogen, phosphorus, and (doubtless) sulfur, produced by metabolic breakdown within the coral and which would otherwise be excreted by the corals. They greatly aid in the formation of the coral skeleton by increasing the speed with which the carbon dioxide produced in coral metabolism is removed and the speed with which the skeletal calcium carbonate is formed. Corals also may gain some nutrient from their zooxanthellae, but they probably do not need the oxygen produced during photosynthesis.

The productivity of reefs is a current focus of interest. A constant supply of food in the form of zooplankton is essential to reef corals, which are carnivorous. The zooplankton supply is dependent on an adequate phytoplankton supply, and the phytoplankton, in turn, require an adequate supply of plant nutrients dissolved in the water. An atoll in the open ocean may be compared to an oasis in a desert, as a localized centre of high productivity. In the warm, well-lit, and well-mixed lagoon waters there is a rapid turnover of the endemic planktonic (floating or swimming) and benthic (bottom-dwelling) population, perhaps 12.5 times per year. Carbonate skeletal matter is accumulated perhaps 1,000 times faster on the summit of an atoll than in the surrounding deeps.

Certain biological factors may contribute to the destruction of coral reefs: the fish and invertebrates that feed on the soft tissues of reef builders and the organisms that bore into coral rock. Of the former, the most destructive such enemy yet known is *Acanthaster planci*, the crown-of-thorns starfish, which, during the 1960s multiplied spectacularly and removed the soft tissues from large areas of many reefs in the southwest Pacific. *A. planci* feeds by everting its stomach and liquifying and absorbing the tissues of the corals. By the late 1970s it had become apparent, however, that the sudden spread of *A. planci* was part of the organism's natural life cycle and that the coral reefs could regenerate rapidly after such an infestation. Coral-rock borers include boring algae, boring sponges (of great significance), various polychaete and sipunculid worms, and many bivalves and a few gastropods. These organisms usually penetrate the rock mechanically but in some cases do so chemically. Extensive damage is caused both by their own activities and by the assistance they give to the erosive action of the sea. A phenomenon known as "bleaching" caused extensive devastation among coral reefs in the east Pacific during the early 1980s and in the Caribbean during the mid- to late 1980s. It is called bleaching because zooxanthellae (principally the so-called brown algae) are expelled, leaving the white coral exposed. While the cause of bleaching is not yet fully known and extensive research is under way, it is believed that the most likely factor is unusually high seawater temperatures, approximately 30° C.

Darwin's theory of atolls

Origin and development of reefs. Charles Darwin concluded in 1842 that barrier reefs began as reefs fringing the land around which they now form a barrier and that oceanic atoll reefs began as reefs fringing a volcanic island. Subsidence of the land fringed was thought to allow the reef to grow upward (and outward over its own forereef debris). Maximum growth would occur at the seaward edge, and lagoons would develop between the ascending barrier, or atoll, reef and the land or volcanic cone. When the volcanic cone became completely submerged, the atoll lagoon would contain only coral islands (see Figure 49).

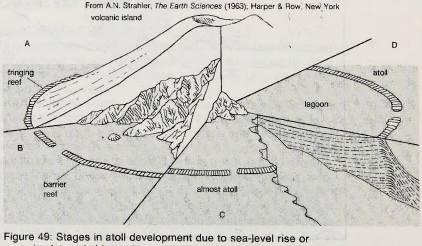


Figure 49: Stages in atoll development due to sea-level rise or oceanic plate subsidence (see text).

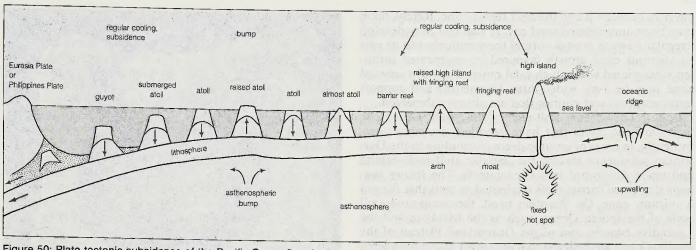


Figure 50: Plate-tectonic subsidence of the Pacific Ocean floor in the evolution of mid-ocean coral reefs. The arrows indicate either subsidence or uplift.

From A. Guilcher, *Coral Reef Geomorphology*, copyright © 1988 John Wiley & Sons, Inc., reprinted by permission of John Wiley & Sons, Inc.

Fundamentally, Darwin's concept is still valid, although many consider submergence by the rise of sea level, following melting of Pleistocene ice sheets, to be a better explanation of the latest upward growth of many reefs, particularly on continental shelves. Mid-ocean stages of coral reef development are explained by plate tectonic theory (see above *Ocean basins*), according to which the ocean floor subsides as it spreads outward from oceanic ridges. The Hawaiian Islands, with barrier reefs in the southeast grading to atolls in the northwest, is a good example of this (see Figure 50).

Planing by marine erosion

A reef whose surface lies above high-tide mark, either by uplift or by eustatic regression of the sea (which is determined by ice sheet-sea level relations), is subject to planing by marine erosion. If planing off is complete, a flat-topped submerged platform results; if subsidence or eustatic submergence intervenes, a wave-cut terrace is left around the reef. Terraces that may have formed in this way are known around many reefs. Some annular reefs may develop without relation to subsiding volcanic cones. When reef platforms have been uplifted above sea level, they are subjected to subaerial erosion. Surface slope, or gradient, determines the amount of runoff and is a prime factor in this erosion. Two secondary processes also are involved: (1) case hardening of steep, bare limestone surfaces by recrystallization caused by alternate wetting and drying, so that walls or knife edges result from weathering; and (2) continuous subsoil solution, if surfaces are nearly horizontal and runoff is diminished. These processes combine to produce a prominent rim and a saucer-shaped interior in emerged limestone islands. With submergence, algal and coral growth resumes, the fastest growth being on the rim and on any pinnacles that may be left. Thus an atoll or annular reef may develop along the rim around the low-lying central region, which becomes a lagoon, and coral knolls grow on former pinnacles in the lagoonal area.

Types of reefs. *Fringing reefs.* Fringing reefs form a veneer in the shallow water near or at the shore of the mainland or of islands. On shorelines where bays receive large quantities of terrestrial mud, sand, and fresh water, fringing reefs are intermittent and are restricted to promontories. Along limestone coasts, however, coastal erosion is by solution, little mud or sand is supplied, and coral growth may be almost continuous along the shore. Fringing reefs may extend as far as 1,500 metres from shore. They show ecological zonation parallel to the shore. Along mainland shores with easily eroded rocks, mudbanks may be washed into the reef flat and colonized by mangroves. Inner parts of reef flats may show low mesas of middle Holocene (perhaps 5,000 years old) and even Pleistocene emergent reefs that have not been quite planed off by marine abrasion. Geomorphologic features peculiar to raised fringing reefs have been described for the Solomon Islands, and there the complex problem of dead reefs and dead patches on reefs arises. The surface of the reef flat is mostly of dead coral, but pools occur in which live coral colonies flourish. An algal rim formed by red calcareous algae may develop on a fringing-reef margin if the reef faces strong waves and swells the year around. This rim is commonly less spectacular than the algal ridge of the windward edge of Pacific oceanic atolls and may be developed merely as an algal platform, or pavement, on which algal encrustation over corals is thin. The seaward slope of a fringing reef, like the seaward slope of an atoll reef, is characterized by a zone of grooves and spurs to a depth of perhaps 15 metres, and it is in this zone that vigorous growth occurs. Under stable shelf conditions, the fringing reef will extend outward as the spurs elongate and the grooves fill or roof over with coral and algal growth. In tectonic belts (zones of uplift and deformation), fringing reefs may be uplifted intermittently, resulting in parallel, stepped subaerial terraces such as those of the Finsch Coast of Papua New Guinea.

Platform and patch reefs. Platform and patch reefs are characteristic of continental shelves; they may or may not lie behind a barrier reef. Reefs grow actively outward as well as upward, especially in the stable conditions of a continental shelf. Any given reef, having depth and temperature fixed by its location, will have its shape determined by the direction and force of the water currents that bring food and by the shape of the base on which it grows. Where the forces of growth are equal in all directions, radial expansion results in platformlike reefs. With further radial growth, lagoonal platform reefs develop. If the reef grows on a sand bank, elongation may result. The shape of an elongated platform reef may be determined by the orientation of rising and falling tidal currents. These may be directly opposed to each other. The boat-shaped reefs of Torres Strait, between Australia and New Guinea, apparently developed in such a pattern. Where wave-generated currents are asymmetric, horseshoe reefs develop, with convexity facing the current and the leeward ends curving round to partly surround a lagoon. Low Islets, made famous by the Great Barrier Reef Expedition of 1928–29, is the best-known example of this type. A sand cay (or cays) commonly develops on one or both of the leeward wings. Those parts of Pacific platform reefs that face strong and persistent currents characteristically have a low algal rim from which radiate grooves and spurs.

Horseshoe reefs

Barrier and ribbon reefs. Barrier reefs commonly present to the ocean and the trade winds a steep wall, some dropping abruptly 1,000 to 5,000 metres. On the lagoon side they grade off gently with a wedge of sediment dotted by small patch reefs, coral knolls, and coral heads. Depths in the lagoon may reach 50 to 80 metres. According to the strength of surf and swell, an algal ridge, rim, platform, or pavement develops and is commonly emergent at low water. The seaward slope has radial grooves and spurs, the grooves forming surf channels and imparting great wave resistance. Ecological zonation on the seaward slope is difficult to study because of the dangers of surf and swell. Zonation on the reef flat, parallel to the wall, is dependent mainly on depth.

A characteristic form in a barrier reef system is the wall, or ribbon, reef, emergent at low tide, such as Yonge Reef, Queens., Australia. A ribbon reef flat is commonly only 300 to 450 metres wide from the seaward wall to its lagoonward edge. Its ends may curve leeward and border the

passages between it and the next reefs in line. Rarely, there may be an unvegetated sand cay. A wall reef may develop irregular leeward prongs normal (perpendicular) to its axis by vigorous coral growth favoured by augmented turbulence associated with a high tidal range. The open leeward zone may be very wide (three kilometres) and support large, scattered reef clumps that are always submerged.

Atolls. The oceanic atoll reefs of the Pacific Ocean rise from volcanic cones that have subsided, probably intermittently, in areas of oceanic deeps. According to the Darwinian subsidence theory, the annular atoll reefs extend and grade downward into barrier reefs. The theory suggests that these barrier reefs originated as reefs that fringed a volcanic cone. On the other hand, the compound atoll reefs of the Indian Ocean, such as the Maldives and the Laccadive Islands, and of the Queensland Plateau of the Pacific Ocean are believed to have grown above foundered continental (rather than oceanic) crustal segments. The Nicaraguan atolls rise from a seafloor of 1,000 metres or more in a volcanic province. (D.Hi./M.L.S.)

LAGOONS

Lagoons are, as indicated above, areas of relatively shallow water situated in a coastal environment and having access to the sea but separated from the open marine conditions by a barrier. The barrier may be either a sandy or shingly wave-built feature, or it may be a coral reef. Thus, there

Principal types

are two main types of lagoons: (1) elongated or irregular stretches of water that lie between coastal barrier islands and the shoreline and (2) circular or irregular stretches of water surrounded by coral atoll reefs or protected by barrier coral reefs from direct wave action (Figure 51). Lagoons of the first type are characterized by quiet water conditions, fine-grained sedimentation, and, in many cases, brackish marshes. Water movements are related to discharge of river flow through the lagoon and to the regular influx and egress of tidal waters through the inlets that normally separate the barrier islands. Lagoons of the second type are best exemplified by the roughly circular quiet waters that are surrounded by coral atoll reefs. Lagoon depths are maintained at a moderate level by sedimentation, and this compensates for the subsidence that commonly attends reef formation (see above *Coral reefs, coral islands, and atolls*). Because the reef is an organic structure, the lagoonal sediments contain much calcareous material. The sheltered waters support a distinctive flora and fauna.

Coastal lagoons are widely distributed throughout the world and have been estimated to constitute about 13 percent of the total world coastline. Lagoons are more common on coasts with moderate to low tidal ranges; for example, they occur widely on low coasts of the southern Baltic, southeast North Sea, Black Sea, Caspian Sea, and Mediterranean Sea and of the southeastern United States and the Gulf of Mexico. Lagoon coasts also occur along southern Brazil; the east coast of Madagascar; northeastern Russia; Japan; and isolated parts of Africa, India, Australia, and New Zealand.

Typical sites

Lagoons are generally characteristic of coasts of low or moderate energy, occurring especially on the east coasts of continents where the swells are less violent and in high latitudes where offshore ice provides some protection. They also are associated with low coasts and rarely occur where high cliffs form the coast. They can form only where there is abundant sediment for construction of the protective barrier islands. Too much sediment from the mainland, however, can lead to delta formation rather than lagoons, although lagoons frequently occur along the outer delta margin and between delta distributaries (see also RIVERS: *Rivers as agents of landscape evolution: Deltas*).

Coral lagoons are restricted to tropical open seas that provide the conditions necessary for coral growth. They occur widely in the western Pacific, in parts of the Indian Ocean, and in isolated places in the Caribbean, mainly within 25° latitude of the equator. Coral lagoons are of great importance to many island communities in the Pacific, particularly where they provide the only quiet water for use as harbours, although the passage through the reef into the lagoon is often perilous.

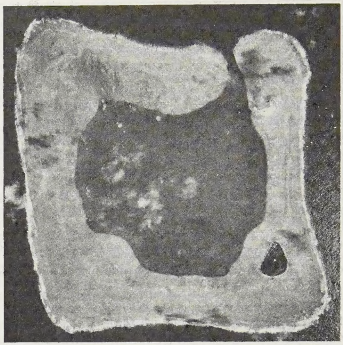


Figure 51: Lagoon in the middle of the Rose Atoll, American Samoa.
Air Survey Hawaii

The clarity of the water of the coral lagoon contrasts with the considerable amount of fine sediment in barrier island lagoon water. The extensive growth of salt marsh in such lagoons often detracts from or precludes their use as harbours, but where the water is deep enough these lagoons provide sheltered anchorage and good conditions for small-boat sailing in quiet water.

Nature of the lagoon environment. *Dimensions.* Coral lagoon dimensions range from small atolls to those so wide that the coral reefs on the far side cannot be seen across the lagoon. Atoll widths range from about 2.5 to nearly 100 kilometres, but the mean value is about 20 kilometres. Depths rarely exceed 60 metres and many are shallower, usually less than 20 metres deep. The lagoon of Mayotte island in the Comoro archipelago in the Indian Ocean attains a maximum depth of about 92 metres but generally is shallower, and it is about 16 kilometres in width at its widest point; this lagoon lies behind a barrier reef that encircles the island, forming an atoll about 55 kilometres in diameter.

Barrier island lagoons are usually elongated, though irregular ones may occur where river estuaries flood behind barriers. This occurs on the east coast of the United States, where lagoons extend intermittently for nearly 1,500 kilometres along the coast. The Gippsland Lakes in Victoria, Australia, exemplify a complex lagoon system formed behind a 149-kilometre beach. Elongated lagoons up to 64 kilometres in length lie behind the beach barrier, and larger lagoons, such as Lake Wellington, lie behind the southwestern end. Postglacial subsidence has flooded the lowland in this area. The lagoons are shallow: Lake Wellington is less than 3.5 metres in depth, and much of Lake King is less than 6 metres deep. Scour holes as deep as 16.5 metres do occur, however. The elongated lagoon behind the barrier is only 1 to 1.5 metres deep, typical of barrier island lagoons.

Water circulation. The degree of water circulation depends on the width of the tidal inlets, the tidal range, and the amount of runoff from adjacent land areas. Maximum velocities are attained at the points where the water passes through the barriers. In the entrance to the Gippsland lagoons, for example, tidal currents reach 5.6 kilometres per hour, but river floods that escape to the ocean can raise the velocity to 13 kilometres per hour. Water may be blown into the lagoon by strong winds; the increased level results in an outflowing current when the wind drops. Seiches can be created in this way. Small waves can be generated within lagoons when the wind blows along their maximum dimension. These may reach 1.25 metres in height and 1.5 to 9 metres in length in the Gippsland lagoons. In coral atoll lagoons there is little or no runoff, and seawater moves in and out through the passes, where

Current velocities

tidal currents reach their maxima. Velocities of 19 to 22 kilometres per hour have been recorded in the Hao Channel of the Tuamotu Archipelago.

Water temperature and salinity. In the Mayotte Lagoon the water has the same properties as the upper layers of the open ocean. The salinity is close to 35 parts per thousand, and the temperature varies between 27° and 24° C. This is typical of coral lagoons, but the temperature and salinity of barrier island lagoons are more variable because of their wider climatic range. Because they are shallow, lagoon waters approximate the air temperature: colder than the open ocean in winter, warmer in summer.

Brackish, marine, or hypersaline waters

Salinities decrease as a function of the amount of runoff entering the lagoon in relation to the tidal influx. Lagoons may be considered brackish, marine, or hypersaline. Brackish lagoons receive much runoff, and salinity increases toward the tidal inlets. The Gippsland lagoons exemplify this type. The salinity at the inner end varies from 0.5 to 5 parts per thousand according to season, and central values vary between 5 and 20 parts per thousand. Hypersaline lagoons occur where evaporation exceeds inflow. Laguna Madre in Texas and Sivash Sound in the Black Sea have salinities of 65 and 132 parts per thousand, respectively. Salt deposits may form in these conditions. The denser saline water tends to lie beneath the fresher water where mixing is not severe.

Equilibrium bottom profiles. Lagoons behind coastal barriers normally are zones of fine sedimentation. Their bottom profiles, therefore, build up gradually with time. Typical depths of the Texas lagoons vary between 1.25 and 3.5 metres, and their floors are flat. Early theories that attempted to relate the form of the offshore and lagoon profile are no longer held; and, because the lagoon profile changes with continued deposition, it cannot be used to establish the process of lagoon formation. The profile is usually gently undulating, but it may contain deeper channels, especially near the tidal inlets. Profiles across coral lagoons either are smooth and flat from calcareous sedimentation or contain knolls of growing or dead coral. There are 2,300 such knobs in the Enewetak Lagoon in the Marshall Islands.

Factors involved in lagoon formation. The essential feature that causes the lagoon to exist is the barrier that separates it from the ocean. In the coral lagoon the formation of the reef depends on the existence of suitable conditions for reef growth, which have already been mentioned briefly.

Barrier bars and sediment sources. The barrier island lagoons, on the other hand, depend not on organic processes but on waves, which act in a suitable environment on an adequate supply of bottom material, most commonly sand. Barrier islands are formed in those areas where the coastal slope is flatter than the equilibrium slope required by the long constructive swells—*i.e.,* the waves that build up the foreshore in front of their breakpoint. They are, therefore, found on low coasts. They may occur in areas of subsidence, stability, or emergence, wherever sufficient sand exists.

Lagoonal sediments and depositional rates

The barrier islands that form the lagoons are made of sand, but the sediments of the lagoon are usually finer, as conditions are quieter. The lagoonal muds differ from shelf muds. Glauconite is rare in lagoon muds, but oyster reefs may occur as along the Texas coast. The muds are found away from the channels, in which only coarse sediment can rest, owing to strong currents. Flocculation in the saline lagoon water expedites clay deposition. The source of the fine sediment is from inland areas, and transport is by rivers. The details of lagoon sedimentation vary with the nature of the river load. Sedimentation rates are much greater in the lagoon than the adjacent open ocean, because deposition is continuous over much of the lagoon and is often aided by flocculation and vegetation. In the Texas lagoons from 1875 to 1936, 20 centimetres of deposition occurred in spite of 30 centimetres of subsidence; the sedimentation rate, therefore, was about one metre in 100 years.

Waves, tides, and surf. The barrier islands are formed by the waves, which build up their equilibrium profile by deposition on a gradient that is too flat. The level of the growing accumulation may be raised by the wind, forming dunes. Where the land behind the growing barrier is low, it will become flooded to form a lagoon if sea level rises slowly. Such a rise of sea level has occurred during the past 20,000 years. As long as the barrier island can maintain its level above the sea, as sea level rises, the lagoon will exist until it is filled with sediment. Not all lagoons and barrier island complexes have formed during rising sea-level conditions, but, where there is evidence that no open-sea foreshore sediments lie on the land side of the barrier, this hypothesis seems likely. In some barriers, however, outbuilding of material by glacial outwash streams or rivers may provide a suitably low gradient and enough sediment to form a barrier, as along the south coast of Iceland. In other areas material carried alongshore to form a spit may develop into a bay-mouth barrier, enclosing a lagoon. Such features can be of sand or shingle. The Fleet, a brackish body of water behind Chesil Beach in southern England, is an example of the latter type.

Waves within the lagoon may develop cuspate spits along the land side of the barrier and the inner edge of the lagoon. These features may eventually break the lagoon into almost circular or oval water bodies. Examples occur in the Chukchi Sea lagoons in northeastern Russia and elsewhere where vegetation does not form marshland.

Storms and catastrophic events. Storms or tsunamis exert an effect on lagoons when they breach or overtop the barriers around a lagoon by raising the water level temporarily. Major changes in configuration can occur in a short time. Hurricanes, for instance, can cause serious effects on the coast of Texas, along which lagoons are common. Padre Island was lowered to below mean low tide at its southern end by one storm in 1919, and several washover channels were cut. The mainland shore also suffered erosion. Deposition may also occur; saline marls have been laid down on freshwater marsh, and small beach ridges may be built inside the lagoon where the high water level drives sand inland over mud. Coral reefs are more resistant to storms than are mobile sandy barriers.

The effect of time. Lagoons of both types change with time. In both, a relative rise of sea level with time is important in the development of the lagoon. In coral atolls there is evidence from deep boring that Darwin's original subsidence hypothesis of atoll formation, via barrier reefs from fringing reefs around a subsiding volcanic peak, is substantially correct in many cases. As long as the coral can maintain its growth at a suitable level as its foundation subsides, the atoll will continue to enclose a lagoon, which is floored by coral or calcareous sediment derived from the reef and which maintains its depth by growth or deposition.

The postglacial rise of sea level also has influenced barrier island formation in many instances. When sea level rises too fast, a barrier may be drowned and its lagoon will cease to exist. (C.A.M.K./M.L.S.)

ESTUARIES

Estuaries are partially enclosed bodies of coastal water that have access to the sea and that receive a freshwater contribution from the land. Physically this would include river mouths (as in the case of the Mississippi River) or coastal plain drainage areas of low relief (*e.g.,* Chesapeake Bay) that are flooded by a rise in sea level. Other examples would be submerged fjords (Scoresby Sound, Greenland) and structural basins (San Francisco Bay) and the bodies of water behind spits (Hurst Castle spit, Eng.) or barrier beaches (Ninety Mile Beach, Australia). In the case of spits and barrier beaches, the definitions of lagoons and estuaries overlap (see above *Lagoons* and especially the article RIVERS: *Rivers as agents of landscape evolution: Estuaries*).

Basic types of estuaries

Classification of estuaries is made on the basis of salinity distribution. Though the terminology differs slightly among different specialists on the subject, there are several basic types. They are as follows: (1) vertically mixed, wherein salinity, while constant from top to bottom at any site, increases from land to sea; (2) slightly stratified, in which saline water circulates in at the bottom, mixes with fresh water, and then flows out at the top (salinity thus increases with depth and out toward the sea); (3) highly

stratified, which is similar to the slightly stratified type, but is limited to the upper layer of water above the outer sill of a fjord; and (4) salt wedge, where saline seawater intrudes in as a wedge at the bottom, while fresh water flows out and over it at the top. (M.L.S.)

GULFS AND BAYS

Any concavity of a coastline or reentrant of the sea, regardless of size, depth, configuration, and geologic structure, may be called a gulf or bay. The nomenclature for features of this type is far from uniform; names that may refer to sizable gulfs and bays in various places include bight, firth, sound, and fjord. A number of pronounced concavities of oceanic margins have no proper name at all.

The problem of terminology extends to the difference between gulfs and seas. There are many small seas, such as the Sea of Marmara (11,000 square kilometres) and the Sea of Azov (38,000 square kilometres), which, strictly speaking, are really gulfs of the ocean or other seas (the Sea of Azov is a gulf of the Black Sea). The Gulf of Aden (about 270,000 square kilometres), another example, is part of the Arabian Sea, and these water bodies have a common regime (similar tides, precipitation, evaporation, and so forth). The narrow sound of Bab el-Mandeb connects the gulf with the vast Red Sea (438,100 square kilometres) and exhibits a number of specific geomorphic features. The Red Sea, in turn, has two small gulfs to the north—namely, those of Suez and Aqaba.

The Bay of Bengal and the Arabian Sea are approximately the same size and have the same monsoonal water circulation. The Bay of Bengal is, in fact, the largest of the gulfs and bays, with a surface area of 2,172,000 square kilometres and a length of 1,850 kilometres (see Table 10). The width of a gulf may exceed its length. The Great Australian Bight has the widest mouth (2,800 kilometres). The Gulf of Guinea is the deepest; its maximum depth (6,363 metres) exceeds that of the Bay of Bengal by more than 1,000 metres.

The shape and bottom topography of gulfs and bays are amazingly diverse. They are determined by the geologic structure and development of the region. Homogeneous bedrock of low strength or resistance results in simple shapes and shallow depths. The Gulf of Riga (at the Baltic Sea) is a possible example of the type. Long narrow arms with approximately parallel shores of the south Kara Sea extend inland for about 800 kilometres. They occupy troughs that originated by erosion during a period of lower sea level (Baidaratskaya Bay, Obskaya Bay with Tazovskaya Bay tributary, Yenisey Bay, Gydanskaya Bay). Deep, angular gulfs, on the other hand, are created along fractures, faults, and rifts (*e.g.*, Varanger Fjord); they usually have irregular bottom topography. Parallel fractures form extremely deep, narrow gulfs with parallel shores, such as the Gulf of California. Genuine fjord-gulfs are notable for their very high length to width ratios (up to 50:1). In regions that have undergone nonuniform deformation and uplift, gulfs and bays of complicated and irregular shape and bottom topography are consequently formed; the Gulf of St. Lawrence is an example.

Gulfs are connected with the sea by means of one or more straits. Sometimes there may be an archipelago in the mouth of the gulf, as in the Gulf of Bothnia. There are some gulfs that open into the sea or into another gulf on opposite sides (Baffin Bay, Gulf of Aden, and the Gulf of Oman).

Single gulfs usually are formed along linear shores of the continent. If the shoreline is irregular and has a complex geologic structure, groups of gulfs of a similar nature may occur. Most shorelines have small reentrants of various size that are called bays. These features are strongly influenced by local conditions, and they are not described or classified within the context of this section, which treats major water bodies of the world. For additional information on the dynamics of water within gulfs and bays, see above *Waves of the sea*.

Factors that affect the characteristics of gulfs and bays. These bodies of water may differ from the adjacent ocean (or sea) by virtue of water properties and dynamics and processes of sedimentation. Such differences are determined by the size and the shape of a given gulf, by the depth and bottom topography, and, to a considerable extent, by the degree of isolation from the ocean. Climatic conditions also are important. Isolation from an adjacent ocean depends on the ratio of width of mouth to total surface area of a gulf or on the cross section of the mouth to total water volume. If there is a sill (a submarine ridge or rise), the ratio of depth above the sill to the depth of the gulf is of great importance. No extensive comparisons of these ratios have been made to date; hence any analysis of controlling variables must remain somewhat qualitative.

Climatic conditions, sills, and river runoff

A high sill hampers the water exchange between an ocean and gulf and may lead to stagnation (oxygen deficiency) as is found in some fjords of Norway, in the Red Sea, and, particularly, in the Black Sea. Also, the presence of a sill causes independent circulation of gulf waters, generated by local winds and the runoff of rivers. Sills are not indispensable for the formation of an independent circulation, however. A narrow mouth, as in the Gulf of Bothnia, leads to the same result.

In humid climates, the waters of gulfs are freshened by river runoff. Salinity is particularly low in the gulfs of the Baltic Sea and along the southern coast of the Kara Sea. Water becomes almost fresh in their heads, especially in the spring when snow begins to thaw. Gulfs of the arid zone suffer from intensive evaporation and receive little river runoff. Thus, salinity increases markedly in this climatic regime—up to 60 parts per thousand in the Persian Gulf and up to 350 parts per thousand in the Kara-Bogaz-Gol (a gulf of the Caspian Sea). In addition to its effect on salinity, river runoff delivers organic matter and nutrient salts that may determine the specific features of life in the gulfs. The number of genera and species of organisms is small, but the organisms present tend to develop in quantities. That is why shrimp, oyster, and other fisheries are concentrated in many gulfs.

Funnel-shaped gulfs, in which the depth gradually decreases headward, usually have resonant tides. The tidal range at the head of such gulfs is several times greater than that in the open ocean (*e.g.*, Bristol Channel, Río de la Plata, Mezenskaya Bay, Shelikhova Gulf). The world maximum tidal range has been registered in the Bay of Fundy (18 metres). The regularity (magnitude and frequency) of the flood tide may be distorted in such instances, and the duration of the flood tide may become much shorter than that of the ebb tide. This may cause the phenomenon of tidal bore, in which a steep wave will move rapidly upstream for dozens of kilometres.

Occurrence of seiches

Gulfs of simple shape with a narrow mouth and a high degree of isolation from the ocean are often subject to seiches. These free oscillations can result from rapid changes of atmospheric pressure and, of course, from tectonic movements such as earthquakes. Seiches gradually decrease, but some oscillation continues long after their cause disappears. A high rise of the water (storm surge) occurs in long and shallow gulfs if winds from the sea are protracted. Such phenomena are difficult to predict, and the high water levels may cause floods. Seiches commonly occur at the heads of Helgoländer Bay in the North Sea and in the Gulf of Finland.

Certain aspects of sedimentation are affected by the isolation of gulfs from the ocean and river runoff. The rate of sediment accumulation in gulfs of limited area may be very high. This, of course, is a function of river discharge; sediment composition is usually similar to that of the load transported by entering rivers. Deposition of calcium carbonate often occurs in shallow gulfs in the arid zones where few if any perennial streams exist. The bottoms of long gulfs (or gulfs having sills) are usually covered with silt even at the shallowest depths (*e.g.*, Hudson Bay, the Po Hai, the inlets or *gubas* of the Kara Sea, the Gulf of Riga). Only strong tidal currents can prevent this siltation and, in some cases, cause the opposite phenomenon of bottom erosion. Currents maintain the existence of or actively deepen bottom troughs in narrow-mouthed gulfs whose depths are over 200 metres, whereas depths of adjacent parts of the open ocean are only on the order of some dozens of metres.

Waves of the open ocean either do not penetrate into

comparatively isolated gulfs or, if they do, they become greatly reduced after entry. Small local waves that are related to gulf size prevail there. This tends to make gulfs quite navigable, and seaports and harbours have generally been situated on them.

Classification of gulfs and bays. The geologic structure and developmental history of gulfs and bays are as varied as are those of the continents or oceans proper. The factors discussed above influence the morphological peculiarities of gulfs, and the latter, in turn, permit some general division or classification of these features to be made. The several groups in one possible scheme are discussed here using typical gulfs and bays of each group as examples (Table 10).

Open concavities and isolated areas

Areas situated in open concavities of the continental coast (Gulf of Alaska, Bay of Biscay, Gulf of Guinea, Great Australian Bight, Bay of Bengal, Gulf of Tehuantepec, for example) are classified as the A_1 group. The depth of these gulfs in the region of the mouth usually is on the order of kilometres. The continental shelf and slope are generally pronounced. The general shape of such gulfs is simple; width of mouth usually exceeds its length. Water circulation and its physical properties are similar to those of the oceans. The character of the marine faunas does not differ from that of oceanic areas.

Large areas considerably isolated from oceans, such as the Gulf of Mexico and Baffin Bay, are designated as group A_2 (Table 10). The former includes a geosynclinal hollow, founded in Mesozoic time (from 245 to 66.4 million years ago) and finally shaped during the Tertiary Period (from 66.4 to 1.6 million years ago). It is connected with the ocean by the narrow and relatively shallow straits of Florida and Yucatán. Baffin Bay is a rift hollow that is connected by straits with the Atlantic.

Ocean gulfs, such as the gulfs of Oman, California, Aden, and some others, have smaller areas and are isolated to a lesser degree. These gulfs and bays, in group A_3, have shapes that are determined by young faults and fractures. Depths in these gulfs generally exceed one kilometre. Unlike the previous group, in which gulfs might be of composite geologic structure, these occupy areas that have undergone only a single episode of deformation.

Gulfs situated on the continental shelf, such as the Bay of Fundy, Hudson Bay, Río de la Plata, San Matías Gulf (off Argentina), and others, are in group B. The depth of such gulfs is up to 200 metres or more, and their configuration is determined by geologic conditions. Because shelf areas repeatedly became dry land when the sea level fell during the ice ages, these gulfs received their final shape during the Pleistocene Epoch. The Gulf of St. Lawrence is included in this group (Table 10), though it is really intermediate between groups A_3 and B. It contains both a pronounced shelf and a long trough up to 530 metres deep.

Marginal and shelf areas

Gulfs of intercontinental and marginal seas are considered to be a third category. These may be divided into group C_1, which consists of gulfs of basin seas, including the deepwater part only (Aqaba) or both the deepwater and the shelf parts (the Gulf of Honduras), and group C_2, the shelf gulfs of the same seas (*e.g.*, the Persian Gulf, Gulf of Suez, Anadyrsky Gulf, Bristol and Norton channels, Shelikhova Gulf).

Finally, there are the gulfs of the shelf seas (*gubas* of the Arctic seas of Russia, gulfs of the Baltic and the White Seas, the Gulf of Carpentaria, the Po Hai, and many others), which are placed in group D (Table 10). The shallow character of the shelf seas influences the water dynamics of the gulfs. Water exchange is weakened, and sediments may accumulate in the gulf mouths, thus forming submarine barriers and further reducing exchange.

(V.P.Z./M.L.S.)

Economic aspects of the oceans

The sea is generally accepted by scientists as the place where life began on Earth. Without the sea, life as it is known today could not exist. Among other functions, it acts as a great heat reservoir, leveling the temperature extremes that would otherwise prevail over the Earth and expand the desert areas. The oceans provide the least expensive form of transportation known, and the coasts serve as a major recreational site. More importantly, the sea is a valuable source of food and a potentially important source of energy and minerals, all of which are required in ever-increasing quantities by industrialized and developing nations alike.

MEDIUM FOR TRANSPORTATION AND COMMUNICATIONS

From the beginning of recorded history, people have used the sea as a means of transporting themselves and their goods. The bulk of the tonnage of products transported throughout the world today continues to be moved in ocean vessels. The size of these vessels ranges from small boats capable of carrying a few tons to bulk carriers (*e.g.*, supertankers) capable of transporting more than 500,000 tons of oil. The cost of transporting goods on the ocean depends on the product, the form of shipment, and the type of vessel. As the per capita consumption of materials increases, the outlook for marine transportation is one of ever-increasing tonnages and size of carrying vessels.

Since the laying of the transatlantic cable in the 19th century, the oceans have served as a major means of communication between continents and islands. Hundreds of seafloor cables connect many large centres of world population. With the development of satellite communications, seafloor cables as a means of communication have decreased somewhat in importance, but they will continue to carry information for many decades to come.

In addition to communications, cable and pipes laid on the seafloor carry electrical energy, oil, and other commodities in many parts of the world.

SOURCE OF FOOD AND WATER

Fishing. Many millions of tons of edible fish and shellfish are taken from the oceans each year. Yet, the food-producing potential of the sea is considerably greater than this. The prevailing methods by which fish are taken from the oceans are inefficient. This problem is compounded by the fact that only a handful of varieties are targeted by commercial fishermen (anchovy, sardine, herring, cod, mackerel, and pollack constitute more than half of the total annual catch). Overfishing of certain varieties and areas along the continental shelves also has resulted in declining numbers and therefore catches. The decimation of the whale (many species are near extinction) and also of the anchovy off the waters of Peru exemplifies the consequences of overfishing.

Several alternatives have been proposed to enhance productivity. One possibility is to extract protein concentrates from all types of fish, promoting the use of varieties formerly ignored for foodstuff. It has been estimated that this procedure could produce a sustained yield of two billion tons of food annually for the world's populace. Another alternative would be to use protein derived from algae cultivated on the continental shelf. A more viable course of action would be to develop the continental shelf for fish- and kelp-farming. The Japanese have instituted a substantial research program in this area. They have farmed oysters in their oceanic bays for many years and more recently raised sea bream and shrimp in protected environments. Commercial shrimp and oyster farms also have been developed in the United States in the shallow waters of estuaries and bays. Both commercial and research aquaculture projects have been conducted to raise abalone, Maine lobster, salmon, and edible forms of seaweed under controlled conditions. Limited sea farming has been practiced in France and Italy, both of which raise mussels and certain other shellfish.

Fish-farming

Various marine organisms have been artificially cultivated as sources for therapeutic drugs. For example, during the late 1980s American researchers set out to grow microalgae that have a high concentration of a substance called beta-carotene, which is believed to prevent or retard the spread of certain forms of cancer.

Desalination. The need for water for domestic and agricultural purposes has grown steadily throughout the latter half of the 20th century. As a result, increasing attention has been given to desalting ocean waters and brackish waters in inland seas. Throughout the world, more than

Table 10: Physical-Geographic Features of Some Gulfs and Bays*

names of gulfs and bays	surface area (10^6 sq km)	volume (10^6 cu km)	length in kilometres	width in kilometres		depth in metres			tidal range in metres	surface water temperature (Celsius)		surface salinity (parts per thousand)		river runoff
				max.	mouth	max.	mean	sill		max.	min.	max.	min.	
A_1 Group														
Gulf of Alaska	1.327	3.226	325	1,650	1,650	5,659	2,431	none	12.0†	12	<0	33	32	small
Bay of Bengal	2.172	5.616	1,850	1,720	1,720	5,258	2,586	none	10.7	27	25	34	18	large
Bay of Biscay	0.194	0.332	400	500	500	5,120	1,715	none	6.7	20	5	35.5	34	medium
Gulf of Guinea	1.533	4.592	540	1,900	1,900	6,363	2,996	none	2.7	27	25	35	31	large
A_2 Group														
Baffin Bay	0.689	0.593	>1,000	600	340‡	>2,300	861	466‡	4.2	5	<0	33.5	30	none
Gulf of Mexico	1.543	2.332	1,330	1,780	445	4,029	1,512	800	1.7	29	17	36.7	33	large
A_3 Group														
Gulf of Aden	...	...	900	335	335	3,328	...	none	2.9	>30	25	36.5	36	none
Gulf of California	0.177	0.132	1,200	200	200	3,660	813	none	5.2§	30	16	35.5	35	medium
Gulf of Oman	...	...	450	330	325	3,474	...	none	3.5	32	22	38	37	none
B Group														
Bay of Fundy	...	...	300‖	100	100	214	75	none	18.0	17	2	32	30	medium
Hudson Bay	0.819	0.092	1,560	1,140	190	274	112	none	7.9	14	<0	28.5	23	large
Río de la Plata	...	...	220	95	95	10	5–7	6	1.0	21	11	33	20	large
Gulf of St. Lawrence	0.238	0.030	¶	¶	¶	530	127	none	5.9	20	−1.8	32	26	large
C_1 Group														
Gulf of Aqaba	...	...	180	28	6	1,828	...	462	0.7#	26	24	42	41	none
Sirt Gulf	...	...	200	450	450	1,627	...	none	0.3	27	14	38	38	none
C_2 Group														
Anadyrsky Gulf	...	...	350	460	460	110	60–70	none	3.0	10	<0	30	28	medium
Persian Gulf	0.241	0.010	1,000	350	56	170	40	71	4.7	33	15	60△	30	small
Gulf of Suez	...	...	325	58	58	82	40–60	none	1.8	28	23	43	41	none
Shelikhova Gulf	...	...	750	300	190	495	100–150	none	12.9	14	<0	33	31	small
Gulf of Thailand	...	...	830	550	370	83	45.5	58	0.8▽	31	27	32.5	30.5	large
D Group														
Gulf of Bothnia	0.117	...	668	240	155◇	294	21	none	0.6	14	0	5.5	1	medium
Po Hai	0.0827	0.0017	480▲	285	105	38	15–20	none	4.4	28	<0	31	22	large
Gulf of Carpentaria	0.4116	...	675	650	530	70	40–50	none	3.6	29	23	35.5	35	small
Mezenskaya Bay	...	...	105	97	97	31	10–20	none	10.0	16	<0	32	15	large
Obskaya Bay	...	...	800	90	60	18	10–12	7	0.7	14	0	15	1	large
Gulf of Finland	0.030	...	420	125	70	110	50–60	86	0.1	17	0	5	2	medium

*Data in this table may differ from those given elsewhere in set for some features because of differing definitions of geographic limits of each feature. Data adapted in part from Kossina (1921, 1933) and Lyman (1961). †Cook Inlet's head. ‡Davis Strait. §Up to 10 metres at the Colorado River mouth. ‖Up to the Minas Bay head. ¶Not given because of the complicated outlines. #The value for the Red Sea's head. △In shallow parts along the south coast. ▽Up to 4 metres at the Mae Nam Chao Phraya. ◇South Quarken Strait—40 kilometres. ▲From the Laichow Wan head to Liao-tung Wan's head.

3,500 land-based desalination plants, producing a total of more than eight billion litres (in excess of two billion gallons) per day, were in operation by the early 1990s. In general, the desalination plants are located in areas where the population has outstripped the onshore water supply and where high-cost desalinated water can be afforded. This situation tends to arise in coastal-desert areas or on densely populated islands, because the cost of pumping water through pipelines to interior areas would add prohibitively to the basic cost at the sites of desalination. The United States operates a little more than 30 percent of the world's desalination facilities. Another 20 percent or so are found in the Middle East, chiefly in Israel, Saudi Arabia, Kuwait, and the United Arab Emirates. Together, these Middle Eastern plants produce roughly two-thirds of the world's desalinated water.

A population usually can afford to pay about 10 times as much for water for domestic purposes as it does for agricultural water. Large-scale nuclear desalination facilities promise to lower the cost of desalted water at the desalination sites to a level that most industries and some agricultural enterprises can afford. As yet, however, no such plant has been constructed.

Methods of choice

At present desalination is accomplished primarily by distillation and membrane processes. Distillation processes involve some form of evaporation and subsequent condensation. Many of the largest commercial desalination plants use multiple-effect distillation (*i.e.*, multistage-flash distillation). Roughly half of the world's desalting facilities employ distillation processes and account for approximately 75 percent of all desalinated water produced annually.

Membrane processes for desalting include reverse osmosis and electrodialysis. Of the two, reverse osmosis is the more widely used, particularly for desalting brackish waters from inland seas. In this method the natural process of osmosis is reversed by applying pressure to brine that is in contact with an osmotic membrane. The membrane impedes the passage of salt ions while allowing fresh water to move through. In electrodialysis an electric potential is used to drive positive and negative ions of the dissolved salts through membranous filters, thereby sharply reducing the salt content of the water between the filters.

In the future, it can be expected that the ocean will become an increasingly important source of fresh water. If production and transportation costs can be lowered sufficiently, it may be possible to produce fresh water to irrigate large areas that border the oceans in many parts of the world.

ENERGY RESOURCES

There are several recognized techniques by which energy can be extracted from the sea. The major problem in taking energy resources from the sea is that they tend to be diffused over a large lateral area. A point-concentration energy source is necessary if it is to be exploited economically.

Tidal power generation. Hydraulic turbine-generator units are presently used to extract energy from ocean tides, although on a very limited scale. As of the late 1980s, operating tidal units included a 240-megawatt plant in France on the estuary of the Rance River and several smaller installations, as, for example, a 40,000-kilowatt pilot plant in Russia on the Barents Sea. There are few sites throughout the world that are suitable for harnessing tidal energy without constructing prohibitively expensive damlike structures. (These are required to trap water at high tide and release it at low tide to turn the hydraulic turbine; see ENERGY CONVERSION: *Major energy-conversion devices and systems: Turbines: Water turbines.*) Thus, it seems unlikely that tidal power will play a more significant role in the coming years.

Utilization of temperature differentials

Ocean thermal energy conversion. Another more promising technology, known as ocean thermal energy conversion (OTEC), makes use of the temperature differential between the warm surface waters of the oceans, heated by solar radiation, and the deeper cold waters to generate power in a conventional heat engine. The difference in temperature between the surface and lower water

layer can be as large as 50° C over vertical distances of as little as 90 metres in some ocean areas. To be economically practical, the temperature differential should be at least 20° C in the first 1,000 metres below the surface.

The OTEC concept was first proposed in the early 1880s by the French engineer Jacques-Arsène d'Arsonval. His idea called for a closed-cycle system, a design that has been adapted for most present-day OTEC pilot plants. Such a system employs a secondary working fluid (a refrigerant) such as ammonia. Heat transferred from the warm surface ocean water causes the working fluid to vaporize through a heat exchanger. The vapour then expands under moderate pressures, turning a turbine connected to a generator and thereby producing electricity. Cold seawater pumped up from the ocean depths to a second heat exchanger provides a surface cool enough to cause the vapour to condense. The working fluid remains within the closed system, vaporizing and reliquefying continuously.

Some researchers have centred their attention on an open-cycle OTEC system that employs water vapour as the working fluid and dispenses with the use of a refrigerant. In this kind of system warm surface seawater is partially vaporized as it is injected into a near vacuum. The resultant steam is expanded through a low-pressure steam turbogenerator to produce electric power. Cold seawater is used to condense the steam, and a vacuum pump maintains the proper system pressure.

During the 1970s and '80s the United States, Japan, and several other countries began experimenting with OTEC systems in an effort to develop a viable renewable energy source. In 1979 American researchers put into operation the first OTEC plant able to generate usable amounts of electric power—about 15 kilowatts of net power. This unit, called Mini-OTEC, was a closed-cycle system mounted on a U.S. Navy barge a few kilometres off the coast of Hawaii. In 1981–82 Japanese companies tested another experimental closed-cycle OTEC plant. Located on the Pacific Island republic of Nauru, this facility produced 35 kilowatts of net power. Since that time researchers have continued developmental work to improve heat exchangers and to devise ways of reducing corrosion of system hardware by seawater.

The prospects for the commercial application of OTEC technology seem bright, particularly on islands and in developing nations in the tropical regions where conditions are most favourable for OTEC plant operation. It has been estimated that the tropical ocean waters absorb solar radiation equivalent in heat content to that of about 170 billion barrels of oil each day. Removal of this much heat from the ocean would not significantly alter its temperature, but would permit the generation of about 10 million megawatts of electricity on a continuous basis.

SOURCE OF MINERALS AND OTHER RAW MATERIALS

Petroleum. In the mid-1950s, the production of oil and gas from oceanic areas was negligible. By the early 1980s, about 14 million barrels per day, or about 25 percent of the world's production, came from offshore wells, and the amount continues to grow. More than 500 offshore drilling and production rigs were at work by the late 1980s at more than 200 offshore locations throughout the world drilling, completing, and maintaining offshore oil wells (see FUELS, FOSSIL). Estimates have placed the potential offshore oil resources at about 2 trillion barrels, or about half of the presently known onshore potential oil sources.

It was once thought that only the continental-shelf areas contained potential petroleum resources, but discoveries of oil deposits in deeper waters of the Gulf of Mexico (about 3,000 to 4,000 metres) have changed that view. It is now believed that the continental slopes and neighbouring seafloor areas contain large oil deposits, thus enhancing potential petroleum reserves of the ocean bottom.

Offshore drilling is not without its drawbacks, however. Not only is it difficult and expensive to drill on the continental shelf and in deeper water, but there is also the risk of accidental discharges of oil that can cause serious damage to the environment and marine organisms. In spite of technological advances, inadvertent leaks, sometimes on a large scale, continue to occur.

Minerals. The rivers of the world dump billions of tons of material into the oceans each year. Seafloor springs and volcanic eruptions also add many millions of tons of elements. Even the winds contribute solid materials to the oceans in appreciable quantities. Most of these sediments rapidly settle to the seafloor in nearshore areas, in some cases forming potentially valuable placer mineral deposits. The dissolved load of the rivers, however, mixes with seawater and is gradually dispersed over the total oceanic envelope of the Earth. Because of the nature of the minerals and their mode of formation, it is convenient to consider the occurrence of ocean deposits in several environments—marine beaches, seawater, continental shelves, sub-seafloor consolidated rocks, and the marine sediments of the deep-sea floor. Minerals are mined from all of these environments except for the deep-sea floor, which was only recently recognized as a repository for mineral deposits of unbelievable extent and significant economic value.

Placer and near-shore deposits

Minerals that resist the chemical and mechanical processes of erosion in nature and that possess a density greater than that of the Earth's common minerals have a tendency to concentrate in gravity deposits known as placers. During the Pleistocene glaciations, sea level was appreciably lowered as the ocean water was transferred to the continental glaciers. Because of the cyclical nature of the ice ages and the intervening warm periods, a series of beaches were formed in nearshore areas both above and below present sea level. Also, when sea level was lowered in past ages, the streams that today flow into the sea coursed much further seaward, carrying placer minerals to be deposited in channels that are now submerged. With geophysical exploration techniques, these channels and beaches can be easily delineated, even though these features are totally covered by Holocene sediments—*i.e.,* those deposited during the past 10,000 years.

Sand and gravel are mined from a number of offshore locations around the world, generally with hydraulic dredges. They are used primarily for construction purposes or for beach replenishment or nearshore fills.

Sulfur, which is taken from salt domes in the Gulf of Mexico, is mined by a process in which pressurized hot water is pumped into the sulfur-containing cap of the dome, melting the sulfur and forcing it to the surface. Compressed air is also used to pump sulfur to the surface; the still-molten sulfur is then conveyed to the shore through insulated pipelines.

Of considerable interest are the seafloor phosphorite deposits on the coastal shelves of many nations. The phosphorite off California occurs as nodules that vary in shape from flat slabs a few metres across to small spherical forms termed oolites. The nodules commonly are found as a single layer at the surface of coarse-grained sediments. Phosphorite composition from the California offshore area is surprisingly uniform and contains potentially economically attractive amounts of phosphorus. Another type of phosphate deposit has been discovered off the west coast of Mexico. It contains as much as 40 percent apatite (common phosphate mineral), and some experts have speculated that up to 20 billion tons of recoverable phosphate rock exist in the deposit.

Deep-sea deposits

Mineral deposits of enormous size and potential economic significance have been discovered on the deep ocean floor. Minerals formed in the deep sea are frequently found in high concentrations because there is relatively little clastic material generated in these areas to dilute the chemical precipitates.

An estimated 10^{16} tons of calcareous oozes, formed by the deposition of calcareous shells and skeletons of planktonic organisms, cover some 130 million square kilometres of the ocean floor. In a few instances, these oozes, which occur within a few hundred kilometres of most nations bordering the sea, are almost pure calcium carbonate; however, they often show a composition similar to that of the limestones used in the manufacture of portland cement.

Covering about 39 million square kilometres of the ocean floor in great bands across the northern and southern ends of the Pacific Ocean and across the southern ends of the Indian and Atlantic oceans are other oozes, consisting of the siliceous shells and skeletons of plankton animals and

plants. Normally these oozes could serve in most of the applications for which diatomaceous earth is used, for fire and sound insulation, for lightweight concrete formulations, as filters, and as soil conditioners.

An estimated 10^{16} tons of red clay covers about 104 million square kilometres of the ocean floor. Although compositional analyses are not particularly exciting, red clay may possess some value as a raw material in the clay-products industries, or it may serve as a source of metals in the future. The average assay for alumina is about 15 percent, but red clays from specific locations have assayed as high as 25 percent alumina; copper contents as high as 0.20 percent also have been found. A few hundredths of a percent of such metals as nickel and cobalt and a percent or so of manganese also are generally present in a micronodular fraction of the clays and in all likelihood can be separated and concentrated from the other materials by a screening process or by some other physical method.

Underlying the hot brines in the Red Sea are basins containing metal-rich sediments that potentially may prove to be of considerable significance. It has been estimated that the largest of several such pools, the Atlantis II Deep, contains several billion dollars worth of copper, zinc, silver, and gold in relatively high grades. These pools lie in about 2,000 metres of water midway between The Sudan and the Arabian Peninsula. Because of their gellike nature, pumping these sediments to the surface may prove relatively uncomplicated. These deposits are forming today under present geochemical conditions and are similar in character to certain major ore deposits on land.

The discovery, in 1978, of polymetallic sulfides at the mid-ocean spreading centres has aroused much interest. As noted elsewhere in the article, these sulfides include sediments enriched in iron and manganese. Sites of rich deposits have been located at the Galápagos spreading centre in the Gulf of California and at the East Pacific Rise.

Manganese nodules

From an economic standpoint, the most interesting oceanic sediments are manganese nodules—small, black to brown, friable lumps found to be widely distributed throughout the major oceans in the late 19th century by the Challenger and Albatross expeditions.

Many theories have been proposed to account for the formation of manganese nodules, the best probably being that the ocean is saturated at its present state of acidity-alkalinity in iron and manganese. For this reason, these elements precipitate as colloidal particles that gradually increase in size and filter down to the seafloor. Colloids of manganese and iron oxides collect many metals and tend to agglomerate as nodules at the seafloor rather than settle as particles in the general sediments. An estimated 1.5 trillion tons of manganese nodules lie on the Pacific Ocean floor alone. Averaging about four centimetres in diameter and found in concentrations as high as 38,600 tons per square kilometre, these manganese nodules contain as much as 2.5 percent copper, 2.0 percent nickel, 0.2 percent cobalt, and 35 percent manganese. In some deposits, the content of cobalt and manganese is as high as 2.5 percent and 50 percent, respectively. Such concentrations would be considered high-grade ores if found on land, and, because of the large horizontal extent of the deposit, they are a potential source of many important industrial metals.

Relatively simple mechanical cable bucket or hydraulic dredges with submerged motors and pumps can effect the mining of the nodules at rates as high as 10,000 to 15,000 tons per day, from depths as great as 6,000 metres. The estimated costs of mining and processing the nodules indicate that copper, nickel, cobalt, and other metals can be economically produced from this source.

The prospects of mining the manganese nodules and metal-rich sediments have brought home the need to resolve long-standing legal problems relating to the ownership of marine resources. During the 18th century the extent of the territorial sea (and therefore rights) was established as 3 nautical miles (5.6 kilometres) from a nation's shoreline. The area beyond the territorial sea, the so-called high seas, was regarded as open to all nations. By the mid-1940s technological advances had extended offshore oil drilling beyond the territorial limit. This situation, together with the desire of various coastal nations to protect their fishing grounds, eventually resulted in an attempt to codify international law concerning territorial waters, ocean resources, and sea lanes. A 1982 treaty calling for the enactment of the United Nations Law of the Sea Convention was signed by 138 nations; some 30 other states, including the United States, the United Kingdom, and West Germany, refused to sign, however. The treaty extended the territorial limit of each coastal nation to a distance of 12 nautical miles and granted it sovereign rights over natural resources—living and nonliving—within an exclusive economic zone (or EEZ) of 200 nautical miles. The nations that refused to sign the treaty objected to its provisions governing seabed mining. The treaty declared the minerals on the seafloor beneath the high seas the "common heritage of mankind" and stipulated that their exploitation be directed by a global authority. While private and national mining concerns are allowed to conduct exploration and set up extraction operations, the question of seabed mineral ownership and mining rights remains largely unresolved. This situation is viewed in some quarters as the primary obstacle to full and effective utilization of seabed resources.

The Law of the Sea

WASTE DISPOSAL AND OTHER RELATED ACTIONS

One of the least known but most significant uses of the sea is as an enormous dump site. In the past, the oceans were able to assimilate the wastes of society without noticeable adverse effects. However, industrialization and other concomitant developments, along with sharp increases in global population, have given rise to quantities and forms of waste that are now taxing the capacity of the oceans to absorb them. Extensive marginal areas of the oceans have been heavily polluted by human wastes ranging from the raw sewage of urban centres to junked appliances and automobiles. Less apparent but more insidious forms of pollution are toxic chemicals, nuclear wastes, and oily bilges pumped by practically all vessels using petroleum for power (see also above *Chemical and physical properties of seawater: Composition of seawater: Effects of human activities*).

Some other human activities are equally harmful to the marine environment. Massive oil spills from tanker accidents, such as the 1989 mishap involving the *Exxon Valdez* in Prince William Sound, Alaska, have not only disfigured innumerable beaches and estuaries but caused widespread damage to wildlife as well (Figure 52). Large power plants are generally located along coastlines to reduce the costs involved in cooling their condensers by water-circulation systems. Although the whole of the ocean never will be affected by the waste heat dissipated by these plants, detrimental environmental effects can be caused in the immediate area of the power-plant outfall. Herbicides and pesticides (especially the organochlorides still used by

Oil spills

AP/Wide World Photos

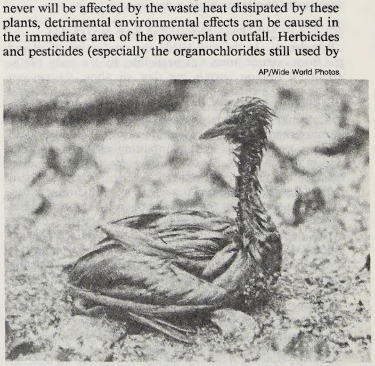


Figure 52: A red-necked grebe coated in crude oil after the tanker *Exxon Valdez* ran aground in Prince William Sound, Alaska, U.S.

some countries) reach the oceans via the wind and rivers and contaminate marine organisms.

The fringes of the oceans—the beaches, lagoons, and bays—are the most sensitive to human action, but the continued dumping of wastes, attended by other abuses, will eventually affect the entire marine environment.

(J.L.M./Ed.)

BIBLIOGRAPHY

General considerations: Broad overviews of the oceans are provided by KEITH STOWE, *Ocean Science,* 2nd ed. (1983); DAVID TOLMAZIN, *Elements of Dynamic Oceanography* (1985); DAVID A. ROSS, *Introduction to Oceanography,* 4th ed. (1988); ALYN C. DUXBURY and ALISON B. DUXBURY, *An Introduction to the World's Oceans,* 3rd ed. (1991); M. GRANT GROSS, *Oceanography, a View of the Earth,* 5th ed. (1990); and TOM BEER, *Environmental Oceanography: An Introduction to the Behaviour of Coastal Waters* (1983). ALASTAIR COUPER (ed.), *The Times Atlas and Encyclopedia of the Sea* (1989), provides a graphic look at all aspects of the ocean. *Ocean Yearbook* (annual), contains essays on resources, transportation, and marine science, among other topics. GUSTAF ARRHENIUS, BIBHAS R. DE, and HANNES ALFVÉN, "Origin of the Ocean," in *The Sea,* vol. 5, *Marine Chemistry,* ed. by EDWARD D. GOLDBERG (1974), pp. 839–861, provides a thorough discussion of the formation of water on the Earth during early geologic history. A. GUILCHER, "Continental Shelf and Slope (Continental Margins)," in *The Sea,* vol. 3, *The Earth Beneath the Sea: History,* ed. by M.N. HILL (1963), describes the distribution and features of the margins of the continents.

(A.C.D.)

Chemical and physical properties of seawater: As an excellent starting point for the reader interested in an integrated account of ocean chemistry, physics, and biology, the classic work by H.U. SVERDRUP, MARTIN W. JOHNSON, and RICHARD H. FLEMING, *The Oceans* (1942, reissued 1970), is highly recommended. An in-depth, but quite readable, account of the general field of marine chemistry is provided by J.P. RILEY and R. CHESTER, *Introduction to Marine Chemistry* (1971). Thorough descriptions of each of the many subdisciplines in chemical oceanography can be found in the multivolume work by J.P. RILEY and G. SKIRROW (eds.), *Chemical Oceanography,* 2nd ed., including DANA R. KESTER, "Dissolved Gases Other than CO_2," vol. 1, ch. 8 (1975), pp. 497–556, a detailed discussion; P.J. LE B. WILLIAMS, "Biological and Chemical Aspects of Dissolved Organic Material in Sea Water," vol. 2, ch. 12 (1975), pp. 301–363, a good survey; J.D. BURTON, "Radioactive Nuclides in the Marine Environment," vol. 3, ch. 18 (1975), pp. 91–191; and KENNETH W. BRULAND, "Trace Elements in Sea-water," vol. 8, ch. 45 (1983), pp. 157–220, a readable account.

The Sea, vol. 1, *Physical Properties of Sea-water,* ed. by M.N. HILL (1962), contains information on all aspects of the behaviour of seawater with changes in temperature, pressure, and salt content, including discussions of density, transmission of light and sound, and sea ice properties. RHODES W. FAIRBRIDGE (ed.), *The Encyclopedia of Oceanography,* (1966), contains many useful entries on this topic, among them ROBERT GERARD, "Salinity in the Ocean," pp. 758–763, discussing spatial and temporal trends of the ocean's salinity and the processes that control it; and J. LAGRULA, "Hypsographic Curve," pp. 364–366, explaining the use of hypsometry and giving the hypsometry of the oceans and their subdivisions. H.W. MENARD and STUART M. SMITH, "Hypsometry of Ocean Basin Provinces," *Journal of Geophysical Research,* 71(18):4305–4325 (September 1966), is also useful.

(A.C.D./R.H.By.)

Circulation of the ocean waters: Useful books include those by OPEN UNIVERSITY OCEANOGRAPHY COURSE TEAM, *Ocean Circulation* (1989); GEORGE L. PICKARD and WILLIAM J. EMERY, *Descriptive Physical Oceanography: An Introduction,* 5th enlarged ed. (1990); STEPHEN POND and GEORGE L. PICKARD, *Introductory Dynamical Oceanography,* 2nd ed. (1983); HENRY STOMMEL, *A View of the Sea* (1987); and HENRY STOMMEL and DENNIS W. MOORE, *An Introduction to the Coriolis Force* (1989). The following journal articles are also of use: JAMES F. PRICE, ROBERT A. WELLER, and REBECCA R. SCHUDLICH, "Wind-driven Ocean Currents and Ekman Transport," *Science,* 238:1534–1538 (Dec. 11, 1987); W.D. NOWLIN, JR. and J.M. KLINCK, "The Physics of the Antarctic Circumpolar Current," *Reviews of Geophysics and Space Physics,* 24(3):469–491 (1986); and BRUCE A. WARREN, "Deep Circulation of the World Ocean," in BRUCE A. WARREN and CARL WUNSCH (eds.), *Evolution of Physical Oceanography* (1981), pp. 6–41.

(A.L.G.)

Waves of the sea: Books discussing waves and tides include OPEN UNIVERISTY OCEANOGRAPHY COURSE TEAM, *Waves, Tides, and Shallow-water Processes* (1989); ALBERT DEFANT, *Ebb and Flow: The Tides of Earth, Air, and Water* (1958; originally published in German, 1953); BLAIR KINSMAN, *Wind Waves: Their Generation and Propagation on the Ocean Surface* (1965, reprinted 1984); M. GRANT GROSS, *Oceanography,* 5th ed. (1990), ch. 8, "Waves," and ch. 9, "Tides," pp. 193–241; and DAVID T. PUGH, *Tides, Surges, and Mean Sea-level* (1987). C. GARRETT and W. MUNK, "Internal Waves in the Ocean," *Annual Review of Fluid Mechanics,* vol. 11, pp. 339–369 (1979), is also of interest.

(A.L.G.)

Density currents in the oceans: H.W. MENARD, "Turbidity Currents," ch. 9 in his *Marine Geology of the Pacific* (1964), pp. 191–222, provides a review of turbidity currents, written from the point of view of marine geology. A.H. BOUMA and A. BROUWER (eds.), *Turbidities* (1964), is a symposium with an extensive bibliography on sediments deposited by turbidity currents. More recent studies include GERARD V. MIDDLETON and MONTY A. HAMPTON, "Subaqueous Sediment Transport and Deposition by Sediment Gravity Flows," ch. 11 in DANIEL JEAN STANLEY and DONALD J.P. SWIFT (eds.), *Marine Sediment Transport and Environmental Management* (1976), pp. 197–218, which describes and discusses turbidity currents, grain flows, fluidized sediment flows, and debris flows; GARY PARKER, YUSUKE FUKUSHIMA, and HENRY M. PANTIN, "Self-accelerating Turbidity Currents," *Journal of Fluid Mechanics,* 171:145–181 (1986); and RICHARD J. SEYMOUR, "Nearshore Auto-suspending Turbidity Flows," *Ocean Engineering,* 13(5):435–447 (1986).

(G.V.M./Ed.)

Impact of ocean-atmosphere interactions on weather and climate: Overviews of the general relationship between air and ocean may be found in A.H. PERRY and J.M. WALKER, *The Ocean Atmosphere System* (1977); ADRIAN E. GILL, *Atmosphere-Ocean Dynamics* (1982); and NEIL WELLS, *The Atmosphere and Ocean: A Physical Introduction* (1986).

Seasonal and interannual ocean-atmosphere interactions are discussed by C.K. FOLLAND, T.N. PALMER, and D.E. PARKER, "Sahel Rainfall and Worldwide Sea Temperatures, 1901–85," *Nature,* 320:602–607 (April 17, 1986); JOHN M. WALLACE and QUANRONG JIANG, "On the Observed Structure of the Interannual Variability of the Atmosphere/Ocean Climate System," in HOWARD CATTLE (ed.), *Atmospheric and Oceanic Variability* (1987), pp. 17–43; T.N. PALMER and SUN ZHAOBO, "A Modelling and Observational Study of the Relationship Between Sea Surface Temperature in the North-west Atlantic and the Atmospheric General Circulation," *Quarterly Journal of the Royal Meteorological Society,* 111(470):947–975 (October 1985); JEROME NAMIAS, "Negative Ocean-Air Feedback Systems Over the North Pacific in the Transition from Warm to Cold Seasons," *Monthly Weather Review,* 104(9):1107–1121 (September 1976); and R.E. DAVIS, "Predictability of Sea Level Pressure Anomalies Over the North Pacific Ocean," *Journal of Physical Oceanography,* 8(2):223–233 (March 1978).

The formation of tropical cyclones is covered in WILLIAM M. GRAY, "Hurricanes: Their Formation, Structure, and Likely Role in the Tropical Circulation," in D.B. SHAW (ed.), *Meteorology Over the Tropical Oceans* (1979), pp. 155–218; J.F. PRICE, "Upper Ocean Response to a Hurricane," *Journal of Physical Oceanography,* 11(2):153–175 (February 1981); D.A. BROOKS, "The Wake of Hurricane Allen in the Western Gulf of Mexico," *Journal of Physical Oceanography,* 13(1):117–129 (January 1983); and GUY A. FRANCESCHINI and SAYED Z. EL-SAYED, "Effect of Hurricane Inez (1966) on the Hydrography and Productivity of the Western Gulf of Mexico," *The German Hydrographie Journal,* 21(5):193–202 (1968).

The Gulf Stream and Kuroshio systems are described in HENRY STOMMEL, *The Gulf Stream: A Physical and Dynamical Description,* 2nd ed. (1965, reissued 1976); WILLIAM H. MACLEISH, *The Gulf Stream* (1989), including a maritime history of the current as well as scientific data; N.P. FOFONOFF, "The Gulf Stream," in BRUCE A. WARREN and CARL WUNSCH (eds.), *Evolution of Physical Oceanography* (1981), pp. 112–139; and HENRY STOMMEL and KOZO YOSHIDA (eds.), *Kuroshio: Physical Aspects of the Japan Current* (1972). More in-depth essays include HENRY STOMMEL, "Asymmetry of Intercoastic Fresh-water and Heat Fluxes," *Proceedings of the National Academy of Science of the United States of America,* 77(5):2377–2381 (May 1980); CARL WUNSCH, "The Ocean Circulation in Climate," in JOHN T. HOUGHTON (ed.), *The Global Climate* (1984); and ALAN R. ROBINSON (ed.), *Eddies in Marine Science* (1983).

Studies of the El Niño/Southern Oscillation phenomena and their effect on climatic change are found in P.W. GLYNN (ed.), *Global Ecological Consequences of the 1982–83 El Nino-Southern Oscillation* (1990); S. GEORGE PHILANDER, *El Niño, La Niña, and the Southern Oscillation* (1990); WARREN S. WOOSTER and DAVID L. FLUHARTY (eds.), *El Niño North: Niño Effects in the Eastern Subarctic Pacific Ocean* (1985); RICHARD T. BARBER and FRANCISCO P. CHAVEZ, "Biological Consequences of El Niño," *Science,* 222(4629):1203–1210 (Dec. 16, 1983); THOMAS Y. CANBY, "El Niño's Ill Wind," *National Geographic,* 165(2):144–183 (February 1984); M.A. CANE, "El Niño," *Annual Review of Earth and Planetary Sciences,* 14:43–70 (1986); DAVID B. ENFIELD, "El Niño, Past and Present," *Reviews of Geophysics,*

27(1):159–187 (1989); NICHOLAS E. GRAHAM and WARREN B. WHITE, "The El Niño Cycle: A Natural Oscillator of the Pacific Ocean-Atmosphere System," *Science,* 240:1293–1302 (June 3, 1988); S. GEORGE PHILANDER and E.M. RASMUSSON, "The Southern Oscillation and El Niño," *Advances in Geophysics,* vol. 28, part A, pp. 197–215 (1985); and E.M. RASMUSSON, "El Niño and Variations in Climate," *American Scientist,* 73(2):168–177 (March–April 1985). In addition, the entire issue of *Oceanus,* vol. 27, no. 2 (Summer 1984), is devoted to El Niño studies.
(N.C.W./D.B.En.)

Ocean basins: Overviews of the geologic features of the deep-sea floor are given in JAMES P. KENNETT, *Marine Geology* (1982); and ALAN E.M. NAIRN and FRANCIS G. STEHLI (eds.), *The Ocean Basins and Margins,* 7 vol. in 9 (1973–88). Specific topics and geographic areas are studied by ROGER L. LARSON and WALTER C. PITMAN III, "World-wide Correlation of Mesozoic Magnetic Anomalies, and Its Implications," *Geological Society of America Bulletin,* 83(12):3645–3661 (December 1972); KEN C. MACDONALD and BRUCE P. LUYENDYK, "The Crest of the East Pacific Rise," *Scientific American,* 244(5):100–116 (May 1981); ARTHUR D. RAFF and RONALD G. MASON, "Magnetic Survey off the West Coast of North America, 40° N. Latitude to 52° N. Latitude," *Geological Society of America Bulletin,* 72(8):1267–1270 (August 1961); H.W. MENARD, "The Deep-ocean Floor," *Scientific American,* 221(3):126–142 (September 1969); B. PARSONS and J.G. SCLATER, "An Analysis of the Variations of Ocean Floor Bathymetry and Heat Flow," *Journal of Geophysical Research,* 82(5):803–827 (1977); C.M. POWELL, S.R. ROOTS, and J.J. VEEVERS, "Pre-breakup Continental Extension in East Gondwanaland and the Early Opening of the Eastern Indian Ocean," *Tectonophysics,* 155:261–283 (1988); DAVID B. ROWLEY and ANN L. LOTTES, "Reconstructions of the North Atlantic and Arctic: Late Jurassic to Present," *Tectonophysics,* 155:73–120 (1988); T. SIMKINS *et al.,* (1989); and two essays in E.L. WINTERER, DONALD M. HUSSONG, and ROBERT W. DECKER (eds.), *The Eastern Pacific Ocean and Hawaii* (1989), vol. N of the series "The Geology of North America": TANYA ATWATER, "Plate Tectonic History of the Northeast Pacific and Western North America," ch. 4, pp. 21–72; and KEN C. MACDONALD, "Tectonic and Magnetic Processes on the East Pacific Rise," ch. 6, pp. 93–110. (B.P.L.)

Continental margins: FRANCIS P. SHEPARD, *Submarine Geology,* 3rd ed. (1973), is somewhat dated but still considered by many marine geologists to contain the best treatment of continental margins; many people, however, consider the second ed. (1963) to be more general and better-organized. ELIZABETH K. BERNER and ROBERT A. BERNER, *The Global Water Cycle: Biochemistry and Environment* (1987), contains much information on the contribution that rivers make to margin sediments. CREIGHTON A. BURK and CHARLES L. DRAKE (eds.), *The Geology of Continental Margins* (1974), contains a general section on continental margins and many chapters on the details of specific margins, written for the advanced student. FRANCIS P. SHEPARD and ROBERT F. DILL, *Submarine Canyons and Other Sea Valleys* (1966), is the most complete treatment of submarine canyons in one volume, written in a style easily grasped by high school as well as advanced students. LARRY J. DOYLE and ORRIN H. PILKEY (eds.), *Geology of Continental Slopes* (1979); and K.O. EMERY, "The Continental Margins," *Scientific American,* 221(3):106–122 (September 1969), are also useful.
(L.J.D.)

Coastal and nearshore features: General discussions may be found in ERIC C.F. BIRD, *Coasts: An Introduction to Coastal Geomorphology,* 3rd ed. (1984); J.L. DAVIES, *Geographical Variation in Coastal Development,* 2nd ed. (1980); and MAURICE L. SCHWARTZ (ed.), *The Encyclopedia of Beaches and Coastal Environments* (1982).

Further information on coral reefs can be found in CHARLES DARWIN, *The Structure and Distribution of Coral Reefs* (1842, reissued 1984); J.A. FAGERSTROM, *The Evolution of Reef Communities* (1987); and ANDRÉ GUILCHER, *Coral Reef Geomorphology* (1988). For more detailed information on the Great Barrier Reef and coral reef devastation, respectively, see W.G.H. MAXWELL, *Atlas of the Great Barrier Reef* (1968); DAVID HOPLEY, *The Geomorphology of the Great Barrier* (1982); CHARLES BIRKELAND, "The Faustian Traits of the Crown-of-Thorns Starfish," *American Scientist,* 77(2):154–163 (March–April 1989); and E.H. WILLIAMS, C. GOENAGA, and V. VINCENTE, "Mass Bleachings on Atlantic Coral Reefs," *Science,* 238:877–878 (Nov. 13, 1987).

Various aspects of lagoons and estuaries are dealt with in K.H. MANN, *Ecology of Coastal Waters* (1982); GEORGE H. LAUFF (ed.), *Estuaries* (1967); BJÖRN KJERFVE (ed.), *Hydrodynamics of Estuaries,* 2 vol. (1988); STEPHEN P. LEATHERMAN, *Barrier Island Handbook,* 3rd ed. (1988); and T.E. PICKETT and R.L. INGRAM, "The Modern Sediments of Pamlico Sound, North Carolina," *Southeastern Geology,* 11(2):53–83 (1969).

Particular gulfs and bays are described in RHODES W. FAIRBRIDGE (ed.), *The Encyclopedia of Oceanography* (1966); A.J. HUXLEY (ed.), *Standard Encyclopedia of the World's Oceans and Islands* (1962); and ERIC C.F. BIRD and MAURICE L. SCHWARTZ (eds.), *The World's Coastline* (1985). (M.L.S.)

Economic aspects of the oceans: An entire issue of *Oceanus,* vol. 21, no. 4 (Winter 1984/85), is devoted to the impact of the Exclusive Economic Zone, especially as it pertains to the United States; coastal fishing, multiple-use management, marine pollution, and nonliving resources are some of the topics covered. ATULIO PONTECORVO, *The New Order of the Oceans: The Advent of a Managed Environment* (1986), deals with the new ocean regime and with the research and technology of marine resources from a global perspective, particularly emphasizing the international effects of the 1982 United Nations Law of the Sea Convention. The evolution of international marine policy and shipping law is compellingly discussed in EDGAR GOLD, *Maritime Transport* (1981), and "Ocean Shipping and the New Law of the Sea: Toward a More Regulatory Regime," *Ocean Yearbook,* vol. 6, pp. 85–96 (1986).

Discussions of the managed production of aquatic organisms include J.R. MUIR, "Aquaculture—Towards the Future," *Endeavour,* 9(1):52–55 (1985); and JOHN BARDACH, "Aquaculture: Moving from Craft to Industry," *Environment,* 30(2):6–11, 36–41 (March 1988). Desalinization processes are discussed and illustrated in ALAN D.K. LAIRD, "The Potable Sea: Taking the Salt from Saltwater," *Oceans,* 15(5):25–29 (September–October 1982); and ROBERTA FRIEDMAN, "Salt-free Water from the Sea," *Sea Frontiers,* 36(3):49–54 (May–June 1990).

Overviews of various ocean energy resources and technologies may be found in MAXWELL BRUCE, "Ocean Energy: Some Perspectives on Economic Viability," *Ocean Yearbook,* vol. 5, pp. 58–78 (1985); and TERRY R. PENNEY and DESIKAN BHARATHAN, "Power from the Sea," *Scientific American,* 256(1):86–92 (January 1987). Tidal power generation is covered by B. COUNT (ed.), *Power from Sea Waves* (1980), based on conference proceedings; and MICHAEL E. MCCORMICK, *Ocean Wave Energy Conversion* (1981). Summaries of ocean thermal energy conversion techniques and future prospects are provided by R. COHEN, "Energy from the Ocean," *Philosophical Transactions of the Royal Society of London,* Series A, 307:405–437 (1982); and D.E. LENNARD, "Ocean Thermal Energy Conversion—Past Progress and Future Prospects," *IEE Proceedings,* vol. 134, part A, no. 5, pp. 381–391 (May 1987).

Current and future petroleum and mineral resources in the ocean environment and the technologies necessary to recover them are addressed in GERARD J. MANGONE (ed.), *The Future of Gas and Oil from the Sea* (1983); ELISABETH MANN BORGESE, *The Mines of Neptune: Minerals and Metals from the Sea* (1985); DAVID CRONAN, "A Wealth of Sea-floor Minerals," *New Scientist,* 106:34–38 (June 6, 1985); and JAMES M. BROADUS, "Seabed Materials," *Science,* 235:853–860 (Feb. 20, 1987).

Discussions of the use of the ocean as a site for waste disposal, and the problems of marine pollution, include IVER W. DUEDALL *et al.* (eds.), *Wastes in the Ocean,* 6 vol. (1983–85), on industrial and sewage, radioactive, and energy wastes, dredged-material disposal, and deep-sea and nearshore waste disposal; R.B. CLARK, *Marine Pollution,* 2nd ed. (1989); WESLEY MARX, *The Oceans: Our Last Resource* (1981); DAVID K. BULLOCH, *The Wasted Ocean* (1989); and a complete issue of *Oceanus,* vol. 33, no. 2 (Summer 1990). (Ed.)

Olympic Games

Just how far back in history organized athletic contests were first held remains a matter of doubt, but it is reasonably certain that they occurred in Greece, at least, some 3,500 years ago.

However ancient in origin, by the end of the 6th century BC at least four of the Greek sporting festivals, sometimes known as classical games, had achieved major importance. They were the Olympic Games, held at Olympia; the Pythian Games at Delphi; the Nemean Games at Nemea; and the Isthmian Games at Corinth. Later, similar festivals were held in nearly 150 cities as far afield as Rome, Naples, Odessus, Antioch, and Alexandria.

The Olympic Games in particular were to become famous throughout the Greek world. There are records of the champions at Olympia from 776 BC to AD 217. The Games, held every four years, were abolished in AD 393 by the Roman emperor Theodosius I, probably because of their pagan associations. For the first 100 or 200 years, Olympic champions came from a dozen or more Greek cities, the majority from Sparta and Athens, but in the next three centuries, athletes were drawn from 100 cities in the Greek empire. And in the final 100 years or so before the games were discontinued, champions came from as far from Olympus as Antioch, Alexandria, and Sidon.

In 1887 the 24-year-old Baron Pierre de Coubertin conceived the idea of reviving the Olympic Games and spent seven years preparing public opinion in France, England, and the United States to support his plan. At an international congress in 1894, his plan was accepted and the International Olympic Committee was founded. The first modern Olympic Games were held in Athens in April 1896, with 13 nations sending nearly 300 representatives to take part in 42 events and 10 different sports. The revival of the Olympic Games led to the formation of many international bodies controlling their own amateur sports and to the creation of National Olympic Committees in countries throughout the world.

This article is divided into the following sections:

Early history 197
  Greece
  Rome
Revival of the Olympics 198
Organization of the modern Olympics 198
  The International Olympic Committee
  National Olympic committees
  IOC awards
  The Games and participants
  The Olympic Village
  Ceremonies
  The awarding of the Olympic Games
History of the modern Games 200
  First period: 1896–1912
  Second period: 1920–36
  Third period: 1948–68
  Fourth period: 1972 onward
  The Winter Olympic Games
Bibliography 201

EARLY HISTORY

Greece. Of all the games held throughout Greece, those staged at Olympia in honour of Zeus are the most famous. Held every four years between August 6 and September 19, they occupied such an important place in Greek life that time was measured by the interval between them—an Olympiad. Although the first Olympic champion listed in the records was one Coroebus of Elis, a cook, who won the sprint race in 776 BC, it is generally accepted that the Games were probably at least 500 years old at that time. According to one legend they were founded by Heracles, son of Alcmene. The Games, like all Greek games, were an intrinsic part of a religious festival. They were held at Olympia in the city-state of Elis, on a track about 32 metres (35 yards) wide. The racing length was one stade, a distance of about 192 metres (210 yards). In the early Olympics a race, called a stade, covered one length of the track. Horse racing, which became part of the ancient games, was held in the hippodrome, south of the stadium.

The first Olympic champion

At the meeting in 776 BC, there was apparently only one event, the stade, but other events were added over the ensuing decades. In 724 BC a two-length race, *diaulos*, roughly similar to the 400-metre race, was included, and four years later the *dolichos*, a long-distance race possibly to be compared to the modern 1,500- or even 5,000-metre event, was added. Wrestling and the pentathlon were introduced in 708 BC. The latter was an all-around competition consisting of five events—the long jump, javelin throw, discus throw, foot race, and wrestling.

Boxing was introduced in 688 BC, and in 680 a chariot race. In 648 the pancratium (Greek *pankration*), a kind of all-strength, or no-holds-barred, wrestling, was included. Kicking and hitting were allowed; only biting and gouging (thrusting a finger or thumb into an opponent's eye) were forbidden. Between 632 and 616 BC events for boys were introduced. And from time to time further events were added, including contests for fully armed soldiers, for heralds, and for trumpeters. The program must have been as varied as that of the modern Olympics, although the athletics (track and field) events were limited; there was no high jumping in any form and no individual field event, except in the pentathlon.

Until the 77th Olympiad (472 BC) all of the contests took place on one day; later they were spread, with, perhaps, some fluctuation, over four days, with a fifth devoted to the closing-ceremony presentation of prizes and a banquet for the champions. Sources generally agree that women were not allowed as competitors or, except for the priestess of Demeter, as spectators. In most events, the athletes participated in the nude.

The Olympic Games were originally restricted to free-born Greeks. The competitors, including those who came from the Greek colonies, were amateur in the sense that the only prize was a wreath or garland. The athletes underwent a most rigorous period of supervised training, however, and eventually the contestants were true professionals. Not only were there substantial prizes for winning, but the Olympic champion also received adulation and unlimited benefits from his city. Athletes became full-time specialists—a trend that in the modern games has caused a long and bitter controversy over amateurism.

Early professionalism

Rome. Greece lost its independence to Rome in the middle of the 2nd century BC, and the support for the competitions at Olympia and other places fell off considerably in the next century. The Romans looked on athletics with contempt—to strip naked and to contend in public was degrading in the eyes of the Roman citizen. The Romans realized the value of the Greek festivals, however, and Augustus, who had a genuine love for athletics, staged athletic games in a temporary wooden stadium erected near the Circus Maximus. Nero was also a keen patron of the festivals in Greece. By the 4th century AD, Rome, with its population of more than 1,000,000, had well over 150 holidays for games. There was chariot racing in the hippodrome and horse racing in the Circus Maximus, with room for more than 250,000 spectators. In an amphithe-

atre with accommodation for 50,000, animals and human beings were maimed and slaughtered in the name of sport.

Indeed, public games were held in abundance, but for the Romans athletic events occupied a secondary position. The only ones that really interested them were the fighting events—wrestling, boxing, and the pancratium. The main difference between the Greek and Roman attitude was that the Roman festivals were described as *ludi* (games), the Greek as *agônes* (contests). The Greeks originally organized their games for the competitors, the Romans for the public. One was primarily competition, the other entertainment; and it is not unreasonable to suggest that the Greeks took an "amateur" view of sport, the Romans a professional one.

REVIVAL OF THE OLYMPICS

The architect of the modern Olympics was Baron Pierre de Coubertin, born in Paris on New Year's Day, 1863. As a young man he was intensely interested in literature and in education and sociology. Family tradition pointed to an army career or possibly politics, but at the age of 24 Coubertin decided that his future lay in education. At the same time he had the idea of reviving the Olympic Games, and he propounded his desire for a new era in international sport when on November 25, 1892, at a meeting of the Union des Sports Athlétiques in Paris, he said:

> Let us export our oarsmen, our runners, our fencers into other lands. That is the true Free Trade of the future; and the day it is introduced into Europe the cause of Peace will have received a new and strong ally. It inspires me to touch upon another step I now propose and in it I shall ask that the help you have given me hitherto you will extend again, so that together we may attempt to realise, upon a basis suitable to the conditions of our modern life, the splendid and beneficent task of reviving the Olympic Games.

The speech did not produce any appreciable activity, but Coubertin was not fainthearted. At a conference on international sport in Paris in June 1894 at which Coubertin raised the possibility of the revival of the Olympic Games, there were 79 delegates representing 49 organizations from nine countries. Coubertin himself wrote that except for his co-workers Dimítrios Vikélas of Greece, who was to be the first president of the International Olympic Committee, and Professor William M. Sloane of the U.S. from the College of New Jersey (later Princeton University), no one had real interest in the revival of the Games. Nevertheless, and to quote Coubertin again, "a unanimous vote in favour of revival was rendered at the end of the Congress chiefly to please me."

It was at first agreed that the Games should be held in Paris in 1900. Six years seemed a long time to wait, however, and it was decided to change the venue—what better site than Athens, the capital of Greece—and the date, to April 1896. A great deal of indifference, if not opposition, had to be overcome, including a refusal by Athens to stage the Games at all. But Coubertin and his newly elected International Olympic Committee of 14 members won through, and the Games were opened by the King of Greece in the first week of April 1896.

First games of the modern cycle

ORGANIZATION OF THE MODERN OLYMPICS

The International Olympic Committee. At the Congress of Paris in 1894, the control and development of the modern Olympic Games was entrusted to the International Olympic Committee (IOC; Comité International Olympique), with headquarters to be established in Switzerland. It is responsible for maintaining the regular celebration of the Olympic Games; seeing that the Games are carried out in the spirit that inspired their revival; and promoting the development of amateur sport throughout the world. The original committee in 1894 consisted of 14 members and Coubertin, and membership since then has been self-perpetuating.

Convinced that the downfall of the ancient Olympic Games had been caused by outside influences that undermined the spirit of the Games, Coubertin felt that the revived Games would go the same way unless they were in the hands of people whose concern was to keep the spirit of amateur sport alive and who were responsible in no way to any outside influences. Thus IOC members are regarded as ambassadors from the IOC to their national sports organizations. They are in no sense delegates to the committee and may not accept from the government of their country, or from any organization or individual, any instructions that in any way affect their independence.

Independence of International Committee members

The IOC is a permanent organization that elects its own members. Each member—the present membership is about 70—must speak French or English and be a citizen of or reside in a country that has a National Olympic Committee. With a very few exceptions, there is only one member from any one country. Members were originally elected for life, but anyone elected after 1965 must retire at 75.

The IOC elects its president for a period of eight years, at the end of which he is eligible for reelection for further periods of four years each. Juan António Samaranch (Spain) was elected in 1980. Previous presidents were Dimítrios Vikélas (1894–96, Greece), Baron Pierre de Coubertin (1896–1925, France), Count Henri de Baillet-Latour (1925–42, Belgium), J. Sigfrid Edström (1946–52, Sweden), Avery Brundage (1952–72, United States), and Michael Morris, Lord Killanin (1972–80, Ireland).

Presidents of the IOC

The executive board of 11 members holds periodic meetings with the international federations and National Olympic Committees. The IOC as a whole meets annually, and a meeting can be convened at any time that one-third of the members so request.

National Olympic Committees. Each country that desires to participate in the Olympic Games must have an Olympic committee accepted by the IOC. By 1988 there were 167 such committees.

A National Olympic Committee is composed of at least five national sporting federations, each affiliated to an appropriate international federation. The ostensible purpose of these National Olympic Committees is the development and promotion of the Olympic movement and of amateur sport. National Olympic Committees arrange to equip, transport, and house their country's representatives at the Olympic Games. According to the rules of the committee, they must be not-for-profit organizations; must not associate themselves with affairs of a political or commercial nature; and must be completely independent and autonomous and in a position to resist all political, religious, or commercial pressure.

A person who has ever competed in sports as a professional, who has ever coached sports competitors for payment, or who is engaged in or connected with sport for personal profit is not eligible to serve on a national committee. The rules provide that

> exceptions to these categories may be made by the Executive Board of the I.O.C. on the recommendation of the National Olympic Committee concerned.

National Olympic Committees that do not conform to IOC rules and regulations forfeit their recognition and their right to send participants to the Olympic Games.

IOC awards. In individual Olympic Games events, the award for first place is a gold (silver-gilt, with six grams of fine gold) medal, for second place a silver medal, and for third place a bronze medal. Solid gold medals were last given in 1912. Diplomas are awarded for fourth, fifth, sixth, seventh, and eighth places. All competitors and officials receive a commemorative medal.

The IOC presents two noncompetitive awards, the Olympic Cup and the Olympic Order. The former was instituted by Baron de Coubertin in 1906. It is awarded to an association or institution that has a general reputation for merit and integrity, that has eminently served the cause of amateur sport, and that has made a substantial contribution to the Olympic movement. The first recipient of the Olympic Cup was the Touring Club de France.

The Olympic Order, created in 1974, is intended for living persons, excluding active members of the IOC. There are two degrees of award—gold and silver—to be made "to any person who has illustrated the Olympic ideal through his action, has achieved remarkable merit in the sporting world, or has rendered outstanding services to the Olympic cause, either through his own personal achievement or his contribution to the development of sport." The first gold

Olympic Order

medal was awarded posthumously to Avery Brundage, for 20 years president of the IOC. Among the silver awards that have been made were those to Jesse Owens, winner of four gold medals at the 1936 Olympics, and to Dan Ferris, for more than 50 years a prominent administrator in U.S. and world track and field. Bronze medals had also been awarded but were discontinued in 1981.

Five previous awards were suspended in 1974: the Olympic Diploma of Merit, first awarded in 1905; the Sir Thomas Fearnley Cup and the Mohammed Taher Trophy, first awarded in 1951; the Count Alberto Bonacossa Trophy, established in 1955; and the Tokyo Trophy, first awarded in 1967.

The Games and participants. The Olympic Games (Summer) are held every four years. There is no age limit for competitors, unless one has been established by a sport's international federation. No discrimination is allowed on grounds of race, religion, or political affiliation. The Games are contests between individuals and not between countries. Separate Winter Games have been held since 1924.

The Olympic Games celebrate an Olympiad, or period of four successive years. The first Olympiad of modern times was celebrated in 1896, and subsequent Olympiads are numbered consecutively, even when no Games take place (as was the case in 1916, 1940, and 1944).

In 1986 the IOC voted to alternate the Summer and Winter Games every two years, beginning in 1994. The Winter Games would be held in 1992 and again in 1994 and thereafter would be held every four years; the Summer Games would maintain their original four-year cycle.

The maximum number of entries permitted for individual events is three per nation. The number is fixed (but can be varied) by the IOC in consultation with the international federation concerned. In team events only one team per country is allowed. In general, a National Olympic Committee may enter only a citizen of the country concerned.

Problems of eligibility. To be allowed to compete, a competitor must meet the eligibility rules as defined by the international body of the particular sport and also by the rules of the IOC.

In recent years there has been a tendency away from the strict definition of amateur status. In 1971 the IOC decided to eliminate the term amateur from the Olympic Charter. Subsequently the eligibility rules were amended to permit "broken-time" payments to compensate athletes for time spent away from work during training and competition. The IOC also legitimized the sponsorship of athletes by National Olympic Committees, sports organizations, and private businesses. In some sports, including track and field, figure skating, and skiing, athletes must place income from commercial endorsements and sponsorships in a restricted trust fund, which is controlled by their national federation.

In 1986 the IOC adopted rules that permit the international federation governing each Olympic sport to decide whether to permit professional athletes in Olympic competition. Professionals in ice hockey, tennis, soccer, and equestrian sports were permitted to compete in the 1988 Olympics, although their eligibility was subject to some restrictions, and professionals from sports other than track and field were eligible for Olympic track and field events.

It now appears to some of those who have been closely associated with the Games that they have become too vast, too nationalistic, too expensive, and too commercial. It is felt that no longer is the important thing "to take part" but that winning a gold medal is the only thing that matters. The primary consideration appears to be no longer the competitors, more than 95 percent of whom cannot win gold medals, but the public. Entertainment now seems to supersede the enjoyment of competition, and the demand for entertainment may well be the death of amateur sport.

The Olympic sports

Programs and events. A Summer Olympic program includes the following sports: archery, association football (soccer), athletics (track and field), badminton, baseball, basketball, boxing, canoeing, cycling, diving, equestrian sports, fencing, field hockey, gymnastics, team handball, judo, modern pentathlon, rowing, shooting, softball, swimming, table tennis, tennis, volleyball, water polo, weight lifting, wrestling, and yachting. Women may participate in all the sports except baseball, boxing, modern pentathlon, water polo, weight lifting, and wrestling. Men do not compete in softball. An Olympic program may also include up to two demonstration sports and in addition must include national exhibitions and demonstrations of fine arts (architecture, literature, music, painting, sculpture, photography, and sports philately).

The particular events included in the different sports are a matter for agreement between the IOC and the international federations. For many years there were 24 events for men in track and field. In 1976, however, the men's 50,000-metre walk was excluded (on the grounds that the total number of competitors in the Games must be reduced), while women's rowing events were included for the first time. It appears that, although the IOC feels that in some way there must be a limit placed upon the total number of competitors, it is unwilling to take the severe action of excluding certain sports.

The Olympic Village. The Olympic Village was first introduced at Los Angeles in 1932. The organizing committee provides the village so that competitors and team officials can be housed together and fed at a reasonable price. The village is located as close as possible to the main stadium and other facilities and has separate accommodations for men and women. Only competitors and officials may live in the village, and the number of team officials is limited.

Ceremonies. *The Olympic flag.* In the stadium and its immediate surroundings, the Olympic flag is flown freely together with the flags of the nations taking part. The Olympic flag presented by Baron Coubertin in 1914 is the prototype: it has a white background and in the centre there are five interlaced rings—blue, yellow, black, green, and red. The blue ring is on the left next to the pole. These rings represent the five continents joined together in the Olympic Movement. The Olympic motto is *Citius—Altius—Fortius* ("Faster—Higher—Stronger").

The opening ceremony. The form of the opening ceremony is laid down by the IOC in great detail, from the moment when the chief of state of the host country is received by the president of the IOC and the organizing committee at the entrance to the stadium to the end of the proceedings when the last team files out. The rules provide that participants are not permitted to carry cameras into the arena, but this provision is always ignored.

When the head of state has reached his place in the tribune, he is greeted with the national anthem of his country, and the parade of competitors begins. The Greek team is always the first to enter the stadium, and, except for the host team, which is always last, the other nations follow in alphabetical order as determined by the language of the organizing country. Each contingent, dressed in its official uniform, is preceded by a shield with the name of its country, while an athlete carries its national flag. At the 1980 games, some of the countries protesting the Soviet Union's involvement in Afghanistan carried the Olympic flag in place of their national flag. The competitors march around the stadium and then form in groups in the centre facing the tribune.

The president of the organizing committee then delivers a brief speech of welcome, followed by another brief speech from the president of the IOC, who asks the chief of state to proclaim the Games open.

A fanfare of trumpets is sounded as the Olympic flag is slowly raised; pigeons are released, symbolically to fly to the countries of the world with the news that the Games are open.

The Olympic flame

The Olympic flame is then carried into the stadium by the last of the runners who have brought it from Olympia, Greece. The runner circles the track, mounts the steps, and lights the Olympic fire that burns night and day during the Games. In 1968 a woman carried the flame into the stadium, and in 1976 the flame was borne jointly by a male and a female athlete. In 1984, at the Games in Los Angeles, a female runner brought the flame into the stadium but passed it to a male, who ran up the steps to light the Olympic fire.

Victory ceremony. Medals are presented during the Games at the various venues and usually soon after the conclusion of the event. The competitors who have won the first three places proceed to the rostrum, with the winner (gold medalist) in the centre, the silver medalist on his or her right, and the bronze medalist on the left. The medals, attached to a chain or ribbon, are hung around the necks of the winners by a member of the IOC, and the flags of the nations concerned are raised to the top of the flagpoles while an abbreviated form of the national anthem of the winner is played. The spectators are expected to stand and face the flags, as do the three successful athletes.

Closing ceremony. The closing ceremony takes place after the final event, which is usually the equestrian Prix des Nations. Since the Melbourne Olympics in 1956, the closing ceremony has been a less formal affair, and after certain formalities have been observed, the athletes taking part stage their own demonstrations.

The ceremonies include a parade of athletes, six from each nation, marching eight or 10 abreast without distinction of nationality, and signifying the friendly bonds of Olympic sport. The president of the IOC calls the youth of the world to assemble in four years to celebrate the Games of the next Olympiad. A fanfare is sounded, the Olympic fire is extinguished, and to the strains of the Olympic anthem the Olympic flag is lowered and the Games are over.

The awarding of the Olympic Games. The honour of holding the Olympic Games is entrusted to a city and not to a country. The choice of the city lies solely with the IOC. Application to hold the Games is made by the chief authority of the city, with the support of the national government.

Applications must state that no political meetings or demonstrations will be held in the stadium or other sports grounds or in the Olympic Village, and it must be promised that every competitor shall be given free entry without any discrimination on grounds of religion, colour, or political affiliation. This involves the assurance that the national government will not refuse visas to any of the competitors. At the Montreal Olympics in 1976, however, the Canadian government refused visas to the representatives of Taiwan because they were unwilling to forgo the title of the Republic of China, under which their National Olympic Committee was admitted to the IOC. This Canadian decision, in the opinion of the IOC, did great damage to the Olympic Games, and it was later resolved that any country in which the Games are organized must undertake strictly to observe the rules. Enforcement would be difficult and even the use of severe penalties by the IOC might not guarantee elimination of infractions.

HISTORY OF THE MODERN GAMES

Since the revival of the Games, growth in the numbers of competitions, of competitors, and of participating countries has been almost continuous. The history of the modern Games may be conveniently divided into four periods: (1) from the revival in 1896 to the outbreak of World War I, (2) the period between the two World Wars, (3), from the resumption of the Games in 1945 through the 1968 Games, and (4) from the 1972 games onward.

First period: 1896–1912. It took some time for the Games to find their continuing pattern. Of the five celebrations in the first period, the first three—Athens in 1896, Paris in 1900, and St. Louis, Missouri, in 1904—were somewhat haphazard affairs. The entries were unlimited and hardly ever "national" in the sense of representing each nation's best performers. The actual events included varied considerably, and it is not possible to determine with any degree of confidence which of the events were really of Olympic significance. Women competed at golf and lawn tennis in the Paris Games and at archery in the St. Louis Games.

London, 1908

The Games of 1908 in London were the first to be organized by the various sporting bodies concerned and were not regarded as an appendage to a world fair and controlled by the promoters. Though the holding of a Franco-British Exhibition at Shepherd's Bush, London, made possible the construction of a new stadium—the famous White City, with a capacity of 66,000—the governing bodies of sport were in control. Twenty-two nations sent more than 2,000 athletes (but only 36 women) to compete in more than 100 events in 17 different sports. Archery and lawn tennis were included in the 1908 program.

At Stockholm in 1912, the number of nations increased to 28, the number of competitors exceeded 2,500, and for the first time three swimming events for women were included. Because of World War I, no Games were held in 1916.

Second period: 1920–36. This period saw the golden age of athletes—from the Finnish domination of middle- and long-distance running to the performances in the sprints, especially that of Jesse Owens. In this period, too, there were five celebrations. In less than two years after the Armistice in November 1918, Belgium organized the first post-World War I Olympics, opened at Antwerp by King Albert I. Twenty-nine nations, including many new ones, sent more than 2,500 competitors to take part in more than 150 events in 19 sports. Women competitors were still a mere handful (about 60). Four years later, in Paris, the number of countries increased to 44, with more than 3,000 competitors for 137 events, a decrease because of a reduction in the excessive number of shooting and yachting events. More than 100 women took part, competing in fencing as well as swimming and lawn tennis. After 1924 a question of eligibility in lawn tennis caused the IOC to omit the sport from future Games.

Increase of women's events

In 1928 in Amsterdam, women's track and field competitions (five events) and one event for women in gymnastics were introduced. Again there were 3,000 athletes, from 46 countries, but the number of women (290) was more than double that at Paris (136). Four years later the Games for the second time crossed the Atlantic, to Los Angeles, and there was a falling off in the number of competitors—less than 1,500 from 37 countries. The long journey from Europe and the heavy cost during a major depression were responsible for the decreases. The return to Europe for the Berlin Games produced 4,000 athletes in 1936.

Third period: 1948–68. In this period there was a significant rise in the number of women competitors and the number of events open to women. Sports other than track and field, such as swimming, came into prominence. Most of all an increase in television coverage of both Winter and Summer Games increased the audience. Because of World War II, no Games were held in 1940 or 1944. London was given the task of organizing the first postwar Games in 1948 with a bare three years available. This was accomplished despite difficult postwar conditions, with rationing of various materials and supplies still in effect and with expenditures limited. The number of competing countries was 59, with 4,000 competitors, of whom 385 were women. Track and field events for women were increased to nine by the addition of the 200-metre run, the broad (now long) jump, and the shot put.

Finland was host to the 1952 Games, at Helsinki. The number of countries participating rose to 69, competitors totalled within 100 of 6,000, and women topped the 500 mark. In 1956 for the first time the venue was in the Southern Hemisphere. The Games in Melbourne were celebrated in November and December, and for the first time one of the sports had to take place in another country. Because of quarantine regulations that prohibited the importation of horses to Australia, the equestrian events were decided in Stockholm. Again the distance from Europe reduced the competitors in Melbourne by 1,500.

In Rome four years later, however, the number of competitors passed the 5,000 mark, the number of nations rose to more than 80, and the number of events rose to 150. Originally Rome was to have staged the 1908 Games.

First Asian Games

The Games were in Asia for the first time in 1964, at Tokyo, again in a city where the previous award (in 1940) had not been fulfilled. Once more there were record figures: 94 countries, 162 events, and 5,500 competitors.

There was considerable criticism when Mexico City was awarded the 1968 Games. Controversy surrounded charges that the altitude, more than 2,300 metres (7,500 feet), would adversely affect the majority of distance runners, who would not have been able to spend many months liv-

ing in a comparably rarefied atmosphere. The number of competitors rose to more than 6,500 from 125 countries. Women numbered 800, the total events 172.

Fourth period: 1972 onward. This period has been marked by political violence and controversy. In Munich in 1972 there were nearly 200 different events, with more than 7,000 competitors from 122 countries. Tragedy struck the Games when Arab (Palestinian) terrorists invaded the Olympic Village and killed two Israeli athletes and seized nine as hostages for the release of 200 Arab prisoners in Israel; all nine, five of their captors, and a West German policeman were slain when police rescue attempts failed.

The Munich tragedy

In Montreal in 1976 the number of entries exceeded those of Munich, but the last-minute withdrawals by countries protesting the presence of the New Zealand team because of their rugby affiliation with South Africa, whose apartheid policies were held to be objectionable, reduced the actual competitors by more than 400.

The games of the XXII Olympiad at Moscow in 1980 were marred by the lack of participation of some 60 eligible nations in protest against the Soviet Union's invasion of Afghanistan in December 1979. More than 5,000 competitors representing 80 nations chose not to take part, but the Games were further diminished by the intrusion of two other issues—alleged bias in judging by eastern European officials and the use of anabolic steroids by athletes to improve performance. The latter issue was exacerbated just before the Games when seven athletes caught using the drugs in 1979 were made eligible for the Olympics against the rules of the International Amateur Athletic Federation.

Preparations for the 1984 Games at Los Angeles were complicated by the fear of injurious effects of the infamous Los Angeles smog on athletes, by jurisdictional disputes over security by various law enforcement agencies, by legal suits over scheduling of endurance events for women athletes, and by the decision of the Soviet Union and a number of its allies not to send teams to participate in the Games.

Nearly 8,500 athletes from 159 countries participated in the 1988 Games at Seoul. The program included 26 sports, with 237 separate events. The use of steroids and other drugs continued to be a major issue; 10 athletes, 5 of them medalists, were disqualified for using banned substances.

The 1992 Summer Olympic Games, hosted by Barcelona, Spain, featured 169 teams with more than 9,000 athletes in 257 medal events. It marked the last time that athletes from the former Soviet Union took part in the Olympic Games as a single team.

The 1996 Games in Atlanta, Georgia, were attended by more than 10,000 athletes from 197 nations. The number of medal events grew to 27, and the quality of the venues and performances were exceptional. However, sponsorships and vendors licenses were liberally awarded in order to cover the expense of the Games, and the excessive commercial presence in Atlanta drew considerable criticism.

The IOC was faced with scandal leading up to the 2000 Games in Sydney, Australia. Investigations revealed that committees bidding to host the 2000 Summer Games and 2002 Winter Games had offered extravagant gifts and grants to IOC members in exchange for votes. As a result, several IOC members were expelled and revisions to the bid process were made.

The 2004 Games were scheduled for Athens, Greece.

The Winter Olympic Games. While some skating events were included in the 1908 and 1920 Games, the Winter Games were accepted as a celebration comparable to the Summer Games and given the official blessing of the IOC in 1924. The first Winter Games were held at Chamonix, France, and consisted of 16 events. There were 16 participating countries, and the participants numbered less than 300. Subsequent Winter Games were held at St. Moritz, Switzerland (1928); Lake Placid, New York (1932); Garmisch-Partenkirchen, Germany (1936); St. Moritz (for a second time; 1948); Oslo (1952); Cortina d'Ampezzo, Italy (1956); Squaw Valley, California (1960); Innsbruck, Austria (1964); Grenoble, France (1968); Sapporo, Japan (1972); Innsbruck (for a second time; 1976); Lake Placid (for a second time; 1980); Sarajevo, Yugoslavia (1984); Calgary, Alberta (1988); Albertville, France (1992); Lillehammer, Norway (1994; the IOC had decided to hold the Winter Olympics two years after the Summer Olympics starting in 1994); and Nagano, Japan (1998). The 2002 Winter Olympics were scheduled for Salt Lake City, Utah, and the 2006 Winter Olympics for Turin, Italy.

The first Winter Games

BIBLIOGRAPHY. H.A. HARRIS, *Greek Athletes and Athletics* (1964, reissued 1979); E. NORMAN GARDINER, *Greek Athletic Sports and Festivals* (1910, reissued 1970), and *Athletics of the Ancient World* (1930, reprinted 1987); and M.I. FINLEY and H.W. PLEKET, *The Olympic Games: The First Thousand Years* (1976), are histories of ancient games. PIERRE DE COUBERTIN, *Olympic Memoirs* (1979; originally published in French, 1931); BILL HENRY and PATRICIA HENRY YEOMANS, *An Approved History of the Olympic Games,* new ed. (1984); JOHN KIERAN, ARTHUR DALEY, and PAT JORDAN, *The Story of the Olympic Games: 776 B.C. to 1976,* rev. ed. (1977); and NICOLAOS YALOURIS (ed.), *The Eternal Olympics,* trans. from Greek (1979), are reports of the Games up to modern times. JOHN E. FINDLING and KIMBERLEY D. PELLE, *Historical Dictionary of the Modern Olympic Movement* (1996), discusses each of the Olympic Games, from the ancient games through the 2000 Olympics in Sydney, in their historical contexts. JEFFREY O. SEGRAVE and DONALD CHU (eds.), *The Olympic Games in Transition* (1988), explores the development of and changes in the modern Games, and *Olympism* (1981), focuses on the philosophy of the Olympics. The philosophy of "Olympism" is further analyzed in JOHN HOBERMAN, *The Olympic Crisis: Sport, Politics, and the Moral Order* (1986). DAVID B. KANIN, *A Political History of the Olympic Games* (1981), is an important study of the political aspects. An analysis of politics as well as athletics is found in ALLEN GUTTMANN, *The Olympics: A History of the Modern Games* (1992). JOHN A. LUCAS, *Future of the Olympic Games* (1992), addresses various problems of the modern Games and proposes solutions. Other Olympic histories written from a political perspective include CHRISTOPHER R. HILL, *Olympic Politics* (1992, rev. ed. 1996), and ALFRED ERICH SENN, *Power, Politics, and the Olympic Games* (1999). KRISTINE TOOHEY and A.J. VEAL, *The Olympic Games: A Social Science Perspective* (2000), is written from a sociological perspective. JOHN J. MACALOON, *This Great Symbol: Pierre de Coubertin and the Origins of the Modern Olympic Games* (1981), is a scholarly biography. ALLEN GUTTMANN, *The Games Must Go On: Avery Brundage and the Olympic Movement* (1984), concentrates on Brundage's tenure as president of the International Olympic Committee (1948–72). LORD KILLANIN (MICHAEL MORRIS, BARON KILLANIN), *My Olympic Years* (1983), provides a personal account of Killanin's time as IOC president (1972–80). There are official reports of every celebration since 1904. PETER LOVESEY and TOM MCNAB, *The Guide to British Track and Field Literature, 1275–1968* (1969), includes 50 works on the Olympic Games. BILL MALLON, *The Olympics: A Bibliography* (1984), covers the Games from 1896 to 1980. A visual history of the 1996 Atlanta Games is given in ANNIE LEIBOVITZ, *Olympic Portraits* (1996).

(H.M.A./Ed.)

Principles of Optics

Originally, the term optics was used only in relation to the eye and vision. Later, as lenses and other devices for aiding vision began to be developed, these were naturally called optical instruments, and the meaning of the term optics eventually became broadened to cover any application of light, even though the ultimate receiver is not the eye but a physical detector, such as a photographic plate or a television camera. Within the present century optical methods have been applied extensively to regions of the electromagnetic radiation spectrum not visible to the eye, such as X-rays, ultraviolet, infrared, and microwave radio waves, and to this extent these regions are now often included in the general field of optics.

In the present article the image-forming properties of lenses, mirrors, and other devices that make use of light are considered. The wave and quantum nature of light, its velocity, wavelength, polarization, diffraction, and interference may be found in LIGHT. The interrelations between light and electricity, which are involved in light sources and detectors, are treated in ELECTRONICS: *Electron tubes*. The analysis of light into its component colours by prisms and gratings forms the basis of the extensive field of spectroscopy, the principles of which are discussed in ANALYSIS AND MEASUREMENT: *Spectroscopy*. For information about the reception of light by the retina of the eye and the interpretation of images by the brain, see SENSORY RECEPTION: *Vision*.

This article is divided into the following sections:

Geometrical optics 202
- General considerations 202
  - The optical image
  - Historical background
- Reflection and refraction 203
  - Reflection
  - The law of refraction
  - Dispersion
  - Total internal reflection
- Ray-tracing methods 205
  - Graphical ray tracing
  - Trigonometrical ray tracing
- Paraxial, or first-order, imagery 205
  - Magnification: the optical invariant
  - The Gauss theory of lenses
  - The thin lens
  - Chromatic aberration
  - Longitudinal magnification
  - Image of a tilted plane
- Optical systems 207
  - System components
  - Nonclassical imaging systems
- Lens aberrations 209
  - Seidel sums
  - Spherical aberration
  - Coma
  - Astigmatism
  - Petzval curvature
  - Distortion
- Image brightness 210
  - General relations
  - Distribution of illumination over an image
  - Visual brightness

Optics and information theory 211
- General observations 211
  - Temporal frequency response
  - Nonlinear optical systems
- Image formation 212
  - Impulse response
  - Transfer function
- Partially coherent light 213
- Optical processing 214
  - Coherent optical systems
  - Filtering
- Holography 215
  - Theory
  - Applications
- Nonlinear optics 216

Geometrical optics

GENERAL CONSIDERATIONS

The optical image. An optical image may be regarded as the apparent reproduction of an object by a lens or mirror system, employing light as a carrier. An entire image is generally produced simultaneously, as by the lens in a camera, but images may also be generated sequentially by point-by-point scanning, as in a television system or in the radio transmission of pictures across long distances in space. Nevertheless, the final detector of all images is invariably the human eye, and, whatever means is used to transmit and control the light, the final image must either be produced simultaneously or scanned so rapidly that the observer's persistence of vision will give him the mental impression of a complete image covering a finite field of view. For this to be effective the image must be repeated (as in motion pictures) or scanned (as in television) at least 40 times a second to eliminate flicker or any appearance of intermittency.

Historical background. To the ancients, the processes of image formation were full of mystery. Indeed, for a long time there was a great discussion as to whether, in vision, something moved from the object to the eye or whether something reached out from the eye to the object. By the beginning of the 17th century, however, it was known that rays of light travel in straight lines, and in 1604 Johannes Kepler, a German astronomer, published a book on optics in which he postulated that an extended object could be regarded as a multitude of separate points, each point emitting rays of light in all directions. Some of these rays would enter a lens, by which they would be bent around and made to converge to a point, the "image" of the object point whence the rays originated. The lens of the eye was not different from other lenses, and it formed an image of external objects on the retina, producing the sensation of vision.

Kepler's theory

There are two main types of image to be considered: real and virtual. A real image is formed outside the system, where the emerging rays actually cross; such an image can be caught on a screen or piece of film and is the kind of image formed by a slide projector or in a camera. A virtual image, on the other hand, is formed inside an instrument at the point where diverging rays would cross if they were extended backward into the instrument. Such an image is formed in a microscope or telescope and can be seen by looking into the eyepiece.

Kepler's concept of an image as being formed by the crossing of rays was limited in that it took no account of possible unsharpness caused by aberrations, diffraction, or even defocussing. In 1957 the Italian physicist Vasco Ronchi went the other way and defined an image as any recognizable nonuniformity in the light distribution over a surface such as a screen or film; the sharper the image, the greater the degree of nonuniformity. Today, the concept of an image often departs from Kepler's idea that an extended object can be regarded as innumerable separate points of light, and it is sometimes more convenient to regard an image as being composed of overlapping patterns of varying frequencies and contrasts; hence, the quality of

a lens can be expressed by a graph connecting the spatial frequency of a parallel line object with the contrast in the image. This concept is investigated fully under *Optics and information theory* below.

Optics had progressed rapidly by the early years of the 19th century. Lenses of moderately good quality were being made for telescopes and microscopes, and in 1841 the great mathematician Carl Friedrich Gauss published his classical book on geometrical optics. In it he expounded the concept of the focal length and cardinal points of a lens system and developed formulas for calculating the position and size of the image formed by a lens of given focal length. Between 1852 and 1856 Gauss's theory was extended to the calculation of the five principal aberrations of a lens (see below), thus laying the foundation for the formal procedures of lens design that were used for the next 100 years. Since about 1960, however, lens design has been almost entirely computerized, and the old methods of designing lenses by hand on a desk calculator are rapidly disappearing.

Lens theories

By the end of the 19th century numerous other workers had entered the field of geometrical optics, notably an English physicist, Lord Rayleigh (John William Strutt), and a German physicist, Ernst Karl Abbe. It is impossible to list all their accomplishments here. Since 1940 there has been a great resurgence in optics on the basis of information and communication theory, which is treated at length below.

Light rays, waves, and wavelets. A single point of light, which may be a point in an extended object, emits light in the form of a continually expanding train of waves, spherical in shape and centred about the point of light. It is, however, often much more convenient to regard an object point as emitting fans of rays, the rays being straight lines everywhere perpendicular to the waves. When the light beam is refracted by a lens or reflected by a mirror, the curvature of the waves is changed, and the angular divergence of the ray bundle is similarly changed in such a way that the rays remain everywhere perpendicular to the waves. When aberrations are present, a convergent ray bundle does not shrink to a perfect point, and the emerging waves are then not truly spherical.

In 1690 Christiaan Huygens, a Dutch scientist, postulated that a light wave progresses because each point in it becomes the centre of a little wavelet travelling outward in all directions at the speed of light, each new wave being merely the envelope of all these expanding wavelets. When the wavelets reach the region outside the outermost rays of the light beam, they destroy each other by mutual interference wherever a crest of one wavelet falls upon a trough of another wavelet. Hence, in effect, no waves or wavelets are allowed to exist outside the geometrical light beam defined by the rays. The normal destruction of one wavelet by another, which serves to restrict the light energy to the region of the rectilinear ray paths, however, breaks down when the light beam strikes an opaque edge, for the edge then cuts off some of the interfering wavelets, allowing others to exist, which diverge slightly into the shadow area. This phenomenon is called diffraction, and it gives rise to a complicated fine structure at the edges of shadows and in optical images (see LIGHT).

The pinhole camera. An excellent example of the working of the wavelet theory is found in the well-known pinhole camera. If the pinhole is large, the diverging geometrical pencil of rays leads to a blurred image, because each point in the object will be projected as a finite circular patch of light on the film. The spreading of the light at the boundary of a large pinhole by diffraction is slight. If the pinhole is made extremely small, however, the geometrical patch then becomes small, but the diffraction spreading is now great, leading once more to a blurred picture. There are thus two opposing effects present, and at the optimum hole size the two effects are just equal. This occurs when the hole diameter is equal to the square root of twice the wavelength (λ) times the distance (f) between the pinhole and film—*i.e.*, $\sqrt{2\lambda f}$. For $f = 100$ millimetres and $\lambda = 0.0005$ millimetre, the optimum hole size becomes 0.32 millimetre. This is not very exact, and a 0.4-millimetre hole would probably be just as good in practice. A pinhole, like a camera lens, can be regarded as having an *f*-number, which is the ratio of focal length to aperture. In this example, the *f*-number is 100/0.32 = 310, designated *f*/310. Modern camera lenses have much greater apertures, in order to achieve light-gathering power, of around *f*/1.2–*f*/5.6.

Resolution and the Airy disk. When a well-corrected lens is used in place of a pinhole, the geometrical ray divergence is eliminated by the focussing action of the lens, and a much larger aperture may be employed; in that case the diffraction spreading becomes small indeed. The image of a point formed by a perfect lens is a minute pattern of concentric and progressively fainter rings of light surrounding a central dot, the whole structure being called the Airy disk after George Biddell Airy, an English astronomer, who first explained the phenomenon in 1834. The Airy disk of a practical lens is small, its diameter being approximately equal to the *f*-number of the lens expressed in microns (0.001 millimetre). The Airy disk of an *f*/4.5 lens is therfore about 0.0045 millimetre in diameter (ten times the wavelength of blue light). Nevertheless, the Airy disk formed by a telescope or microscope objective can be readily seen with a bright point source of light if a sufficiently high eyepiece magnification is used.

Airy disk

The finite size of the Airy disk sets an inevitable limit to the possible resolving power of a visual instrument. Rayleigh found that two adjacent and equally bright stars can just be resolved if the image of one star falls somewhere near the innermost dark ring in the Airy disk of the other star; the resolving power of a lens can therefore be regarded as about half the *f*-number of the lens expressed in microns. The angular resolution of a telescope is equal to the angle subtended by the least resolvable image separation at the focal length of the objective, the light-gathering lens. This works out at about four and a half seconds of arc divided by the diameter of the objective in inches.

The Rayleigh limit. As noted above, when a perfect lens forms an image of a point source of light, the emerging wave is a sphere centred about the image point. The optical paths from all points on the wave to the image are therefore equal, so that the expanding wavelets are all in phase (vibrating in unison) when they reach the image. In an imperfect lens, however, because of the presence of aberrations, the emerging wave is not a perfect sphere, and the optical paths from the wave to the image point are then not all equal. In such a case some wavelets will reach the image as a peak, some as a trough, and there will be much destructive interference leading to the formation of a sizable patch of light, much different from the minute Airy disk characteristic of a perfectly corrected lens. In 1879 Rayleigh studied the effects of phase inequalities in a star image and came to the conclusion that an image will not be seriously degraded unless the path differences between one part of the wave and another exceed one-quarter of the wavelength of light. As this difference represents only 0.125 micron (5×10^{-6} inch), it is evident that an optical system must be designed and constructed with almost superhuman care if it is to give the best possible definition.

REFLECTION AND REFRACTION

Reflection. The use of polished mirrors for reflecting light has been known for thousands of years, and concave mirrors have long been used to form real images of distant objects. Indeed, Isaac Newton greatly preferred the use of a mirror as a telescope objective to the poor-quality lenses available in his time. Because there is no limit to the possible size of a mirror, all large telescopes today are of this type.

When a ray of light is reflected at a polished surface, the angle of reflection between ray and normal (the line at right angles to the surface) is exactly equal to the angle of incidence. It can be seen that a convex mirror forms a virtual image of a distant object, whereas a concave mirror forms a real image. A plane mirror forms a virtual image of near objects, as in the familiar looking glass. Plane mirrors are often used in instruments to bend a beam of light into a different direction.

Law of reflection

The law of refraction. When a ray of light meets the surface of separation between two transparent media, it is sharply bent or refracted. Because rays are really only directions and have no physical existence, the passage of light waves through a surface must be considered if refraction is to be understood. Refraction effects are based on the fact that light travels more slowly in a denser medium. The ratio of the velocity of light in air to its velocity in the medium is called the refractive index of the medium for light of a particular colour or wavelength. The refractive index is higher for blue light than for light at the red end of the spectrum.

In Figure 1, AA' represents a plane wave of light at the instant that A' meets the plane refracting surface $A'B$ separating two media having refractive indices n and n', respectively. During the time taken by the light to travel from A to B in material n, light travels from A' to B' in material of refractive index n', forming the new wave BB' in the second material, proceeding in direction BC. Hence, the relationship $n'/n = AB/A'B'$ can be obtained; and dividing numerator and denominator by BA' gives

$$\frac{n'}{n} = \frac{AB/BA'}{A'B'/BA'} = \frac{\sin I}{\sin I'} . \qquad (1)$$

The angles I and I' are called the angle of incidence and angle of refraction between the refracting surface and the incident and refracted waves, respectively.

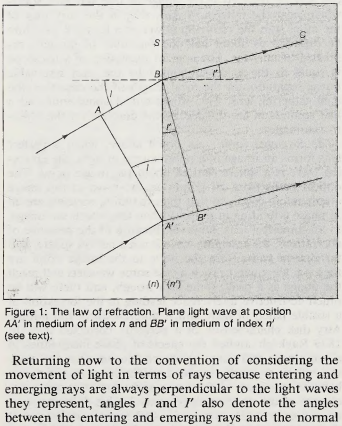


Figure 1: The law of refraction. Plane light wave at position AA' in medium of index n and BB' in medium of index n' (see text).

Returning now to the convention of considering the movement of light in terms of rays because entering and emerging rays are always perpendicular to the light waves they represent, angles I and I' also denote the angles between the entering and emerging rays and the normal (perpendicular) line to the refracting surface at B.

Equation (1), known as the law of refraction, is generally written: $n' \sin I' = n \sin I$.

Dispersion. The difference between the refractive indices of a transparent material for a specific blue light and a specific red light is known as the dispersion of the material. The usual choices of blue and red lights are the so-called "F" and "C" lines of hydrogen in the solar spectrum, named by Fraunhofer, with wavelengths 4861 and 6563 angstroms (the angstrom unit, abbreviated Å, is 10^{-8} centimetre), respectively. It is generally more significant, however, to compare the dispersion with the mean refractive index of the material for some intermediate colour such as the sodium "D" Fraunhofer line of wavelength 5893 angstroms. The dispersive power (w) of the material is then defined as the ratio of the difference between the "F" and "C" indices and the "D" index reduced by 1, or,

Dispersive power

$$w = \frac{n_F - n_C}{n_D - 1} .$$

Hundreds of different types of optical glass are currently available from manufacturers. These may be represented graphically on a plot of mean refractive index against dispersive power (Figure 2).

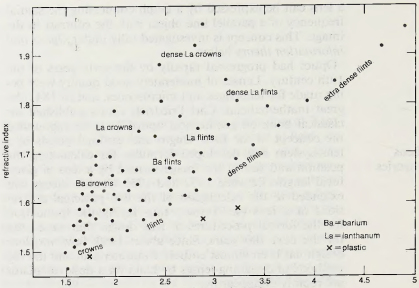


Figure 2: Relationships between refractive indices and dispersive powers of several representative optical glasses and plastics.

At first lenses were made from selected pieces of window glass or the glass used to make blown tableware. In the early 1800s, the manufacture of clear glass that was intended specifically for lenses began in Europe. The glass was slowly stirred in the molten state to remove striations and irregularities, and then the whole mass was cooled and broken up into suitable pieces for lens making. Subsequently, the pieces were placed in molds of the approximate size of the lens, slowly remelted to shape, and carefully annealed; *i.e.*, allowed to cool slowly under controlled conditions to reduce strains and imperfections. Various chemicals were added in the molten state to vary the properties of the glass: addition of lead oxide, for example, was found to raise both the refractive index and the dispersive power. In 1884 it was discovered that barium oxide had the effect of raising the refractive index without increasing the dispersion, a property that proved to be of the greatest value in the design of photographic lenses known as anastigmats (lenses devoid of astigmatic aberration). In 1938 a further major improvement was achieved by the use of various rare-earth elements, and since 1950 lanthanum glass has been commonly used in high-quality photographic lenses.

The cost of optical glass varies considerably, depending on the type of glass, the precision with which the optical properties are maintained, the freedom from internal striae and strain, the number of bubbles, and the colour of the glass. Many common types of optical glass are now available in quite large pieces, but as the specifications of the glass become more stringent the cost rises and the range of available sizes becomes limited. In a small lens such as a microscope objective or a telescope eyepiece, the cost of the glass is insignificant, but in large lenses in which every millimetre of thickness may represent an additional pound in weight, the cost of the glass can be very high indeed.

Lenses can be molded successfully of various types of plastic material, polymethyl methacrylate being the most usual. Even multi-element plastic lenses have been manufactured for low-cost cameras, the negative (concave) elements being made of a high-dispersion plastic such as styrene.

Plastic lenses

Total internal reflection. When a ray of light emerges obliquely from glass into air, the angle of refraction between ray and normal is greater than the angle of incidence inside the glass, and at a sufficiently high obliquity the angle of refraction can actually reach 90°. In this case the emerging ray travels along the glass surface, and the sine of the angle of incidence inside the glass, known as the critical angle, is then equal to the reciprocal of the refractive index of the material. At angles of incidence greater than the critical angle, the ray never emerges, and total internal reflection occurs, for there is no measurable

loss if the glass surface is perfectly clean. Dirt or dust on the surface can cause a small loss of energy by scattering some light into the air.

Light is totally internally reflected in many types of reflecting prism and in fibre optics, in which long fibres of high-index glass clad with a thin layer of lower index glass are assembled side-by-side in precise order. The light admitted into one end of each fibre is transmitted along it without loss by thousands of successive internal reflections at the interlayer between the glass and the cladding. Hence, an image projected upon one end of the bundle will be dissected and transmitted to the other end, where it can be examined through a magnifier or photographed. Many modern medical instruments, such as cystoscopes and bronchoscopes, depend for their action on this principle. Single thick fibres (actually glass rods) are sometimes used to transmit light around corners to an otherwise inaccessible location.

RAY-TRACING METHODS

Graphical ray tracing. In 1621 Willebrord Snell, a professor of mathematics at Leiden, discovered a simple graphical procedure for determining the direction of the refracted ray at a surface when the incident ray is given. The mathematical form of the law of refraction, equation (1) above, was announced by the French mathematician René Descartes some 16 years later.

Snell's construction is as follows: The line AP in Figure 3A represents a ray incident upon a refracting surface at P, the normal at P being PN. If the incident and refracted rays are extended to intersect any line SS parallel to the normal, the lengths PQ and PR along the rays will be

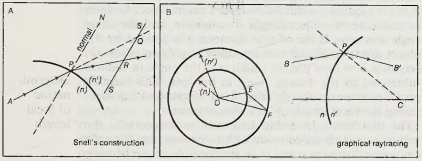


Figure 3: Graphic refraction procedures (see text).

Snell's construction

proportional to the refractive indices n and n'. Hence, if PQ and the indices are known, PR can be found and the refracted ray drawn in.

A convenient modification of Snell's construction can readily be used to trace the path of a ray through a complete lens. In Figure 3B, the incident ray BP strikes a refracting surface at P. The normal to the surface is PC. At any convenient place on the page two concentric circles are drawn about a point O with radii proportional to the refractive indices n and n', respectively. A line OE is now drawn parallel to the incident ray BP extending as far as the circle representing the refractive index n of the medium containing the incident ray. From E a line is drawn parallel to the normal PC extending to F on the circle representing the refractive index n'. The line OF then represents the direction of the desired refracted ray, which may be drawn in at PB'. This process is repeated successively for all the surfaces in a lens. If a mirror is involved, the reflected ray may be found by drawing the normal line EF across the circle diagram to the incident-index circle on the other side.

Trigonometrical ray tracing. No graphical construction can possibly be adequate to determine the aberration residual of a corrected lens, and for this an accurate trigonometrical computation must be made and carried out to six or seven decimal places, the angles being determined to single seconds of arc or less. There are many procedures for calculating the path of a ray through a system of spherical refracting or reflecting surfaces, the following being typical: The diagram in Figure 4 represents a ray lying in the meridian plane, defined as the plane containing the lens axis and the object point. A ray in this plane is defined by its slope angle, U, and by the length of the

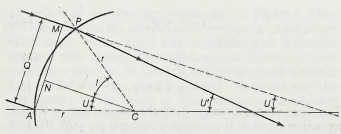


Figure 4: Trigonometrical ray tracing (see text).

perpendicular, Q, drawn from the vertex (A) of the surface on to the ray. By drawing a line parallel to the incident ray through the centre of curvature C, to divide Q into two parts at N, the relation is stated as $AN = r \sin U$, and $NM = r \sin I$. Hence

$$Q = r(\sin U + \sin I). \tag{2}$$

From this the first ray-tracing equation can be derived,

$$\sin I = \frac{Q}{r} - \sin U. \tag{3a}$$

Applying the law of refraction, equation (2), gives the second equation

$$\sin I' = \frac{n}{n'} - \sin I. \tag{3b}$$

Because the angle $PCA = U + I = U' + I'$, the slope of the refracted ray can be written as

$$U' = U + I - I'; \tag{3c}$$

and, lastly, by adding primes to equation (2),

$$Q' = r(\sin U' + \sin I').$$

Having found the Q' of the refracted ray, transfer to the next surface can be performed by

$$Q_2 = Q_1' - d \sin U_1',$$

in which d is the axial distance from the first to the second refracting surface. After performing this calculation for all the surfaces in succession, the longitudinal distance from the last surface to the intersection point of the emergent ray with the lens axis is found by

$$L' = \frac{Q'}{\sin U'}.$$

Corresponding but much more complicated formulas are available for tracing a skew ray, that is, a ray that does not lie in the meridian plane but travels at an angle to it. After refraction at a surface, a skew ray intersects the meridian plane again at what is called the diapoint. By tracing the paths of a great many (100 or more) meridional and skew rays through a lens, with the help of an electronic computer, and plotting the assemblage of points at which all these rays pierce the focal plane after emerging from the lens, a close approximation to the appearance of a star image can be constructed, and a good idea of the expected performance of a lens can be obtained.

PARAXIAL, OR FIRST-ORDER, IMAGERY

Paraxial rays

In a lens that has spherical aberration, the various rays from an axial object point will in general intersect the lens axis at different points after emerging into the image space. By tracing several rays entering the lens at different heights (*i.e.*, distances from the axis) and extrapolating from a graph connecting ray height with image position, it would be possible to infer where a ray running very close to the axis (a paraxial ray) would intersect the axis, although such a ray could not be traced directly by the ordinary trigonometrical formulas because the angles would be too small for the sine table to be of any use. Because the sine of a small angle is equal to the radian measure of the angle itself, however, a paraxial ray can be traced by reducing the ray-tracing formulas to their limiting case for small angles and thus determining the paraxial intersection point directly. When this is done, writing paraxial-ray data with lowercase letters, it is found that the Q and Q' above both become equal to the height of incidence y, and the formulas (3a), (3b), and (3c) become, in the paraxial limit:

$$i = \frac{y}{r'} - u \tag{4a}$$

$$i' = \frac{n}{n'} i \tag{4b}$$

$$u' = u + i - i'. \tag{4c}$$

The longitudinal distance from the last surface to the intersection point of the emerging paraxial ray with the lens axis becomes $l' = y/u'$.

Because all paraxial rays from a given object point unite at the same image point, the resulting longitudinal distance (l') is independent of the particular paraxial ray that is traced. Any nominal value for the height of incidence, y, may therefore be adopted, remembering that it is really an infinitesimal and y is only its relative magnitude. Thus, it is clear that the paraxial angles in equation (4) are really only auxiliaries, and they can be readily eliminated, giving the object–image distances for paraxial rays:

$$n'(l' - r)u' = n(l - r)u \tag{5}$$

and

$$\frac{n'}{l'} = \frac{n}{l} + \frac{n' - n}{r}. \tag{6}$$

Magnification: the optical invariant. It is frequently as important to determine the size of an image as it is to determine its location. To obtain an expression for the magnification—that is, the ratio of the size of an image to the size of the object—the following process may be used: If an object point B lies to one side of the lens axis at a transverse distance h from it, and the image point B' is at a transverse distance h', then B, B', and the centre of curvature of the surface, C, lie on a straight line called the auxiliary axis. Then, by simple proportion,

$$m = \frac{h'}{h} = \frac{l' - r}{l - r} = \frac{nu}{n'u'}.$$

Hence,

$$h'n'u' = hnu, \tag{7}$$

and the product (hnu) is invariant for all the spaces between the lens surfaces, including the object and image spaces, for any lens system of any degree of complexity. This theorem has been named after the French scientist Joseph-Louis Lagrange, although it is sometimes called the Smith-Helmholtz theorem, after Robert Smith, an English scientist, and Hermann Helmholtz, a German scientist; the product (hnu) is often known as the optical invariant. As it is easy to determine the quantities h, n, and u for the original object, it is only necessary to calculate u' by tracing a paraxial ray in order to find the image height h' for any lens. If the lens is used in air, as most lenses are, the refractive indices are both unity, and the magnification becomes merely $m = u/u'$.

Lagrange theorem

The Gauss theory of lenses. In 1841 Gauss published a now famous treatise on optics in which he demonstrated that, so far as paraxial rays are concerned, a lens of any degree of complexity can be replaced by two principal, or nodal, points and two focal points, the distances from the principal points to their respective focal points being the focal lengths of the lens, and, furthermore, that the two focal lengths are equal to one another when the refractive indices of object and image spaces are equal, as when a lens is used in air.

The principal and focal points may be defined as follows: Figure 5 shows a lens system of any construction, with a bundle of rays entering from the left in a direction parallel to the lens axis. After refraction by the lens each ray will cross the axis at some point, and the entering and emerging portions of each ray are then extended until they intersect at a point such as Q. The locus of all the points Q is a surface of revolution about the lens axis known as the equivalent refracting locus of the lens. The point where this locus crosses the axis is called the principal point, P_2, and the central portion of the locus in the neighbourhood of the axis, which is virtually a plane perpendicular to the axis, is called the principal plane. The point where the emerging paraxial ray crosses the axis is called the focal point F_2, the distance from P_2 to F_2 being the (posterior) focal length f'. A similar situation exists for a parallel beam of light entering from the right, giving the anterior principal point P_1, the anterior focal point F_1, and the front focal length f. For a lens in air it can be shown that the two focal lengths are equal in magnitude but opposite in direction—*i.e.*, if F_2 is to the right of P_2, then F_1 must lie to the left of P_1, as in the case of an ordinary positive lens (one that gives a real image). In a negative lens (one that gives a virtual image), F_2 lies to the left of P_2, and the posterior focal length f' is negative.

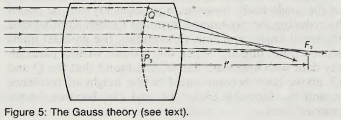


Figure 5: The Gauss theory (see text).

The relation between the distances of object and image from a lens can be easily stated if the positions of the two principal points and the two focal points are known. (In using these expressions, distances are considered positive or negative depending on whether they are measured to the right or to the left from their respective origins.) For a lens in air: (a) If the conjugate distances measured from the respective focal points are x and x', and if m is the image magnification (height of image divided by height of object), then $m = -x'/f' = f'/x$ and $xx' = -f'^2$. (b) If the conjugate distances measured from the respective principal points are p and p' and if m is the image magnification, then $m = p'/p$ and $1/p' = 1/p + 1/f'$. The Lagrange equation (7) requires modification for a distant object because in that case the object height h is infinite, and the slope angle u is zero. If the off-axis distance h is divided by the object distance L, and u is multiplied by L, equation (7) becomes $h' = (n/n')\varphi$, in which φ is the angle in radians subtended by the distant object at the lens. This formula provides a means for defining focal length and for measuring the focal length of an unknown lens.

Determination of focal length

The thin lens. In a thin lens such as a spectacle, the two principal planes coincide within the lens, and then the conjugate distances p and p' in the formula above become the distances of object and image from the lens itself.

The focal length of a thin lens can be computed by applying the surface-conjugate formula (6) to the two surfaces in succession, writing the l of the first surface as infinity and the l of the second surface equal to the l' of the first surface. When this is done, the lens power (P) becomes

$$P = \frac{1}{f'} = (n - 1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right).$$

If a number of thin lenses are assembled together in contact, assuming that the combination is still virtually a thin lens, then the power ($1/f'$) of the combination will be equal to the sum of the powers of the separate lenses. Thus, if a thin negative lens is placed in close contact with a thin positive lens of equal power, the combination will have zero power and will no longer act like a lens. This is the basis of a process used by optometrists to ascertain the power of an unknown lens by combining it with one of a series of known lenses from a trial case.

Chromatic aberration. Because the refractive index of glass varies with wavelength, every property of a lens that depends on its refractive index also varies with wavelength, including the focal length, the image distance, and the image magnification. The change of image distance with wavelength is known as chromatic aberration, and the variation of magnification with wavelength is known as chromatic difference of magnification, or lateral colour. Chromatic aberration can be eliminated by combining a strong lens of low-dispersion glass (crown) with a weaker lens made of high-dispersion (flint) glass. Such a combination is said to be achromatic. This method of removing chromatic aberration was discovered in 1729 by Chester Hall, an English inventor, and it was exploited vigorously in the late 18th century in numerous small telescopes. Chromatic variation of magnification can be eliminated by

achromatizing all the components of a system or by making the system symmetrical about a central diaphragm. Both chromatic aberration and lateral colour are corrected in every high-grade optical system.

Longitudinal magnification. If an object is moved through a short distance δp along the axis, then the corresponding image shift $\delta p'$ is related to the object movement by the longitudinal magnification ($\overline{m}$). Succinctly,

$$\overline{m} = \delta p'/\delta p = m^2 ,$$

in which m is the lateral magnification. The fact that the longitudinal magnification is equal to the square of the transverse magnification means that m is always positive; hence, if the object is moved from left to right, the image must also move from left to right. Also, if m is large, then $\overline{m}$ is very large, which explains why the depth of field (δp) of a microscope is extremely small. On the other hand, if m is small, less than one as in a camera, then $\overline{m}$ is very small, and all objects within a considerable range of distances (δp) appear substantially in focus.

Limitation on depth of field

Image of a tilted plane. If a lens is used to form an image of a plane object that is tilted relative to the lens axis, then the image will also be tilted in such a way that the plane of the object, the plane of the image, and the median plane of the lens all meet. This construction can be derived by the use of the lateral and longitudinal magnification relations just established above. With a tilted object the magnification at any point is given by the ratio of the distances of image and object from the lens at that point in the image, and, consequently, m varies progressively from one end of the image to the other. This arrangement is frequently used in view cameras equipped with "swings" to increase depth of field and in enlargers to rectify the convergence of parallel lines caused by tilting the camera, for example, in photographing tall buildings. The rule finds extensive application in photogrammetry and in the making of maps from aerial photographs.

OPTICAL SYSTEMS

System components. An optical system consists of a succession of elements, which may include lenses, mirrors, light sources, detectors, projection screens, reflecting prisms, dispersing devices, filters and thin films, and fibre-optics bundles.

Lenses. All optical systems have an aperture stop somewhere in the system to limit the diameter of the beams of light passing through the system from an object point. By analogy with the human eye, this limiting aperture stop is called the iris of the system, its images in the object and image spaces being called the entrance pupil and exit pupil, respectively. In most photographic lenses the iris is inside the objective, and it is often adjustable in diameter to control the image illumination and the depth of field. In telescope and microscope systems the cylindrical mount of the objective lens is generally the limiting aperture or iris of the system; its image, formed behind the eyepiece where the observer's eye must be located to see the whole area being observed, called the field, is then the exit pupil.

Entrance and exit pupils

The pupils of a lens system can be regarded as the common bases of oblique beams passing through the system from all points in an extended object. In most systems, however, the mounts of some of the lens elements cut into the oblique beams and prevent the beams from being perfectly circular, and the pupils are then not fully filled with light. This effect is known as vignetting and leads to a reduction in illumination in the outer parts of the field of view.

A common feature of many optical systems is a relay lens, which may be introduced to invert an image or to extend the length of the system, as in a military periscope. An example of the use of a relay lens is found in the common rifle sight shown diagrammatically in Figure 6. Here the front lens A is the objective, forming an inverted image of the target on the cross wire or reticle at B. The light then proceeds to the relay lens C, which forms a second image, now erect, at D. Beyond this image is the eyepiece E to render the light parallel so that the image may be seen sharply by the observer. Unfortunately, the oblique beam from the objective will usually miss the relay lens, and so a field lens must be inserted at or near the first image B to bend the oblique beams around and redirect them toward the relay lens. The power of the field lens is chosen so that it will form an image of the objective lens aperture on the relay lens aperture. The iris and entrance pupil of this system coincide at the objective; there is an internal pupil at the relay lens, and the exit pupil lies beyond the eyepiece as shown in Figure 6.

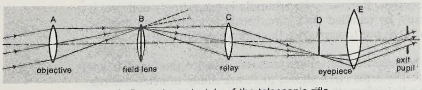


Figure 6: Operating principle of the telescopic rifle sight (see text).

Mirrors. Mirrors are frequently used in optical systems. Plane mirrors may be employed to bend a beam of light in another direction, either for convenience or to yield an image reversed left for right if required. Curved mirrors, concave and convex, may be used in place of lenses as image-forming elements in reflecting telescopes. All of the world's largest telescopes and many small ones are of the reflecting type. Such telescopes use a concave mirror to produce the main image, a small secondary mirror often being added to magnify the image and to place it in a convenient position for observation or photography. Telescope mirrors are commonly made parabolic or hyperbolic in section to correct the aberrations of the image. Originally telescope mirrors were made from polished "speculum metal," an alloy of copper and tin, but in 1856 Justus von Liebig, a German chemist, invented a process for forming a mirror-like layer of silver on polished glass, which was applied to telescope mirrors by the German astronomer C.A. von Steinheil. Today most mirrors are made of glass, coated with either a chemically deposited silver layer or more often one made by depositing vaporized aluminum on the surface. The aluminum surface is as highly reflective as silver and does not tarnish as readily.

Use in reflecting telescopes

A large astronomical mirror presents many problems to the optical engineer, mainly because even a distortion of a few microns of the mirror under its own weight will cause an intolerable blurring of the image. Though many schemes for supporting a mirror without strain have been tried, including one to support it on a bag of compressed air, the problem of completely eliminating mirror distortion remains unsolved. A metal mirror, if well ribbed on the back, may be lighter than a glass mirror and therefore easier to handle, but most metals are slightly flexible and require just as careful support as glass mirrors. Since temperature changes can also cause serious distortion in a mirror, astronomers try to hold observatory temperatures as constant as possible.

Light sources. Many types of optical instruments form images by natural light, but some, such as microscopes and projectors, require a source of artificial light. Tungsten filament lamps are the most common, but if a very bright source is required, a carbon or xenon arc is employed. For some applications, mercury or other gas discharge tubes are used; a laser beam is often employed in scientific applications. Laser light is brilliant, monochromatic, collimated (the rays are parallel), and coherent (the waves are all in step with each other), any or all of these properties being of value in particular cases.

Detectors. The image formed by an optical system is usually received by the eye, which is a remarkably adaptable and sensitive detector of radiation within the visible region of the electromagnetic spectrum (see LIGHT). A photographic film, another widely used detector, has the advantage of yielding a permanent record of events. Since about 1925 many types of electrical detectors of radiation, both within the visible region and beyond it, have been developed. These include photoelectric cells of various kinds in which either a voltage or a resistance is modified by light falling on the device. Many new types of detectors are sensitive far into the infrared spectrum and are used to detect the heat radiated by a flame or other hot object. A number of image intensifiers and converters, particularly for X-ray or infrared radiation, which have appeared

since World War II, embody a radiation detector at one end of a vacuum tube and an electron lens inside the tube to relay the image on to a phosphor screen at the other end. This arrangement produces a visible picture that may be observed by eye or photographed to make a permanent record.

Television camera tubes detect real images by electronic scanning, the picture on the viewing tube being a replica of the image in the original camera. The combined application of electronics and optics has become common (see below). An extreme example of electro-optics appears in some space cameras, in which the film is exposed, processed, and then scanned by a tiny point of light; the light passing through the film is picked up by a photocell and transmitted to Earth by radio, where it is made to control the brightness of another point of light scanning a second piece of film in exact synchronism with the scanning spot in the camera. The whole system thus produces a picture on Earth that is an exact replica of the picture photographed in space a few minutes earlier.

Projection screens. The simplest screen for the projection of slides or motion pictures is, of course, a matte white surface, which may be on a hard base as in outdoor theatres or on a stretched cloth indoors. A theatre screen is often perforated to transmit sound from loudspeakers placed behind it.

Improved screen materials have been developed to increase the brightness of the picture to suit the particular shape of the auditorium. A screen covered with tiny beads tends to send the light back in the general direction of the projector, and is suitable for use at one end of a long, narrow auditorium. Another type of screen is covered with fine embossed vertical grooves; this tends to distribute the light in a horizontal band across the audience with little or no vertical spread. A real advantage of these highly reflective screens is that they tend to reflect ambient room light away from the viewer as by a mirror, so that the pictures appear almost as bright and clear by day as in a darkened room.

Reflecting prisms. Reflecting prisms are pieces of glass bounded by plane surfaces set at carefully specified angles. Some of these surfaces transmit light, some reflect light, while some serve both functions in succession. A prism is thus an assembly of plane reflectors at relatively fixed angles, which are traversed in succession by a beam of light.

The simplest prism is a triangular block of glass with two faces at right angles and one at an angle of 45°. The face at 45° deflects a beam of light through a right angle. The common Porro prism used in a pair of binoculars contains four 45° reflecting surfaces, two to reverse the beam direction in the vertical plane and two in the horizontal plane (Figure 7). These reflecting faces could be replaced by pieces of mirror mounted on a metal frame,

Prism binoculars

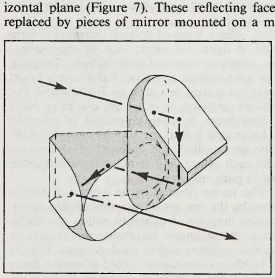


Figure 7: Porro prism.

but it is hard to hold mirrors rigidly and harder still to keep them clean. Some microscopes are equipped with a 45° deflection prism behind the eyepiece; this prism may provide two or three reflections depending on the type of image inversion or left-for-right reversal required.

Prisms containing a semireflecting, semitransmitting surface are known as beam splitters and as such have many uses. An important application is found in some colour television cameras, in which the light from the lens is divided by two beam splitters in succession to form red, green, and blue images on the faces of three image tubes in the camera.

Dispersing devices. There are two forms of dispersing element used to spread out the constituent colours of a beam of light into a "spectrum," namely a prism and a grating. The prism, known to Newton, is the older; it separates the colours of the spectrum because the refractive index of the glass is lowest for red light and progressively increases through the yellow and green to the blue, where it is highest. Prism spectroscopes and spectrographs are made in a variety of forms and sizes, but in all cases the blue end of the spectrum is greatly spread out while the red end is relatively compressed.

Prism and grating

A diffraction grating is a ruled mirror or transparent plate of glass having many thousands of fine parallel grooves to the inch. It separates the colours of the spectrum by a process of diffraction. Each groove diffracts, or scatters, light in all directions, and in the case of light of one particular wavelength, there will be one direction in which the light wave from one groove lags behind the light wave from the next groove by precisely one or more whole wavelengths. This results in a strong beam of diffracted light in that direction and darkness in all other directions. Since each spectral colour corresponds to a different wavelength, the grating spreads out the spectrum into a fan where it can be observed or photographed. The red rays are bent most and the blue rays least, the opposite of the situation with a prism.

Although a prism or grating is the essential dispersing element in a spectrograph, a fine slit and additional lenses or focussing mirrors must be used to form a sharply defined spectrum. Prism spectroscopes are, of course, limited to those wavelengths for which the prism material is transparent; a reflecting grating can be used for any wavelength that the material will reflect.

Filters and thin films. A colour filter is a sheet of transparent material that modifies a light beam by selective absorption of some colours in relation to others. A neutral filter absorbs all wavelengths equally and merely serves to reduce the intensity of a beam of light without changing its colour.

Filters may be made from sheets of coloured glass, plastic, or dyed gelatin, and in some cases glass cells filled with liquid have been used. Since World War II, another type of filter depending on the interference of light has been developed in which one or more metallic or other types of films of controlled thickness have been deposited on a glass plate, the layers being so thin as to cause selective interference of some wavelengths in relation to others and thus act as a nonabsorbing filter. In this case the rejected colours are reflected instead of being absorbed.

Polarizing filters have the property of transmitting light that vibrates in one direction while absorbing light that vibrates in a perpendicular direction. These filters are used extensively in scientific instruments. In sunglasses and when placed over a camera lens, polarizing filters reduce unwanted reflections from nonmetallic surfaces. Polarizing spectacles have been used to separate the left-eye and right-eye beams in the projection of stereoscopic pictures or movies.

Fibre-optics bundles. As noted earlier, a thin rod or fibre of glass or other transparent material transmits light by repeated internal reflections, even when the rod is somewhat curved. An ordered bundle of rods or fibres is thus capable of taking an image projected upon one end of the bundle and reproducing it at the other end. A fibre-optics bundle can be fused together into a rigid channel, or it may be left flexible, only the ends being rigidly fastened together. Because a fibre bundle is exceedingly delicate, it must be handled with care; breaking a fibre would cause a black dot to appear in the reproduced image.

(R.K./Ed.)

Nonclassical imaging systems. Besides the familiar optical systems cited above, there are many nonclassical optical elements that are used to a limited extent for special purposes. The most familiar of these is the aspheric (non-

Aspheric surfaces

spherical) surface. Because plane and spherical surfaces are the easiest to generate accurately on glass, most lenses contain only such surfaces. It is occasionally necessary, however, to use some other axially symmetric surface on a lens or mirror, generally to correct a particular aberration. An example is the parabolic surface used for the primary mirror of a large astronomical telescope; another is the elliptic surface molded on the front of the little solid glass reflector units used on highway signs.

Another commonly used optical surface is the side of a cylinder. Such surfaces have power only in the meridian perpendicular to the cylinder axis. Cylindrical lenses are therefore used wherever it is desired to vary the magnification from one meridian to a perpendicular meridian. Cylindrical surfaces are employed in the anamorphic lenses used in some wide-screen motion-picture systems to compress the image horizontally in the camera and stretch it back to its original shape in the projected image.

To correct astigmatism in the eye, many spectacles are made with toric surfaces—*i.e.*, with a stronger curvature in one meridian than in the perpendicular meridian, like the bowl of a teaspoon. These surfaces are generated and polished by special machines and are made by the million every year.

Another nonclassical optical system is the bifocal or trifocal spectacle lens. They are made either by forming two or three separate surfaces on a single piece of glass or obtaining additional power by fusing a piece of high-index glass on to the front of the main lens and then polishing a single spherical surface over both glasses.

Two French scientists, Georges-Louis Buffon and Augustin-Jean Fresnel, in the 18th century suggested forming a lens in concentric rings to save weight, each ring being a portion of what would normally be a continuous spherical surface but flattened out. On a large scale, Fresnel lenses have been used in lighthouses, floodlights, and traffic signals, and as cylindrical ship's lanterns. With fine steps a few thousandths of an inch wide, molded plastic Fresnel lenses are often used as condensers in overhead projectors and in cameras as a field lens in contact with a ground-glass viewing screen.

Lenses have occasionally been made with one surface taking the form of a flattened cone. Such lenses produce a long, linear image of a point source, lying along the lens axis; for this reason they are commonly referred to as axicons. They have been used to produce a straight line of light in space for aligning machines and shafting, but since about 1965 the beam from a gas laser has generally been used instead.

Axicons

LENS ABERRATIONS

Seidel sums. If a lens were perfect and the object were a single point of monochromatic light, then, as noted above, the light wave emerging from the lens would be a portion of a sphere centred about the ideal image point, lying in the paraxial image plane at a height above the axis given by the Lagrange theorem. In practice, however, this condition is most unlikely to occur; it is much more probable that the emerging wave will depart slightly from a perfect sphere, the departure varying from point to point over the lens aperture. This departure is extremely small, being of the order of the wavelength of light that is only half a micron, so it would be impossible to show this departure on a drawing. It can be represented mathematically, however, in the following way: The coordinates of a point in the exit-pupil aperture will be represented by x_0 and y_0, the y_0 coordinate lying in the meridian plane containing the object point and the lens axis. The departure of the wave from the ideal sphere is generally called OPD, meaning optical path difference. It can be shown that OPD is related to x_0 and y_0 by five constants S_1 through S_5, and the quantity h'_0.

$$\text{OPD} = S_1(x_0^2 + y_0^2)^2 + S_2 y_0(x_0^2 + y_0^2)h'_0 + \\ + S_3(x_0^2 + 3y_0^2)h'^2_0 + S_4(x_0^2 + y_0^2)h'^2_0 + S_5 y_0 h'^3_0 .$$

Each of these five terms is considered to be a separate "aberration," the coefficients $S_1, \ldots S_5$ being called Seidel sums after the 19th-century German scientist L.P. Seidel, who identified the imperfections. These aberrations are respectively spherical, coma, astigmatism, Petzval field curvature, and distortion. The symbol h'_0 refers to the height of the final image point above the lens axis, and hence it defines the obliquity of the beam.

Calculation of Seidel sums

The five Seidel sums can be calculated by tracing a paraxial ray from object to image through the lens and by tracing also a paraxial principal ray from the centre of the aperture stop outward in both directions toward the object and image, respectively. The angle of incidence i and the ray slope angle u of each of these paraxial rays at each surface are then listed and inserted into the following expressions for the five sums. The angle u'_0 represents the final emerging slope of the paraxial ray.

The calculation starts by determining the radius A of the exit pupil by $A = \sqrt{x_0^2 + y_0^2}$ and also the quantity K at each surface by

$$K = \tfrac{1}{2} y\, n \left(\frac{n}{n'} - 1\right)(i - u').$$

The corresponding K_{pr} for the paraxial principal ray is also determined at each surface. Then, the five aberrations may be written

$$S_1 = \frac{1}{4A_4} \sum K i^2 \qquad S_2 = \frac{1}{A^2 h'^2_0} \sum K i i_{pr}$$

$$S_3 = \frac{1}{2A^2 h'^2_0} \sum K i_{pr}^2 \qquad S_4 = \frac{u'^2_0}{4A^2} \sum \frac{n - n'}{nn'\, r}$$

$$S_5 = \frac{1}{A h'^3_0} \sum \left[K_{pr} i i_{pr} + \tfrac{1}{2} h'_0 u'^2_0 (u'^2_{pr} - u_{pr}^2) \right].$$

To interpret these aberrations, the simplest procedure is to find the components x', y' of the displacement of a ray from the Lagrangian image point in the paraxial focal plane, by differentiating the OPD expression given above. The partial derivatives $\partial\text{OPD}/\partial x_0$ and $\partial\text{OPD}/\partial y_0$ represent respectively the components of the slope of the wave relative to the reference sphere at any particular point (x_0, y_0). Hence, because a ray is always perpendicular to the wave, the ray displacements in the focal plane can be found by

$$x' = f \frac{\partial \text{OPD}}{\partial x_0} \quad \text{and} \quad y' = f \frac{\partial \text{OPD}}{\partial y_0},$$

in which f is the focal length of the lens. The aggregation of rays striking the focal plane will indicate the kind of image that is characteristic of each aberration.

This procedure will be applied to each of the five aberration terms separately, assuming that all the other aberrations are absent. Obviously, in a perfect lens x' and y' are zero because OPD is zero. It must be remembered, however, that by using rays instead of waves, all fine-structure

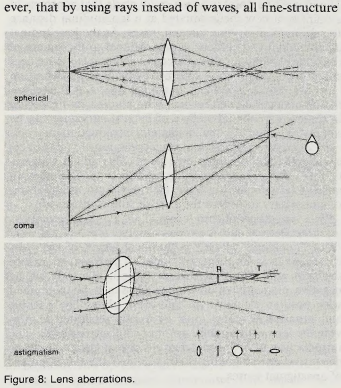


Figure 8: Lens aberrations.

effects caused by diffraction will be lost, and only the macroscopic image structure will be retained.

Spherical aberration. The first term in the OPD expression is $\text{OPD} = S_1(x_0^2 + y_0^2)^2$. Hence

$$x' = f\frac{\partial \text{OPD}}{\partial x_0} = 4fA^2S_1 \cdot x_0 \quad \text{and} \quad y' = f\frac{\partial \text{OPD}}{\partial y_0} = 4fA^2S_1 \cdot y_0 .$$

These displacements can both be eliminated simultaneously by applying a longitudinal shift L to the focal plane. This changes x' by $-Lx_0/f$ and y' by $-Ly_0/f$; hence, if L is made equal to $4f^2A^2S_1$, both ray displacements vanish. The aberration, therefore, represents a condition in which each zone of the lens has a different focus along the axis, the shift of focus from the paraxial image being proportional to A^2. This is known as spherical aberration (see Figure 8).

Coma. The S_2 term in the OPD expression represents the aberration called coma, in which the image of a point has the appearance of a comet. The x' and y' components are as follows:

$$x' = fh_0'S_2(2x_0y_0)$$
$$y' = fh_0'S_2(x_0^2 + 3y_0^2).$$

When this aberration is present, each circular zone of the lens forms a small ringlike image in the focal plane, the rings formed by successive concentric zones of the lens fitting into two straight envelope lines at 60° to each other (Figure 8). Because the brightness of this image is greatest at the tip, coma tends to form a one-sided haze on images in the outer parts of the field.

Astigmatism. If only the S_3 term is present, then

$$x' = 2fh_0'^2S_3(x_0)$$
$$y' = 2fh_0'^2S_3(3y_0).$$

For any one zone of the lens, x' and y' describe a vertical ellipse with major axis three times the minor axis. The images formed by all the smaller zones of the lens fit into this ellipse and fill it out with a uniform intensity of light. If the image plane is moved along the axis by a distance L, as in focussing a camera, then, at $L = 2f^2h_0'^2S_3$, the ellipse shrinks to a radial focal line (R). Twice this displacement yields a circle; three times this L gives a tangential focal line (T), which is followed by an ellipse with its major axis in the x direction, as in Figure 8, bottom. The usual effect of astigmatism in an image is the appearance of radial or tangential blurring in the outer parts of the field.

Effect of astigmatism

Petzval curvature. For the S_4 term taken alone,

$$x' = 2fh_0'^2S_4 \cdot x_0$$
$$y' = 2fh_0'^2S_4 \cdot y_0.$$

The image of a point is now a small circle that contracts to a point at a new focus situated at a longitudinal distance $L = 2f^2h_0'^2S_4$ from the paraxial image. As the longitudinal displacement of the focus is proportional to the square of the image height h_0', this aberration represents a pure field curvature without any accompanying loss of definition (all lines remain sharp). It is named after the Hungarian mathematician József Petzval, who studied its properties in the early 1840s. The effect of Petzval curvature can be somewhat offset by the deliberate introduction of sufficient overcorrected astigmatism, as was done in all the preanastigmat photographic objectives. This added astigmatism is, of course, undesirable, and in order to design an anastigmat lens having a flat field free from astigmatism, it is necessary to reduce the Petzval sum S_4 drastically.

For a succession of thin lenses (1, 2, 3, . . . etc.) in a system, the Petzval sum becomes simply $1/f_1n_1 + 1/f_2n_2 + 1/f_3n_3 + \ldots$ etc., in which f is the focal length of each element and n is its refractive index. Therefore, to reduce the sum and minimize this aberration, relatively strong negative elements of low-index glass can be combined with positive elements of high-index glass. The positive and negative elements must be axially separated to provide the lens with a useful amount of positive power. The introduction of high-index barium crown glass with a low dispersive power in the 1880s initiated the development of anastigmat lenses.

Distortion. For the S_5 aberration,

$$x' = 0$$
$$y' = fh_0'^3S_5 .$$

When this aberration is present, the entire image point is displaced toward or away from the axis by an amount proportional to the third power of the transverse distance h_0' of the image from the axis. This leads to the formation of an image of a square that is either barrel-shaped or a cushion-shaped figure.

It is to be noted that the five Seidel aberrations represent the largest and most conspicuous defects that can arise in an uncorrected optical system. Even in the best lenses in which these five aberrations have been perfectly corrected for one zone of the lens and for one point in the field, however, there will exist small residuals of these aberrations and of many other higher order aberrations also, which are significantly different from the classical types just described. The typical aberration figures shown in Figure 8 are, of course, grossly exaggerated, and actually it requires some magnification of a star image to render these appearances clearly visible. Nevertheless, they are important enough to require drastic reduction in high-quality lenses intended to make sharp negatives capable of considerable enlargement.

Other aberrations

IMAGE BRIGHTNESS

General relations. All photometric concepts are based on the idea of a standard candle, lamps having accurately known candle power being obtainable from the various national standards laboratories. The ratio of the candle power of a source to its area is called the luminance of the source; luminances range from about 2,000 candles per square millimetre at the surface of the Sun down to about 3×10^{-6} candle per square centimetre (3×10^{-6} stilb) for the luminous paint on a watch dial. Ordinary outdoor scenes in daylight have an average luminance of several hundred candles per square foot. The quantity of light flux flowing out from a source is measured in lumens, the lumen being defined as the amount of flux radiated by a small "point" source of one candle power into a cone having a solid angle of one steradian. When light falls upon a surface it produces illumination (*i.e.*, illuminance), the usual measure of illuminance being the footcandle, which is one lumen falling on each square foot of receiving surface.

It is often important to be able to calculate the brightness of an image formed by an optical system, because photographic emulsions and other light receptors cannot respond satisfactorily if the light level is too low. The problem is to relate the luminance of an object with the illuminance in the image, knowing the transmittance and aperture of the optical system. A small area A of a plane object having a luminance of B candles per square unit will have a normal intensity of AB candles. This source radiates light into a cone of semi-angle U, limited, for example, by the rim of a lens. The light flux (F) entering the cone can be found by integration to be

$$F = \pi AB \sin^2 U \text{ lumens.}$$

If the object luminance is expressed as B_L lamberts, the lambert being an alternative luminance unit equal to $1/\pi$ (*i.e.*, 0.32) candle per unit area, the flux (F) is

$$F = AB_L \sin^2 U \text{ lumens,}$$

because there are π times as many lamberts in a given luminance as there are candles per unit area.

A fraction t of this flux finds its way to the image, t being the lens transmittance, generally about 0.8 or 0.9 but less if a mirror is involved. The area of the image is Am^2, in which m, the magnification, is given by

$$m = \frac{\sin U}{\sin U'} .$$

Hence, the image illuminance (E) is

$$E = (t\pi AB \sin^2 U) \div (A \sin^2 U/\sin^2 U') = t\pi B \sin^2 U' \quad (8)$$

or $E = tB_L \sin^2 U'$.

The image illuminance thus depends *only* on the luminance of the source and the cone angle of the beam proceeding from the lens to the image. This is a basic and most important relation underlying all calculations of image illuminance.

It is often more convenient to convert the angle U' into other better known quantities, such as the *f*-number of the lens and the image magnification. The relation here is

$$\sin U' = \frac{1}{2(f\text{-number})(1 + m/m_p)}. \qquad (9)$$

Definition of *f*-number

The *f*-number of the lens is defined as the ratio of the focal length to the diameter of the entrance pupil; m is the image magnification; and m_p is the pupil magnification—*i.e.*, the diameter of the exit pupil divided by the diameter of the entrance pupil. Combining equations (8) and (9) gives

$$\text{Image illuminance} = E = \frac{\pi LB}{4(f\text{-number})^2(1 + m/m_p)^2}.$$

As an example in the use of this relation, if it is supposed that an *f*/2 lens is being used to project an image of a cathode-ray tube at five times magnification, the tube luminance being 5,000 foot-lamberts (1.7 candles per square centimetre), the lens transmittance is 0.8, and the pupil magnification is unity. Then the image illuminance will be

$$E = \frac{0.8 \times 5{,}000}{4 \times 4 \times 36} = 6.9 \text{ foot-candles.}$$

The image is very much less bright than the object, a fact that becomes clear to anyone attempting to provide a bright projected image in a large auditorium.

Distribution of illumination over an image. So far only the illumination at the centre of an image has been considered, but the distribution of illumination over a wide field is often important. In the absence of any lens, the small plane source already considered radiates in a direction inclined at an angle φ to the axis with an intensity $AB \cos \varphi$. This light has to travel farther than the axial light to reach a screen, and then it strikes the screen at another angle φ. The net result is that the oblique illumination on the screen is smaller than the axial illumination by the factor $\cos^4 \varphi$.

The same law can be applied to determine the oblique illumination due to a lens, assuming a uniform extended diffusing source of light on the other side of the lens. In this case, however, the exit pupil will not in general be a perfect circle because of possible distortion of the iris by that part of the optical system lying between the iris and the image. Also, any mechanical vignetting in the lens will make the aperture noncircular and reduce still further the oblique illumination. In a camera this reduction in oblique illumination results in darkened corners of the picture, but, if the reduction in brightness is gradual, it is not likely to be detected because the eye adapts quickly to changing brightness as the eyes scan over the picture area. Indeed, a 50 percent drop in brightness between the centre and corners of an ordinary picture is scarcely detectable.

Visual brightness. The apparent brightness of things seen by the eye follows the same laws as any other imaging system, because the apparent brightness is measured by the illuminance in the image that is projected on the retina. The angle U' in equation (8) inside the eye is determined by the size of the pupil of the eye, which varies from about one millimetre to about eight millimetres, depending on the brightness of the environment. Apart from this variation, retinal illuminance is directly proportional to object luminance, and objects having the same luminance appear equally bright, no matter at what distance they are observed.

Retinal illuminance

From this argument, it is clear that no visual instrument, such as a telescope, can possibly make anything appear brighter than when viewed directly. To be sure, a telescope having a large objective lens accepts more light from an object in proportion to the area of the lens aperture, but it magnifies the image area in the same proportion; so the increased light is spread over an increased area of the retina, and the illuminance remains unchanged. Actually, the telescopic view is always dimmer than the direct view because of light losses in the telescope due to glass absorption and surface reflections and because the exit pupil of the telescope may be smaller than the pupil of the eye, thus reducing the angle U'.

The case of a star being observed through a telescope is quite different, because no degree of magnification can possibly make a star appear as anything other than a point of light. Hence, star images appear brighter in proportion to the area of the telescope objective (assuming that the exit pupil is larger than the eye pupil), and the visibility of a star against the sky background is thus improved in proportion to the square of the diameter of the telescope objective lens.

Optics and information theory

GENERAL OBSERVATIONS

A new era in optics commenced in the early 1950s following the impact of certain branches of electrical engineering—most notably communication and information theory. This impetus was sustained by the development of the laser in the 1960s.

The initial tie between optics and communication theory came because of the numerous analogies that exist between the two subjects and because of the similar mathematical techniques employed to formally describe the behaviour of electrical circuits and optical systems. A topic of considerable concern since the invention of the lens as an optical imaging device has always been the description of the optical system that forms the image; information about the object is relayed and presented as an image. Clearly, the optical system can be considered a communication channel and can be analyzed as such. There is a linear relationship (*i.e.*, direct proportionality) between the intensity distribution in the image plane and that existing in the object, when the object is illuminated with incoherent light (*e.g.*, sunlight or light from a large thermal source). Hence, the linear theory developed for the description of electronic systems can be applied to optical image-forming systems. For example, an electronic circuit can be characterized by its impulse response—that is, its output for a brief impulse input of current or voltage. Analogously, an optical system can be characterized by an impulse response that for an incoherent imaging system is the intensity distribution in the image of a point source of light; the optical impulse is a spatial rather than a temporal impulse—otherwise the concept is the same. Once the appropriate impulse response function is known, the output of that system for any object intensity distribution can be determined by a linear superposition of impulse responses suitably weighted by the value of the intensity at each point in the object. For a continuous object intensity distribution this sum becomes an integral. While this example has been given in terms of an optical imaging system, which is certainly the most common use of optical elements, the concept can be used independent of whether the receiving plane is an image plane or not. Hence, for example, an impulse response can be defined for an optical system that is deliberately defocussed or for systems used for the display of Fresnel or Fraunhofer diffraction patterns. (Fraunhofer diffraction occurs when the light source and diffraction patterns are effectively at infinite distances from the diffracting system, and Fresnel diffraction occurs when one or both of the distances are finite.)

The optical system as a communication channel

Temporal frequency response. A fundamentally related but different method of describing the performance of an electronic circuit is by means of its temporal frequency response. A plot is made of the response for a series of input signals of a variety of frequencies. The response is measured as the ratio of the amplitude of the signal obtained out of the system to that put in. If there is no loss in the system, then the frequency response is unity (one) for that frequency; if a particular frequency fails to pass through the system, then the response is zero. Again, analogously the optical system may also be described by defining a spatial frequency response. The object, then,

to be imaged by the optical system consists of a spatial distribution of intensity of a single spatial frequency—an object the intensity of which varies as $(1 + a \cos \omega x)$, in which x is the spatial coordinate, a is a constant called the contrast, and ω is a variable that determines the physical spacing of the peaks in the intensity distribution. The image is recorded for a fixed value of a and ω and the contrast in the image measured. The ratio of this contrast to a is the response for this particular spatial frequency defined by ω. Now if ω is varied and the measurement is repeated, a frequency response is then obtained.

Nonlinear optical systems. The analogies described above go even further. Many optical systems are nonlinear, just as many electronic systems are nonlinear. Photographic film is a nonlinear optical element in that equal increments of light energy reaching the film do not always produce equal increments of density on the film.

A different type of nonlinearity occurs in image formation. When an object such as two stars is imaged, the resultant intensity distribution in the image is determined by first finding the intensity distribution formed by each star. These distributions must then be added together in regions where they overlap to give the final intensity distribution that is the image. This example is typical of an incoherent imaging system—*i.e.,* the light emanating from the two stars is completely uncorrelated. This occurs because there is no fixed phase relationship between the light emanating from the two stars over any finite time interval.

A similar nonlinearity arises in objects illuminated by light from the Sun or other thermal light source. Illumination of this kind, when there is no fixed relationship between the phase of the light at any pair of points in the incident beam, is said to be incoherent illumination. If the illumination of the object is coherent, however, then there is a fixed relationship between the phase of the light at all pairs of points in the incident beam. To determine the resultant image intensity under this condition for a two point object requires that the amplitude and phase of the light in the image of each point be determined. The resultant amplitude and phase is then found by summation in regions of overlap. The square of this resultant amplitude is the intensity distribution in the image. Such a system is nonlinear. The mathematics of nonlinear systems was developed as a branch of communication theory, but many of the results can be used to describe nonlinear optical systems.

Nonlinear optical systems

This new description of optical systems was extremely important to, but would not alone account for, the resurgence of optical research and development. This new approach resulted in the development of whole new branches of study, including optical processing and holography (see below). It also had an effect, together with the development of digital computers, on the concepts and versatility of lens design and testing. Finally, the invention of the laser, a device that produces coherent radiation, and the development and implementation of the theory of partially coherent light gave the added impetus necessary to change traditional optics into a radically new and exciting subject.

IMAGE FORMATION

Impulse response. An optical system that employs incoherent illumination of the object can usually be regarded as a linear system in intensity. A system is linear if the addition of inputs produces an addition of corresponding outputs. For ease of analysis, systems are often considered stationary (or invariant). This property implies that if the location of the input is changed, then the only effect is to change the location of the output but not its actual distribution. With these concepts it is then only necessary to find an expression for the image of a point input to develop a theory of image formation. The intensity distribution in the image of a point object can be determined by solving the equation relating to the diffraction of light as it propagates from the point object to the lens, through the lens, and then finally to the image plane. The result of this process is that the image intensity is the intensity in the Fraunhofer diffraction pattern of the lens aperture function (that is, the square of the Fourier transform of the lens aperture function; a Fourier transform is an integral equation involving periodic components). This intensity distribution is the intensity impulse response (sometimes called point spread function) of the optical system and fully characterizes that optical system. Figure 9 shows the impulse response for a perfect optical system (no aberration) that is limited only by its aperture.

Intensity distribution

With the knowledge of the impulse response, the image of a known object intensity distribution can be calculated. If the object consists of two points, then in the image plane the intensity impulse response function must be located at the image points and then a sum of these intensity distributions made. The sum is the final image intensity. If the two points are closer together than the half width of the impulse response, they will not be resolved. For an object consisting of an array of isolated points, a similar procedure is followed—each impulse response is, of course, multiplied by a constant equal to the value of the intensity of the appropriate point object. Normally, an object will consist of a continuous distribution of intensity, and, instead of a simple sum, a convolution integral results.

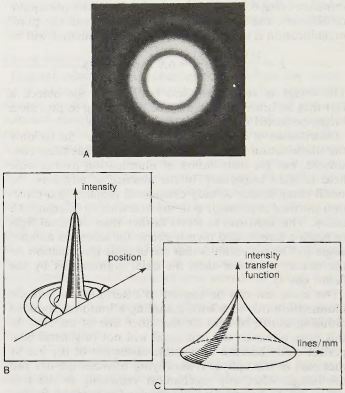


Figure 9: *The intensity impulse response of a perfect (i.e., no aberrations) optical system.*
(A) Photograph; (B) plot; (C) the transfer function of the same aberration-free system.

Transfer function. The concept of the transfer function of an optical system can be approached in several ways. Formally and fundamentally it is the Fourier transform of the intensity impulse response. The transfer function for the optical system for Figure 9B is shown in Figure 9C. Because the impulse response is related to the lens aperture function, so is the transfer function. In particular, the transfer function can be obtained from a knowledge of the aperture function by taking the function and plotting the resultant overlapping areas as the aperture function is slid over itself (*i.e.,* the autocorrelation of the aperture function).

Auto-correlation

Conceptually, however, the transfer function is best understood by considering the object intensity distribution to be a linear sum of cosine functions of the form $(1 + a \cos 2\pi\mu x)$, in which a is the amplitude of each component of spatial frequency μ. The image of a cosine intensity distribution is a cosine of the same frequency; only the contrast and phase of the cosine can be affected by a linear system. The image of the above object intensity distribution can be represented by $[1 + b \cos (2\pi\mu x + \varphi)]$, in which b is the amplitude of the output cosine of frequency μ and φ is the phase shift. The transfer function, $\tau(\mu)$, for that frequency is then given by the ratio of the amplitudes:

$$\tau(u) = \frac{b}{a} e^{j\varphi(u)} .$$

If μ is now varied, the spatial frequency response of the system is measured by determining $\tau(\mu)$ for the various values of μ. It should be noted that $\tau(\mu)$ is in general complex (containing a term with $\sqrt{-1}$), although in the example of Figure 9 it is real and positive [$\varphi(\mu) = 0$ for all μ].

The transfer function, like the impulse response, fully characterizes the optical system. To make use of the transfer function to determine the image of a given object requires that the object be decomposed into a series of periodic components called its spatial frequency spectrum. Each term in this series must then be multiplied by the appropriate value of the transfer function to determine the individual components of the series that is the spatial frequency spectrum of the image—a transformation of this series will give the image intensity. Thus, any components in the object spectrum that have a frequency for which $\tau(\mu)$ is zero will be eliminated from the image.

PARTIALLY COHERENT LIGHT

Image formation is concerned above with incoherent object illumination, which results in an image formed by the addition of intensities. The study of diffraction and interference, on the other hand, requires coherent illumination of the diffracting object, the resulting diffracted optical field being determined by an addition of complex amplitudes of the wave disturbances. Thus, two different mechanisms exist for the addition of light beams, depending upon whether the beams are coherent or incoherent with respect to each other. Unfortunately, this is not the whole story; it is not sufficient to consider only the two situations of strictly coherent and strictly incoherent light. In fact, strictly incoherent fields are only approximately obtainable in practice. Furthermore, the possibility of intermediate states of coherence cannot be ignored; it is necessary to describe the result of mixing incoherent light with coherent light. It was to answer the question How coherent is a beam of light? (or the equivalent one, How incoherent is a beam of light?) that the theory of partial coherence was developed. Marcel Verdet, a French physicist, realized in the 19th century that even sunlight is not completely incoherent, and two objects separated by distances of over approximately $^1/_{20}$ millimetre will produce interference effects. The eye, operating unaided in sunlight, does not resolve this separation distance and hence can be considered to be receiving an incoherent field. Two physicists, Armand Fizeau in France and Albert Michelson in the United States, were also aware that the optical field produced by a star is not completely incoherent, and hence they were able to design interferometers to measure the diameter of stars from a measurement of the partial coherence of the starlight. These early workers did not think in terms of partially coherent light, however, but derived their results by an integration over the source. At the other extreme, the output from a laser can produce a highly coherent field.

Partial coherence

The concepts of partially coherent light can best be understood by means of some simple experiments. A circular uniform distant source produces illumination on the front of an opaque screen containing two small circular apertures, the separation of which can be varied. A lens is located behind this screen, and the resultant intensity distribution in its focal plane is obtained. With either aperture open alone, the intensity distribution observed is such that it is readily associated with the diffraction pattern of the aperture, and it may thus be concluded that the field is coherent over the dimensions of the aperture. When the two apertures are opened together and are at their closest separation, two-beam interference fringes are observed that are formed by the division of the incident wave front by the two apertures. As the separation of the apertures increases, the observed interference fringes get weaker and finally disappear, only to reappear faintly as the separation is further increased. As the separation of the apertures is increased, these results show that (1) the fringe spacing decreases; (2) the intensities of the fringe minima are never zero; (3) the relative intensity of the maxima above the minima steadily decreases; (4) the absolute value of the intensity of the maxima decreases and that of the minima increases; (5) eventually, the fringes disappear, at which point the resultant intensity is just twice the intensity observed with one aperture alone (essentially an incoherent addition); (6) the fringes reappear with a further increase in separation of the aperture, but the fringes contain a central minimum, not a central maximum.

If the intensities of the two apertures are equal, then the results (1) through (5) can be summarized by defining a quantity in terms of the maximum intensity (I_{max}) and the minimum intensity (I_{min}), called the visibility (V) of the fringes—*i.e.*, $V = (I_{max} - I_{min})/(I_{max} + I_{min})$. The maximum value of the visibility is unity, for which the light passing through one aperture is coherent with respect to the light passing through the other aperture; when the visibility is zero, the light passing through one aperture is incoherent with respect to the light passing through the other aperture. For intermediate values of V the light is said to be partially coherent. The visibility is not a completely satisfactory description because it is, by definition, a positive quantity and cannot, therefore, include a description of item (6) above. Furthermore, it can be shown by a related experiment that the visibility of the fringes can be varied by adding an extra optical path between the two interfering beams.

Fringe visibility

The key function in the theory of partially coherent light is the mutual coherence function $\Gamma_{12}(\tau) = \Gamma(x_1, x_2, \tau)$, a complex quantity, which is the time averaged value of the cross correlation function of the light at the two aperture points x_1 and x_2 with a time delay τ (relating to a path difference to the point of observation of the interference fringes). The function can be normalized (*i.e.*, its absolute value set equal to unity at $\tau = 0$ and $x_1 = x_2$) by dividing by the square root of the product of the intensities at the points x_1 and x_2 to give the complex degree of coherence, hence

$$\gamma_{12}(\tau) = \frac{\Gamma_{12}(\tau)}{\sqrt{I(x_1) I(x_2)}} .$$

The modulus of $\gamma_{12}(\tau)$ has a maximum value of unity and a minimum value of zero. The visibility defined earlier is identical to the modulus of the complex degree of coherence if $I(x_1) = I(x_2)$.

Often the optical field can be considered to be quasi-monochromatic (approximately monochromatic), and then the time delay can be set equal to zero in the above expression, thus defining the mutual intensity function. It is often convenient to describe an optical field in terms of its spatial and temporal coherence by artificially separating out the space- and time-dependent parts of the coherence function. Temporal coherence effects arise from the finite spectral width of the source radiation; a coherence time Δt can be defined as $1/\Delta\nu$, in which $\Delta\nu$ is the frequency bandwidth. A related coherence length Δl can also be defined as $c/\Delta\nu = \lambda^2/\Delta\lambda^2$, in which c is the velocity of light, λ is the wavelength, and $\Delta\lambda$ the wavelength bandwidth. Providing that the path differences in the beams to be added are less than this characteristic length, the beams will interfere.

The term spatial coherence is used to describe partial coherence arising from the finite size of an incoherent source. Hence, for the equipath position for the addition of two beams, a coherence interval is defined as the separation of two points such that the absolute value $|\gamma_{12}(0)|$ is some prechosen value, usually zero.

The mutual coherence function is an observable quantity that can be related to the intensity of the field. The partially coherent field can be propagated by use of the mutual coherence function in a similar way to the solution of diffraction problems by propagation of the complex amplitude. The effects of partially coherent fields are clearly of importance in the description of normally coherent phenomena, such as diffraction and interference, but also in the analysis of normally incoherent phenomena, such as image formation. It is notable that image formation in coherent light is not linear in intensity but is linear in the complex amplitude of the field, and in partially coherent light the process is linear in the mutual coherence.

The mutual coherence function

OPTICAL PROCESSING

Coherent optical systems. Optical processing, information processing, signal processing, and pattern recognition are all names that relate to the process of spatial frequency filtering in a coherent imaging system—specifically, a method in which the Fraunhofer diffraction pattern (equivalently the spatial frequency spectrum or the Fourier transform) of a given input is produced optically and then operated upon to change the information content of the optical image of that input in a predetermined way.

The idea of using coherent optical systems to allow for the manipulation of the information content of the image is not entirely new. The basic ideas are essentially included in Abbe's theory of vision in a microscope first published in 1873; the subsequent illustrative experiments of this theory, notably by Albert B. Porter in 1906, are certainly simple examples of optical processing.

Abbe's ideas can be interpreted as a realization that image formation in a microscope is more correctly described as a coherent image-forming process than as the more familiar incoherent process. Thus, the coherent light illuminating the object on the microscope stage would be diffracted by that object. To form an image, this diffracted light must be collected by the objective lens of the microscope, and the nature of the image and the resolution would be affected by how much of the diffracted light is collected. As an example, an object may be considered consisting of a periodic variation in amplitude transmittance—the light diffracted by this object will exist in a series of discrete directions (or orders of diffraction). This series of orders contains a zero order propagating along the optical axis and a symmetric set of orders on both sides of this zero order. Abbe correctly discerned what would happen as the microscope objective accepted different combinations of these orders. For example, if the zero order and one first order are collected, then the information obtained will be that the object consisted of a periodic distribution, but the spatial location of the periodic structure is not correctly ascertained. If the other first order of diffracted light is included, the correct spatial location of the periodic structure is also obtained. As more orders are included, the image more closely resembles the object.

Coherent optical data processing became a serious subject for study in the 1950s, partly because of the work of a French physicist, Pierre-Michel Duffieux, on the Fourier integral and its application to optics, and the subsequent use of communication theory in optical research. The work was initiated in France by André Maréchal and Paul Croce, and today a variety of problems can be attempted by the technique. These include removal of raster lines (as in a TV picture) and halftone dots (as in newspaper illustration); contrast enhancement; edge sharpening; enhancement of a periodic or isolated signal in the presence of additive noise; aberration balancing in which a recorded aberrated image can be somewhat improved; spectrum analysis; cross correlation of data; matched and inverse filtering in which a bright spot of light in the image indicates the presence of a particular object.

Examples of data processing

Filtering. The basic system required for coherent optical processing consists of two lenses (Figure 10). A collimated beam of coherent light is used to transilluminate the object. The first lens produces the characteristic Fraunhofer diffraction pattern of the object, which is the spatial frequency distribution associated with the object. (Mathematically, it is the Fourier transform of the object amplitude distribution.) A filter that consists of amplitude (density) or phase (optical path) variations, or both, is placed in the plane of the diffraction pattern. The light passing through this filter is used to form an image, this step being accomplished by the second lens. The filter has the effect of changing the nature of the image by altering the spatial frequency spectrum in a controlled way so as to enhance certain aspects of the object information. Maréchal gave the descriptive title double diffraction to this type of two-lens system.

Blocking filters

The filters can be conveniently grouped into a variety of types depending upon their action. Blocking filters have regions of complete transparency and other regions of complete opacity. The opaque areas completely remove certain portions of the spatial frequency spectrum of the object. The removal of raster lines and halftone dots is accomplished with this type of filter. The object can be considered as a periodic function the envelope of which is the scene or picture—or equivalently the periodic function samples the picture. The diffraction pattern consists of a periodic distribution with a periodicity reciprocally related to the raster periodicity. Centred at each of these periodic locations is the diffraction pattern of the scene. Hence, if the filter is an aperture centred at one of these locations so that only one of the periodic elements is allowed to pass, then the raster periodicity is removed, but the scene information is retained (see Figure 10). The problem of

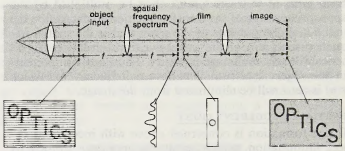


Figure 10: Two-lens coherent optical processing system, showing how the raster periodicity is removed but the scene information is retained (see text).

the removal of halftone dots is the two-dimensional equivalent of the above process. Because the two-dimensional spatial frequency spectrum of an object is displayed in a coherent optical processing system, it is possible to separate out information by means of its orientation. Other applications of blocking filters include band-pass filters, which again have a direct relationship to the band-pass filters in electronic circuits.

Amplitude filters

A second type of filter is an amplitude filter that will consist of a continuous density variation. These filters can be produced to achieve the enhancement of contrast of the object input or the differentiation of the object. They are often constructed by controlled exposure of photographic film or evaporation of metal onto a transparent substrate.

Certain optical processing techniques require that the phase of the optical field be changed, and, hence, a filter with no absorption but varying optical thickness is required. Usually, both the amplitude and the phase have to be modified, however, thus requiring a complex filter. In simple cases the amplitude and phase portions can be made separately, the phase filter being manufactured by using an evaporated layer of transparent material, such as magnesium fluoride. Current practice is to fabricate the complex filter by an interferometric method in which the required complex amplitude function is recorded as a hologram (see below).

The phase-contrast microscope can be considered to be an example of an optical processing system, and the concepts understood by reference to Figure 10. Only the simplest form will be considered here. The spatial frequency spectrum of the phase object is formed and the phase of the central portion of that spectrum changed by $\pi/2$ or $3\pi/2$ to produce positive or negative phase contrast, respectively. To improve the contrast of the image an additional filter covering the same area as the phase filter is used that is partially absorbing (*i.e.*, an amplitude filter). The restriction on this process is that the variations of the phase $\varphi(x)$ are small so that $e^{i\varphi(x)} \cong 1 + i\varphi(x)$. With incoherent light, phase information is not visible, but many biological samples consist only of variations of refractive index, which results in optical path and hence phase, differences. The image in the phase-contrast microscope is such that the intensity in that image relates linearly to, and hence is a display of, the phase information in the object—*e.g.*, $I(x) \propto 1 \pm 2\varphi(x)$ for positive and negative phase contrast, respectively.

Correction of aberrated images

One of the important motivations for the study of optical processing methods is to achieve some correction of aberrated images. Considerable technological advantage can be gained if photographs taken with an aberrated optical system in incoherent light can be corrected by subsequent processing. Within definable limits this can be

accomplished, but the impulse response or the transfer function of the aberrated system must be known. The recorded image intensity distribution is the convolution of the object intensity with the intensity impulse response of the aberrated system. This record is the input to the coherent optical processing system; the diffraction pattern formed in this system is the product of the spatial frequency spectrum of the object and the transfer function of the aberrated system. Conceptually, the filter has to be the inverse of the transfer function in order to balance out its effect. The final image would then ideally be an image of the object intensity distribution. It is critical, however, that the transfer function has a finite value over only a limited frequency range, and only those frequencies that are recorded by the original aberrated system can be present in the processed image. Hence, for these spatial frequencies that were recorded, some processing can be carried out to get a flatter effective transfer function; both the contrast and the phase of the spatial frequency spectrum may have to be changed because the transfer function is, in general, a complex function. Prime examples are for images aberrated by astigmatism, defocussing, or image motion.

HOLOGRAPHY

Theory. Holography is a two-step coherent image-forming process in which an intermediate record is made of the complex optical field associated with the object. The invention of the wave-front reconstruction process (now called holography) was first described in 1948 by Dennis Gabor, a Hungarian-born physicist, with a specific application in mind—to attempt to improve the resolution of images formed with electron beams. The technique has, however, had most of its success to date when light beams are employed particularly in the visible part of the spectrum. The first step in the process is to record (often on high-resolution film) the interference pattern produced by the interaction of the light diffracted by the object of interest and a coherent background or reference wave. In the second step, this record, which is the hologram, is illuminated coherently to form an image of the original object. In fact, two images are usually formed—a real image (often called the conjugate image) and a virtual image (often called the primary image). There are two basic concepts that underlie this process: first, the addition of a coherent background (or reference) beam. Two optical fields may be considered, the complex amplitudes of which vary as the cosine of an angle proportional to the space coordinate and as the modulus (absolute magnitude) of the cosine of the angle, respectively. From a measurement of the intensity of these fields it is impossible to distinguish them because both vary as the cosine squared of the space coordinate. If a second coherent optical field is added to each of these two fields, however, then the resultant fields become $(1 + \cos x)$ and $(1 + |\cos x|)$, respectively. The measured intensities are now different, and the actual fields can be determined by taking the square root of the intensity. The amplitude transmittance of a photographic record is, in fact, the square root of the original intensity distribution that exposed the film. In a more general sense, an optical field of the form $a(x) \exp [i\varphi_1(x)]$, in which $a(x)$ is the amplitude and $\varphi_1(x)$ is the phase, can be distinguished from a field $a(x) \exp [i\varphi_2(x)]$ by adding a coherent background; the phases $\varphi_1(x)$ and $\varphi_2(x)$ are then contained as cosine variations of intensity in the resulting pattern. Hence, the problem of recording the phase information of the optical field is circumvented. When the hologram is illuminated, however, the optical field that originally existed in that plane is recreated. To apply the second basic concept—that of an image-forming property—it is necessary to determine what the hologram of a point object is—in actuality it is a sine-wave zone plate or zone lens. If a collimated beam of light is used to illuminate a zone lens, then two beams are produced; the first comes to a real focus, and the other is a divergent beam that appears to have come from a virtual focus. (By comparison, the more classical zone plate has a multitude of real and virtual focuses, and a real lens has but one.) When the object is other than a point, the zone lens is modified by the diffraction pattern of the object; *i.e.*, each point on the object produces its own zone lens, and the resultant hologram is a summation of such zone lenses.

Steps in the holography process

In Gabor's original system the hologram was a record of the interference between the light diffracted by the object and a collinear background. This automatically restricts the process to that class of objects that have considerable areas that are transparent (see Figure 11A). When the hologram is used to form an image, twin images are formed, as illustrated in Figure 11B. The light associated with these images is propagating in the same direction, and hence in the plane of one image light from the other image appears as an out-of-focus component. This type of hologram is usually referred to as an in-line Fresnel hologram because it is the pattern of the object that interferes with the collinear coherent background. The deleterious effects of the second image can be minimized if the hologram is made in the far field of the object so that it is a Fraunhofer diffraction pattern of the object that is involved. This latter technique has found significant application in microscopy, particularly in the measurement of small particles, and in electron microscopy.

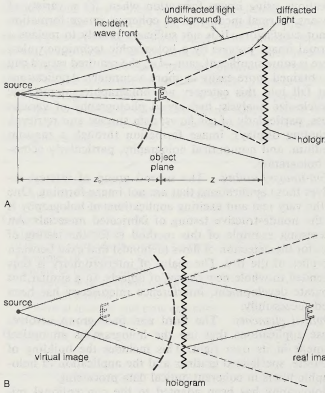


Figure 11: *Holography.*
(A) Formation of a hologram by light diffracted by an object with collinear background (Gabor's original method). (B) Image formation by illumination of the hologram.

A more versatile method of recording the hologram is to add a second beam of light as a reference wave to produce the hologram. The hologram is now the record of the interference pattern produced by the light diffracted by the object and this separate reference wave. The reference wave is usually introduced at an angle to the diffracted beam, hence this method is often called off-axis (or side-band) holography. When the hologram is illuminated, the image-forming beams do not propagate in the same direction but are inclined to each other with an angle twice that between the diffracted beam and the original reference beam. Hence, the light associated with an image is completely separated from the other image.

A further technique that has some value and relates to the earlier discussion of optical processing is the production of the so-called generalized or Fourier transform hologram. Here the reference beam is added coherently to a Fraunhofer diffraction pattern of the object or formed by a lens (as in the first stage of Figure 10).

The generalized hologram

The process described so far has been in terms of transmitted light through the object. The methods involving the separate reference beam can be used in reflected light, and the virtual (primary) image produced from the hologram has all the properties of an ordinary image in terms of three-dimensionality and parallax. Normally, a recorded

image is only a two-dimensional representation of the object. Full-colour holograms can be recorded by essentially recording three holograms simultaneously—one in red light, one in blue, and one in green.

Applications. *Image-forming.* The applications mentioned here are in three groups: image-forming applications, non-image-forming applications, and the hologram as an optical element. It is notable that all three groups relate to the basic use of the process rather than specific holographic techniques. The first group involves those applications using image formation when, for a variety of reasons, normal incoherent or coherent image formation is not satisfactory. It is not sufficient merely to replace a normal image process by a holographic technique unless there is some significant gain—*i.e.,* the required record can be obtained more easily or more accurately. Applications that fall into this category are holographic microscopy; particle-size analysis; high-speed photography of various types, particularly of gas flows; data storage and retrieval, including displays; image formation through a random medium; and non-optical holography, particularly acoustic holography.

Non-image-forming. The second group of interest involves those applications that are not image-forming. One of the very real and exciting applications of holography is to the nondestructive testing of fabricated materials. An interesting example of this method is for the testing of tires for the detection of flaws (debonds) that exist between the plies of the tire. The realm of interferometry is thus extended to whole new classes of objects. In a similar but separate development, interference microscopy has been used successfully.

Optical elements. The third and final group involves those applications that use the hologram as an optical element in its own right. This includes the building of accurate, specialized gratings and the application of holographic filters in coherent optical data processing.

Holography has been adapted to the conventional microscope, which is modified by the inclusion of a separate reference beam so that the light diffracted by the object in the microscope is made to interfere with the light from the reference beam. An increase in the depth of field available is achieved by this type of recording process. The image is produced when the hologram is illuminated again by a coherent beam.

The application of holography to particle-size analysis (*e.g.,* to determine the size distribution of dust and liquid droplets) was really the first of the modern-day applications. In a sense, this, too, can be thought of as microscopy. The principles of Fraunhofer holography were developed to solve this particular problem. Because the particles are in motion, a hologram must be made instantaneously. A pulsed-ruby laser technique is therefore used. The hologram is formed between the light diffracted by the particles or droplets and the coherent background light that passes directly through the sample. In reconstruction, a series of stationary images are formed that can be examined at leisure. Hence, a transient event has been transformed into a stationary image for evaluation.

Transient holograms

Data storage and retrieval is perhaps one of the more important applications of holography, which is in the process of development and refinement. Because the information about the image is not localized, it cannot be affected by scratches or dust particles. Recent advances in materials, particularly those that might be erasable and reusable, have added further interest in holographic optical memories.

Among the non-image-forming applications are interferometry, interference microscopy, and optical processing. Holographic interferometry can be done in several ways. The basic technique involves recording a hologram of the object of interest and then interfering the image produced from this hologram with the coherently illuminated object itself. A variation on this technique would be to form two holograms at different times of the same object as it undergoes testing. The two holograms can then be used together to form two images, which would again interfere. The interference fringes seen would be related to the changes in the object between the two exposures. A third technique uses a time-average hologram, which is particularly applicable to the study of vibrating objects.

There are two applications that come under the heading holographic optical elements—the use of holographic gratings and the use of holographic filters for coherent optical data processing.

NONLINEAR OPTICS

Nonlinear effects in optics are now quite readily observable using the highly coherent and highly energetic laser beams. These effects occur when the output of a system is not linearly related to the input (*e.g.,* a nonlinear electronic amplifier can be built with a gain that increases with signal intensity). The most important nonlinear effect is probably frequency doubling. Optical radiation of a given frequency is propagated through a crystalline material and interacts with that material to produce an output of a different frequency that is twice the input frequency. For example, the 10,600-angstrom infrared output of a neodymium laser can, under suitable conditions, be converted into green light at 5300 angstroms in a crystal of barium strontium niobate. (B.J.T.)

BIBLIOGRAPHY. There are many journals and hundreds of books covering the general field of optics; some of the more familiar books include ARTHUR C. HARDY and FRED H. PERRIN, *The Principles of Optics* (1932); and FRANCIS A. JENKINS and HARVEY E. WHITE, *Fundamentals of Optics,* 4th ed. (1976). At a more advanced level are R.S. LONGHURST, *Geometrical and Physical Optics,* 3rd ed. (1973); L.C. MARTIN, *Technical Optics,* 2 vol. (1960–61); MAX BORN and EMIL WOLF, *Principles of Optics,* 6th ed. (1980, reissued 1993); and EUGENE HECHT and ALFRED ZAJAC, *Optics,* 2nd ed. (1987). In the purely geometric field, the following can be recommended: W.T. WELFORD, *Geometrical Optics* (1962), which is fairly elementary and provides an excellent introduction to the subject; and L.C. MARTIN, *Geometrical Optics* (1955), which is somewhat more advanced. Other useful works include *Applied Optics and Optical Engineering* (irregular), with volumes treating optical devices, optical theory, and the principal types of optical instruments; and A.E. CONRADY, *Applied Optics and Optical Design,* part 1 (1929, reissued 1992), and part 2, ed. and completed by RUDOLF KINGSLAKE (1960, reissued 1992), a full description of the theory and practice of lens design, providing detailed instructions for the design of many types of optical systems. WALTER T. WELFORD, *Useful Optics* (1991), is a succinct review of principles basic to implementing optical tools.

Recommended books on the subject of information theory are EDWARD L. O'NEILL, *Introduction to Statistical Optics* (1963, reissued 1992); JOSEPH W. GOODMAN, *Introduction to Fourier Optics* (1968, reissued 1988); ARNOLD ROY SHULMAN, *Optical Data Processing* (1970); KENDALL PRESTON, *Coherent Optical Computers* (1972); H. LIPSON (ed.), *Optical Transforms* (1972); AJOY GHATAK and K. THYAGARAJAN, *Optical Electronics* (1989); GEORGE W. STROKE, *An Introduction to Coherent Optics and Holography,* 2nd ed. (1969); JOHN B. DEVELIS and GEORGE O. REYNOLDS, *Theory and Applications of Holography* (1967); ROBERT J. COLLIER, CHRISTOPH B. BURCKHARDT, and LAWRENCE H. LIN, *Optical Holography* (1971); and ROBERT JONES and CATHERINE WYKES, *Holographic and Speckle Interferometry,* 2nd ed. (1989). Topics of current research may be found in the periodicals *Journal of the Optical Society of America* (monthly); *Applied Optics* (semimonthly); and *Progress in Optics* (annual).

(R.K./B.J.T./Ed.)

The Mathematical Theory of Optimization

Optimization is a technique for improving or increasing the value of some numerical quantity that in practice may take the form of temperature, air flow, speed, pay-off in a game, political appeal, destructive power, information, monetary profit, and the like. Techniques of optimization assume such varied forms that no one general description is possible. With the advent of modern technology, more and more emphasis has been placed on optimization of various types, and special thinking has developed to the extent that it is meaningful to speak of a mathematical theory of optimization. Computer technology has been critically important in practical applications. Further advances in the optimization and control of complex systems will probably depend more on mathematical theory than on technological invention.

Optimization includes linear and nonlinear programming, cybernetics, and control theory, the various aspects of which are treated in this article. Game theory is often included as well, since early work on optimization was extended by the development of that branch of mathematics.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 10/21, 10/22, and 10/23, and the *Index.* This article is divided into the following sections:

Linear and nonlinear programming (mathematical programming) 217
  General observations 217
  Origins and influences 217
  Linear programming theory 218
    Basic ideas
    The simplex method
    The dual problem
  Nonlinear programming theory 220
    Classification of the problems
    Methods of solution
Cybernetics 221
  Definitions of the term cybernetics 221
  Principles 221
    Information theory
    Automata as information converters
    Cybernetic systems
Control theory 223
  General background 223
  Examples of modern control systems 223
  Principles of control 224
  Control of linear systems 225
    Systems with constant coefficients
    Optimal control
    Optimal filtering and state estimation
  Nonlinear control systems 226
Bibliography 226

Linear and nonlinear programming (mathematical programming)

GENERAL OBSERVATIONS

Mathematical programming may be described in terms of its mathematical structure and computational procedures or in terms of the broad class of important decision problems which can be formulated as the minimization (maximization) of a function of several variables that are subject to a system of side constraints. For example, a linear program is defined as the minimization of a linear "objective" function whose variables satisfy a system of linear inequalities.

The scope of mathematical programming

In practice, mathematical programming usually refers to such linear programs, the general study of nonlinear programs (those in which either the objective function or at least one of the constraint functions is nonlinear), integer programs (that is, linear programs with the additional restriction that some or all of the variables must be integer valued), stochastic programs (those programs involving random variables), and network flow theory (transportation or flow through networks). As such, mathematical programming overlaps, has contributed to, and has been influenced by operations research, mathematical economics, control theory, dynamic programming, and combinatorial theory.

The term programming had its origin in programming (*i.e.,* planning, scheduling) the quantity and timing of the various activities of an organization such as a factory, an airline, the national economy, or world trade. (It is not to be confused with "programming" as used for the task of preparing a sequence of instructions for a computer.) The goal is to find an optimum schedule.

A simple example, "the assignment problem," illustrates the essential difficulty. A factory has 70 men with different qualifications, and it is desirable to assign them to 70 jobs. If a "value" can be attached to assigning a particular man to a particular job, then the problem becomes one of selecting out of 70! (which is the product of integers from 1 to 70) possible ways of permuting the assignments the one that yields the maximum total value to the factory. Because 70! is approximately 10^{100}, it would take a computer executing 1,000,000 operations per second more than 10^{87} years (or many times the projected life of the universe) to examine all the permutations. Such decision problems are common and have resulted in the development of clever formulation of the mathematical models, powerful mathematical methods of solution, and efficient computer algorithms. The solution of linear programs is said to constitute 25 percent of all the time used by computers to solve scientific problems.

ORIGINS AND INFLUENCES

Although widely used now to solve everyday decision problems, linear programming was comparatively unknown before 1947. No work of any significance was carried out before this date, even though the French mathematician Jean-Baptiste-Joseph Fourier seemed to be aware of the subject's potential as early as 1823. A Russian mathematician, Leonid Vitalyevich Kantorovich, who published an extensive monograph in 1939, *Matematicheskie metody organizatsi i planirovaniya proizvodstva* ("Mathematical Methods for Organization and Planning of Production"), is credited with being the first to recognize that certain important broad classes of scheduling problems had well-defined mathematical structures. Unfortunately his proposals remained unknown both in the erstwhile Soviet Union and elsewhere for nearly two decades. Meanwhile, linear programming had developed considerably in the United States and Western Europe. In the period following World War II, officials in the United States government felt that efficient coordination of the energies of a whole nation in the advent of nuclear war would require the use of scientific planning techniques. The advent of the computer made such an approach feasible.

Intensive work began in 1947 in the U.S. Air Force. The linear programming model was proposed because it is simple and practical and yet provided a sufficiently general framework for representing interdependent activities that must share scarce resources. The system (*e.g.,* national economy) is viewed as made up of various activities (*e.g.,* production, training, shipping, disposal, and storing).

The first linear programming model

Each activity is assumed to require a flow of inputs and outputs of various types of items (*e.g.*, men, materials) proportional to the level of the activity. Activity levels are assumed to be representable in terms of positive numbers or zero. The revolutionary feature of the approach, however, was the expressing of the goal, as that of minimizing (maximizing) an objective function (*e.g.*, maximizing sorties in the case of the air force or maximizing profits in industry). Before 1947 all practical planning was characterized by a series of authoritatively imposed rules of procedure and priorities. General objectives were never stated, probably because of the impossibility of performing the calculations necessary to minimize an objective function under constraints. In 1947 a method composed of successive tests for optimality at extreme points and intervening linear movement along polygonal edges called the simplex computational method was introduced which turned out to be indeed efficient. Interest in linear programming grew rapidly and by 1951 its use spread to industry. Today, it is almost impossible to name an industry that is not using mathematical programming in some form, although its use varies greatly, even within the same industry.

The current interest in linear programming by economists appears to be an anachronism. The French economist François Quesnay's attempt in his *Tableau économique* (1758) to interrelate the role of the landlord, the peasant, and the artisan was a crude example of a linear programming model. Léon Walras, another French economist, in 1874 proposed a sophisticated approach that had as part of its structure fixed technological coefficients (as assumed in linear programs). Oddly enough, until the 1930s, the linear-type model was little exploited.

Von Neumann in a 1937 paper analyzed a steadily expanding economy based on alternative methods of production and fixed technological coefficients. As far as mathematical history is concerned, the study of linear inequality systems excited virtually no interest before 1936. In 1911 a vertex-to-vertex movement along edges of a polyhedron (as is done in the simplex method) was suggested as a way to solve a problem that involved optimization, and in 1941 movement along edges was proposed for a problem involving transportation. Credit for laying much of the mathematical foundations should probably go to von Neumann. In 1928 he published his famous paper on *Game theory*. In 1947 he conjectured the equivalence of linear programs and matrix games, introduced the important concept of duality, and made several proposals for the numerical solution of linear programming and game problems. Serious interest by other mathematicians began in 1948 with the rigorous development of duality and related matters.

The general simplex method, already mentioned (see above *Origins and influences*) and to be discussed in detail (see below *The simplex method*), was first programmed in 1951 for the United States Bureau of Standards SEAC computer. Starting in 1952, the simplex method was programmed for use on various models of IBM (International Business Machines) computers and later for those of other companies. These programs turned out to be practical. As a result, commercial applications of linear programs in industry and government grew rapidly. New computational techniques and variations of older techniques continued to be developed.

At the present time there is much interest in solving large linear programs with special structures, for example, corporate models and national planning models that are multistaged, dynamic, and exhibit a hierarchical structure. It is estimated that certain underdeveloped countries have the potential of increasing their GNP (gross national product) anywhere from 10 to 15 percent per year if detailed growth models of the economy could be constructed, optimized, and implemented.

LINEAR PROGRAMMING THEORY

Linear programming in factory production

Basic ideas. A simple problem in linear programming is one in which it is necessary to find the maximum (or minimum) value of a simple function, such as $x_1 + 2x_2$, subject to certain constraints. An example of its application might be that of a factory producing two commodities. In any production run, it produces x_1 of the first type and x_2 of the second. If the profit on the second type is twice that on the first, then $x_1 + 2x_2$ represents the total profit. The function $x_1 + 2x_2$ is known as the objective function.

Clearly the profit will be highest if the factory devotes its entire production capacity to making the second type of commodity. In a practical situation, however, this may not be possible; a set of constraints is introduced by such factors as availability of machine time, manpower, or raw materials. For example, if the second type of commodity requires a raw material that is limited so that no more than five can be made in any batch, then x_2 must be less than or equal to five; *i.e.*, $x_2 \leq 5$. If the first commodity requires another type of material limiting it to eight per batch then $x_1 \leq 8$. If 1 and 2 take equal times to make and the machine time available allows a maximum of ten to be made in a batch, then $x_1 + x_2$ must be less than or equal to 10; *i.e.*, $x_1 + x_2 \leq 10$.

Two other constraints are that x_1 and x_2 must each be greater than or equal to zero, because it is impossible to make a negative number of either; *i.e.*, $x_1 \geq 0$ and $x_2 \geq 0$. These conditions result in a simple linear program that can be stated in concise algebraic form. The problem is to find the values of x_1 and x_2 for which the profit is a maximum. Any solution can be denoted by a pair of numbers (x_1, x_2); for example, if $x_1 = 3$ and $x_2 = 6$, the solution is (3, 6). These numbers can be represented by points plotted on two axes; in Figure 1 the distance along

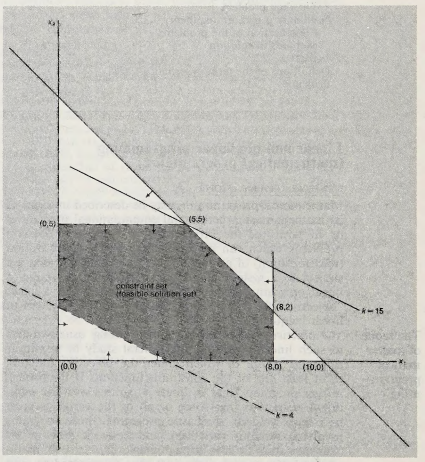


Figure 1: Constraint set bounded by the five lines $x_1 = 0$, $x_2 = 0$, $x_1 = 8$, $x_2 = 5$, and $x_1 + x_2 = 10$. These enclose an infinite number of points that represent feasible solutions.

the horizontal axis represents x_1 and that along the vertical represents x_2. An infinite number of points exist corresponding to an infinite number of solutions, but because of the constraints the feasible solutions must lie within a certain region. For example, the constraint $x_1 \geq 0$ means that points representing feasible solutions lie on or to the right of the x_2 axis. The constraint $x_2 \geq 0$ means that they also lie on or above the x_1 axis. Application of all the constraints leads to the result that points representing solutions lie within the shaded area bounded by a polygon formed by intersection of the lines $x_1 = 0$, $x_2 = 0$, $x_1 = 8$, $x_2 = 5$, and $x_1 + x_2 = 10$. These form the constraint set. For example, to make three items of commodity 1 and four of 2 is a feasible solution since the point (3, 4) lies in this region. To find the best solution, the objective func-

tion $x_1 + 2x_2 = k$ is plotted on the graph for some value of k, say $k = 4$. This value is indicated by the broken line in Figure 1. As k is increased, a family of parallel lines are produced and the line for $k = 15$ just touches the constraint set at the point (5, 5). If k is increased further, the values of x_1 and x_2 lie outside the set of feasible solutions. Thus the best solution is that in which equal quantities of each commodity are made. The points (0, 0), (0, 5), (5, 5), (8, 2), and (8, 0) are the extreme points of the constraint set, and the problem involves finding the extreme point the coordinates of which yield the largest value for k.

The standard form of a linear program is expressed as an instruction to maximize (or minimize) a linear expression in n nonnegative variables, $\{x_i\}$, subject to m linear equations (see 1). In applications, the equations correspond to material balance of various items and the variables to levels of various activities (*e.g.*, production processes). Any system of linear inequalities can be reduced to this standard form by simple substitutions.

Definition of a feasible solution

A set of x_j values that is nonnegative and satisfies the equations is a "feasible" solution and corresponds to a feasible but not necessarily optimal program. Such a solution is viewed as a point in n-dimensional space.

The constraint set—*i.e.*, the set of feasible solutions—has a property known as convexity. In general, a set S is convex if and only if it contains the entire line segment between any two of its elements. An element of a convex set S is called an extreme point if it does not lie on the line segment joining two other points of S. A line segment in S is called an edge if no point on it lies also on a line segment in S that crosses it. The constraint set of a linear program has only a finite number of extreme points.

If a linear programming problem has a unique optimal feasible solution, then it is at an extreme point; if not unique, there is an optimal solution at an extreme point. Thus, the optimal feasible solution is within a finite set of extreme points.

The simplex method. There are certain difficulties involved in the simple graphical approach to solutions. First, the graphical method of solution illustrated by the figure is useful only for systems of inequalities involving two variables. Practical problems often involve hundreds of equations, thousands of variables, however, and a computer is required. Second, the number of extreme points can be quite large. For example, for n equal to 2,000 and a constraint set defined by restricting the sum of x_i and y_j to be 1 for all nonnegative values of the variables and for all n values of j (see 2), there are 2^n, or about 10^{600}, extreme points. It is thus necessary to restrict the number of extreme points to be examined. This is done by using the simplex method. It works in the following way. It is assumed that an extreme point is known. (If no extreme point is given, a variant of the simplex method, called Phase I, is used to find one or determine that there are no feasible solutions.) Using the algebraic specification of the problem, it is easy to test whether that extreme point is optimal. If the test for optimality is not passed, then a movement along some edge to an adjacent extreme point is sought along which the value of the objective function increases at the fastest rate. Sometimes one can move along an edge and make the objective function value increase without bound. If this occurs, the procedure terminates with a prescription of the edge along which the objective goes to positive infinity. If not, a new extreme point is reached having at least as high an objective function value as its predecessor. The sequence described is then repeated. Termination occurs when an optimal extreme point is found or the unbounded case occurs.

The simplex method solves the numerical example given above in the following way: The problem is put into canonical form by converting the linear inequalities into equations by introducing slack variables $x_3 \geq 0$, $x_4 \geq 0$, $x_5 \geq 0$, and the variable x_0 for the value of the objective function; the problem may then be restated as that of finding nonnegative quantities $x_1, \cdots, x_5$ and the largest possible x_0 satisfying the resulting equations (see 3). One obvious solution of the resulting system of equations is found by setting $x_1 = 0$, $x_2 = 0$ (see 4), which corresponds to the extreme point at the origin in the figure. If x_1 (or x_2) were increased from zero while the other one x_2 (or x_1) is fixed at zero, the values of x_0, x_3, x_4, x_5 could be made to satisfy the equations, and the objective value x_0 would increase as desired. The variable x_2 produces the largest increase of x_0 per unit change; so it is used first. Its increase is limited by the nonnegativity requirement on the variables. In particular, if x_2 were increased beyond 5, x_4 would become negative.

Slack variables

At $x_2 = 5$, this situation produces a new solution (see 5) and corresponds to the extreme point (0, 5) in the figure. The system of equations is put into equivalent form (see 6) by solving for the variables x_0, x_2, x_3, x_5, which are nonzero in the above solution in terms of those variables now at zero; *i.e.*, x_1 and x_4. It is now apparent that an increase of x_1 while holding x_4 equal to zero will produce a further increase in x_0. The nonnegativity restriction on x_5 prevents x_1 from going beyond 5. The new solution (see 7) corresponds to the extreme point (5, 5) in the figure. Solving for x_0, x_1, x_2, x_3 in terms of the variables x_4, x_5 (which are currently at zero value) yields a final equivalent system of equations (see 8). It is to be noted that no variable can be changed from its present value and yield a feasible solution with a higher value than $x_0 = 15$ (see 9). Hence, an optimal solution is determined (see 10).

The dual problem. Associated with each linear programming problem is a second linear programming problem known as its dual. In this association, the original problem is referred to as the primal. Von Neumann provided a

$$
\begin{cases}
\text{maximize } p_1x_1 + \cdots + p_nx_n \\
\text{subject to } a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \geq 0 \\
\qquad\qquad a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \geq 0 \\
\qquad\qquad \cdots\cdots\cdots \\
\qquad\qquad a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \geq 0 \\
\qquad\qquad x_1 \geq 0, \cdots, x_n \geq 0
\end{cases} \tag{1}
$$

$$
x_j + y_j = 1, \qquad x_j \geq 0, \qquad y_j \geq 0 \quad \text{for } j = 1, \cdots, n \tag{2}
$$

$$
\begin{cases}
-x_0 + x_1 + 2x_2 = 0 \\
x_1 + x_3 = 8 \\
x_2 + x_4 = 5 \\
x_1 + x_2 + x_5 = 10
\end{cases} \tag{3}
$$

$$
\begin{cases}
x_0 = 0, x_1 = 0, x_2 = 0, x_3 = 8 \\
x_4 = 5, x_5 = 10
\end{cases} \tag{4}
$$

$$
(x_0, x_1, x_2, x_3, x_4, x_5) = (10, 0, 5, 8, 0, 5) \tag{5}
$$

$$
\begin{cases}
-x_0 + x_1 - 2x_4 = -10 \\
x_1 + x_3 = 8 \\
x_2 + x_4 = 5 \\
x_1 - x_4 + x_5 = 5
\end{cases} \tag{6}
$$

$$
(x_0, x_1, x_2, x_3, x_4, x_5) = (15, 5, 5, 3, 0, 0) \tag{7}
$$

$$
\begin{cases}
-x_0 - x_4 - x_5 = -15 \\
x_3 + x_4 - x_5 = 3 \\
x_2 + x_4 = 5 \\
x_1 - x_4 + x_5 = 5
\end{cases} \tag{8}
$$

$$
x_4 \geq 0, x_5 \geq 0, x_0 = 15 - x_4 - x_5 \leq 15 \tag{9}
$$

$$
x_1 = 5, x_2 = 5, x_3 = 3, x_4 = x_5 = 0 \tag{10}
$$

form for the primal problem (see 11) that yields an elegant statement of the dual problem (see 12).

An intimate relationship exists between a linear program and its dual. For any vector x^0 (see ANALYSIS: *Vector and tensor analysis*) satisfying the constraints of the primal and any vector y^0 satisfying the constraints of its dual, the primal objective value is never greater than the dual objective value, namely, $px \leq yb$. One implication of this relationship is that when the dual problem possesses at least one feasible vector, the primal objective function is bounded from above.

If the objective values satisfy $px^0 = y^0b$, then it is easy to show that x^0 and y^0 if feasible are optimal solutions of the primal and dual programs. If a vector x^0 is an optimal solution of the primal problem, there exists an optimal solution of the dual problem such that px^0 equals y^0b. Moreover, if $v^0 = b - Ax^0$ and $u^0 = p - y^0A$, then for optimal primal and dual feasible vectors x^0 and y^0, the relations $y_i^0 v_i^0 = 0$ for $i = 1, \cdots, m$ and $x_j^0 u_j^0 = 0$ for $j = 1, \cdots, n$ hold and are referred to as complementary slackness conditions. Thus, if x_j^0 is positive in an optimal solution, then the jth inequality constraint of the dual holds as an equality; *i.e.*, its slack, measured by u_j^0, satisfies $u_j^0 = 0$.

Complementary slackness conditions

NONLINEAR PROGRAMMING THEORY

A mathematical programming problem is nonlinear if the objective function $f(x)$ to be minimized, in which x is a vector, or any of the constraint functions are nonlinear.

Classification of the problems. One broad class of nonlinear programming problems is that concerning minimizing $f(x)$ subject to no constraints (the unconstrained problem); another is the linearly-constrained nonlinear programs that include as a special subclass quadratic programs concerned with the minimization of a quadratic function (see 13) subject to linear constraints. Another is the chemical equilibrium problem in which the Gibbs function, a measure of the free energy of a chemical system (see 14), is to be minimized subject to (linear) mass-balance equations (see 15) and nonnegativity conditions (see 16). The x_j in the problem represent the unknown number of molecules of different types in a system under constant temperature and pressure. In the mass-balance equations the b_i are the given number of atoms of various types; the a_{ij} are the number of atoms of type i in a single molecule of type j. The c_j in the Gibbs function are related to the constants in the law of mass action and are considered as constants only for dilute solutions.

Classes of nonlinear programming problems

Nonlinear programs are also classified according to whether defining functions have the appropriate convexity property. A function $F(x)$ defined on a convex set S is convex if (and only if) the set of points lying on or above its graph is convex. In analytic terms, convexity is equivalent to an inequality relation (see 17). The function $G(x)$ on S is concave if and only if $F(x) = -G(x)$ is convex. A useful property of any convex function $F(x)$ defined on the convex set S is that the set of all points in S such that $F(x) \leqq 0$ is a convex set. (But the set of vectors x, such that $F(x) = 0$, is not necessarily convex.)

When a quadratic function has the property that its quadratic part (see 18) is nonnegative for all choice of values for x_r, x_s, then $Q(x)$ is a convex function. Linear functions are both convex and concave.

If $f(x)$ and $h_i(x), \cdots, h(x)$ are convex functions, defined for each vector x appearing as the argument of the functions, then the problem of minimizing $f(x)$ subject to $h_i(x) \leq 0 (i = 1, \cdots, l)$ is called the convex programming problem. One important property of convex functions is the sufficiency of local conditions for identifying a minimum. If f is a convex function on the convex set S and x^0 is a point of S such that $f(x^0) \leq f(x)$ for all x in S that are sufficiently close to x^0, then x^0 is also a (global) minimum of f on S; that is, $f(x^0) \leq f(x)$ for all x in S. This property can be exploited in computational procedures aimed at calculating optimal solutions of convex programming problems.

Convex programming problem

Methods of solution. The question of identifying optimal solutions of nonlinear programs is often discussed assuming that the constituent functions are differentiable. The simplest such case is the unconstrained minimization of a differentiable function $f(x)$. If $x^0 = (x_1, \cdots, x_n)$ is an optimal solution, then the gradient vector of f at x^0 vanishes (see 19). Geometrically, this means that the tangent plane to the graph of f at the point $[x^0, f(x^0)]$ is horizontal. For a convex function f, the vanishing of the gradient is enough to guarantee that the point x^0 is a (global) minimum of f. But in general, the vanishing of the gradient can occur at points that are not even local minima, namely, at saddle points and local maxima.

The equality-constrained minimization problem is of the form: minimize $f(x)$ subject to $g_i(x) = 0$ for $i = 1, \cdots, m$ and $m < n$. If the functions f, $g_1, \cdots, g_m$ are differentiable, the attempt to isolate a (local) minimum point x^0 for this problem by finding numbers $\lambda_1, \cdots, \lambda_m$ such that $\nabla f(x^0) = \sum_{i=1}^{m} \lambda_i \nabla g_i(x^0)$ is called the method of Lagrange multipliers (see 20). In this approach, one looks for unconstrained (local) minima of an associated Lagrangian function (see 21).

When all the constituent functions are differentiable, there is an analogous result on the necessary conditions of optimality for inequality-constrained minimization problems (see 22). These necessary conditions are closely related to the Kuhn-Tucker conditions. The mathematicians H.W. Kuhn and A.W. Tucker of the United States showed in a now classic, nonlinear programming paper that if a certain regularity condition holds, a relation is true between the gradient of f and the gradients of subsidiary

(11) PRIMAL: Maximize px subject to $x \geqslant 0$, $Ax \leqslant b$ in which $A = [a_{ij}]$ is an $m \times n$ matrix, $x = (x_1, x_2, \cdots, x_n)^T$, $p = (p_1, p_2, \cdots, p_n)$ and $b = (b_1, b_2, \cdots, b_m)^T$.

(12) DUAL: Minimize yb subject to $y \geqslant 0$, $yA \geqslant p$, in which $y = (y_1, y_2, \cdots, y_m)$.

(13) $$Q(x) = \sum_{j=1}^{n} c_j x_j + \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_i x_j$$

(14) $$F(x) = \sum_{j=1}^{n} c_j x_j + \sum_{j=1}^{n} x_j \log\left(x_j \Big/ \sum_{j=1}^{n} x_j\right)$$

(15) $$\sum_{j=1}^{n} a_{ij} x_j = b_i, \qquad i = 1, \cdots, m$$

(16) $$x_j \geqslant 0, \qquad j = 1, \cdots, n$$

(17) $F(\lambda x + \mu y) \leqslant \lambda F(x) + \mu F(y)$ for all $x \in S$, $y \in S$ and $\lambda \geqslant 0$, $\mu \geqslant 0$ such that $\lambda + \mu = 1$.

(18) $$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_i x_j$$

(19) $$\nabla f(x^0) = \left[\frac{\partial f(x^0)}{\partial x_1}, \cdots, \frac{\partial f(x^0)}{\partial x_n}\right] = (0, \cdots, 0)$$

(20) If f and g_i are differentiable and the m gradient vectors $\nabla g_1(x^0), \cdots, \nabla g_m(x^0)$ are independent, then there exist numbers $\lambda_1, \cdots, \lambda_m$ such that the vector relation $\nabla f(x^0) = \lambda_1 \nabla g_1(x^0) + \cdots + \lambda_m \nabla g_m(x^0)$ holds. The numbers $\lambda_1, \cdots, \lambda_m$ are known as Lagrange multipliers.

(21) $$L(x, \lambda) = f(x) + \lambda_1 g_1(x) + \cdots + \lambda_m g_m(x)$$

functions h_j at an optimal solution x^0. In particular, if f and the functions h_j are convex, then a solution x^0 that satisfies stated conditions for some vector λ of multipliers is a global minimum.

Convex quadratic programming problems

Some methods for solving convex quadratic programming problems are closely related to the simplex method for linear programming and share with it the feature of terminating after only a finite number of iterations. Most of these methods make use of the Kuhn-Tucker conditions because the differentiability and regularity requirements are automatically satisfied. These characteristics are seldom found in nonlinear programming procedures in general. Instead of leading in a finite number of steps to an exact solution, they (at best) yield a point close to a point at which a local minimum is attained.

One important ancestor of nonlinear programming methods is the method of steepest descent due to the 19th-century French mathematician Augustin-Louis Cauchy. Its original application was to solving systems of nonlinear equations. For example, to solve the system of simultaneous equations $g_i(x) = 0$, $(i = 1, \cdots, m)$ a sum of squares of the functions appearing in the equations can be formed (see 23). An attempt can be made to obtain its unconstrained minimum. For every vector x, the functional value $f(x)$ is nonnegative because it is the sum of squares. Consequently, x^0 is a solution if and only if $f(x^0) = 0$.

The method is easily stated: From a starting point (trial solution) $x^{(1)}$, compute the direction of steepest descent, $-\nabla f(x^{(1)})$. Using a fixed step size, symbolized by the Greek letter sigma, σ, minimize the function f along the line segment of length σ issuing from $x^{(1)}$ in the direction of steepest descent. The point $x^{(2)}$ that minimizes $f(x)$ along this line segment becomes the new trial solution. In general, the preceding step is repeated at each trial solution $x^{(k)}$ to obtain the next one $x^{(k+1)}$.

The method of steepest descent is valid under the assumption that the function f is continuously differentiable, and all points satisfying $f(x) \leq f(x^{(1)})$ lie within a fixed finite distance from the origin. The aim of the procedure is actually to locate a point x^* satisfying the necessary local conditions of optimality: $\nabla f(x^*) = 0$. Because such a point might be reached only in a limiting sense, it is customary to terminate the computational process when a point $x^{(k)}$ is reached at which the absolute value $|\, \partial f(x^{(k)})/\partial x_j \,|$ is less than a prescribed value for $j = 1, \cdots, n$. If the function f being minimized is convex, a point x^* at which $\nabla f(x^*)$ is zero must be a global minimum.

Another method deals with the linearly constrained nonlinear programming problem: Minimize $f(x)$ subject to $Ax = b$, $x \geq 0$ in a manner that uses both linear programming and the Kuhn-Tucker conditions. It is assumed that the objective function $f(x)$ is continuously differentiable. It is then natural to replace f by its linear approximation at a feasible point and solve the associated linear program. For example, if $x^{(k)}$ is a feasible point, f can be replaced by its linear approximation (see 24). Assuming that $y^{(k)}$ is the solution of this linear program, a search can then be performed to minimize f along the line joining $x^{(k)}$ to $y^{(k)}$. The point $x^{(k+1)}$ at which the latter minimum is attained becomes the new trial solution at which the linear approximation to f can be formed for the repetition of the process. The computation can be interrupted if a point x^0 is reached at which the solution point y^0 of the associated linear programming problem satisfies $\nabla f(x^0)\,(y^0 - x^0) \geq 0$.

(22)
$$\begin{cases} \text{Minimize } f(x) \text{ subject to } h_j(x) \leq 0 \text{ for } j = 1, \cdots, l. \\ \text{Then } x^0 \text{ is a locally minimizing solution only if there} \\ \text{exist multipliers } \lambda_0, \lambda_1, \cdots, \lambda_l \text{ not all zero such that} \\ \lambda_0 \nabla f(x^0) = \lambda_1 \nabla h_1(x^0) + \cdots + \lambda_l \nabla h_l(x^0) \text{ and} \\ h_j(x^0) \leq 0, \quad \lambda_j \geq 0, \quad \lambda_j \cdot h_j(x^0) = 0 \text{ for } j = 1, \cdots, l. \end{cases}$$

(23)
$$f(x) = [g_1(x)]^2 + \cdots + [g_m(x)]^2$$

(24)
$$F(y) = f(x^k) + \nabla f(x^k)(y - x^k)$$

It can be shown that the point x^0 must be optimal when the objective function f is convex.

Of the many algorithms for the solution of nonlinear programming problems, none is known to be superior to the generalized reduced-gradient method. Its roots lie in the reduced-gradient method (for linearly constrained nonlinear programs) and ultimately in the simplex method itself. (G.B.D./R.W.Co.)

Generalized reduced-gradient method

Cybernetics

The presentation to follow deals only with the principles of cybernetics and includes reference to mathematical aspects such as information theory, automata theory, and cybernetic systems. The technological aspects of cybernetics are multifold and can be found in INFORMATION PROCESSING AND INFORMATION SYSTEMS and AUTOMATION.

DEFINITIONS OF THE TERM CYBERNETICS

The term cybernetics comes from the ancient Greek word *kybernētikos* ("good at steering"), referring to the art of the helmsman. In the first half of the 19th-century, the French physicist André-Marie Ampère, in his classification of sciences, suggested that the still nonexistent science of the control of governments be called cybernetics. This term was soon forgotten until used by the United States mathematician Norbert Wiener as the title for his book published in 1948. In that book Wiener made reference to an 1868 article by the British physicist James Clerk Maxwell on governors and pointed out that "governor" is derived, via Latin, from the same Greek word that gives rise to the term cybernetics. The date of Wiener's publication is generally accepted as the date of the birth of cybernetics as an independent science. Wiener defined cybernetics as "the science of control and communications in the animal and machine." This definition relates cybernetics closely first of all with the theory of automatic control and with physiology, particularly the physiology of the nervous system. Subsequently, the computer and the areas of mathematics related to it (*e.g.*, mathematical logic) had a great influence on the development of cybernetics. The reason is that the computer can be used not only for automatic calculation but also for all conversions of information, including various types of information processing used in control systems. This ability made two different views of cybernetics possible. The narrower view, common in the Western countries, defines cybernetics as the science of control of complex systems of various types (technical, biological, and social). In many countries (for example, in the United States), particular emphasis is given to those aspects of cybernetics that are used in the generation of control systems in technology and in living organisms. In addition to cybernetics, the science of computers and the general rules of information processing have been developed in the Western countries (English—computer science; French—*informatique*).

The broader interpretation of the subject of cybernetics that prevailed in Russia and the other former Soviet republics for many years includes not only control but all forms of information processing. This definition includes Western computer science as one of the component parts of cybernetics.

PRINCIPLES

Information theory. It should be emphasized that the concept of information and its conversion is an important first concept of cybernetics no matter how its subject is defined. Information arises with any act of selection or limitation of possibilities of selection. For example, if it is known that a room can contain 100 men, then the message that the room at a given moment contains 50 men or that the number of people in the room is less than 40 carries information. The message that the number is less than 200 carries no information, because the first, fixed selection is in no way thus limited. In developing the principle of limitation of selection, Claude E. Shannon, an electrical engineer in the United States, introduced the method of changing the quantity of information.

Information is transmitted by signals. In pure cybernetics,

The nature and types of signals

the physical nature of the signals is completely disregarded. It is important only that the signals can be differentiated from each other. The form of the set of possible signals and the nature of their changes are also significant. For example, a signal carrying information on the number of persons in a room cannot take on fractional values. The values of these signals change in jumps. On the other hand, a signal giving information of the temperature of the air in the room cannot change its value, say from 19° C to 20° C, without passing through all intermediate values. Signals of the first type are called discrete and of the second type, continuous. The same terms are used in relation to the information represented by these signals.

In describing continuous information within a certain accuracy, it can always be reduced to discrete information. The usual method of representation of discrete information is as a finite sequence of signals selected from a certain fixed finite set of signals called the (abstract) alphabet, for example the set of letters in the Latin alphabet, the set of decimal numbers, etc. One important but simple fact is the possibility of representing any discrete information in the form of sequences of signals of only two different types, as is done, for example, in the dots and dashes of the well-known Morse telegraph code. Problems of various forms of representation of discrete information make up the subject of a special division of theoretical cybernetics called encoding theory.

Automata as information converters. Encoding is the simplest form of information conversion. In the general case, these conversions are performed by so-called information converters, or data processors. An information converter, referred to in cybernetics as a machine or automaton (see AUTOMATA THEORY), is a device that converts a certain set of signals called the input signals to another set, the set of output signals. The input signals arrive at the converter through the input channels, the output signals leave through the output channels. Furthermore, in the general case, the processor can store signals in the form of a certain set of parameters, the different values of which determine the internal state of the processor, that is, the condition of its memory.

A processor is called continuous if all of the parameters that define it (memory, input, and output signals) are continuous. In the case of discrete processors, all of these signals must be discrete. Processors dealing both with continuous and with discrete signals are called hybrid processors.

The changes in output signals and in the internal state of any processor depend generally both on the input signals and on the internal state of the processor at a given time. For a processor without memory, the values of the output signals at any moment in time depend only on the values of the input signal at the same moment in time. If all of these dependences are fixed by fully defined, unambiguous functions (*e.g.,* single-valued), the processor is called deterministic. If there are random dependences, the processor is called probabilistic.

Information processors in which no functional components and no internal structural parts other than those described above can be distinguished are called elementary. Nonelementary processors, also called cybernetic systems, consist of networks composed of elementary (or in any case simpler) processors. A network is constructed by connecting some (or all) of the output channels of the processors composing the network to some (or all) of their input channels. A portion of the input and output channels of the elementary processors may be appropriately connected to the input and output signals of the entire system.

The concept of elementariness or nonelementariness of a processor is relative and depends on the depth of penetration into the subject. For example, if interest is confined to the functions of the brain, not in its structure, it can be analyzed as an elementary processor (although it is characterized by a tremendous number of parameters). It is known, however, that the brain is an extremely complex system, composed of more elementary processors, the neurons. In turn, upon transition to the molecular level, the neurons themselves can be regarded as extremely complex cybernetic systems.

The cybernetic approach to systems study

Study of systems. The study of systems and particularly of the complex probabilistic systems is one of the most important tasks in cybernetics. The specific feature of the cybernetic approach to the study of systems is that it abstracts itself from the actual nature of the elements of the system and the signals circulating in it. This allows a general mathematical apparatus and general methods of investigation to be developed, suitable for systems of varying nature and purpose. In the study of continuous cybernetic systems, a system of ordinary differential equations is particularly significant.

Theory of algorithms. In the discrete case, the modus operandi is the so-called theory of algorithms. An algorithm is an arbitrary, finite system of rules of any nature, allowing expressions in any alphabet to be converted to new expressions in the same (or any other) alphabet. So-called algorithmic languages are used for precise descriptions of the rules for these conversions (see AUTOMATA THEORY: *Classification of automata*).

It is extremely important that the presence in an algorithmic language of a relatively small number of means of expression can give it the property of universality. Universality of an algorithmic language means the possibility of expressing in this language any conversion of discrete information that can be defined using a finite number of rules, that is, expressed as a certain algorithm in any other algorithmic language.

The fact of existence of universal algorithmic languages means that it is possible to construct universal discrete information processors. To do this, it is sufficient to fix upon some concrete alphabet and to construct a memory device capable of storing any word (expression) in the alphabet and a device capable of performing all of the elementary rules of a certain universal algorithmic language.

Universal digital computers are such universal information processors if the limitations resulting from their finite memory volume are not considered. Until computers appeared, the only universal data processor was the natural information processor, the human brain.

The availability of technical universal data processors is highly significant for the development of cybernetics in at least two ways. First of all, this fact opens unlimited possibilities for the automation not only of the physical, but also of the mental activity of man, and this automation does not require that new technical devices be invented for each case. It is sufficient to study and describe the rules defining the type of activity to be automated and to program them; *i.e.,* express them in one of the algorithmic languages used in existing computers.

Mathematical modelling. Second, the use of universal data processors—computers—gives cybernetics a method of scientific investigation of systems that is new in principle—so-called mathematical modelling. Until this method appeared, scientists actually had only two different methods of study: experimental and theoretical. In the former, experiments were performed only with the system itself, or with an actual, physical model of the system. In the second, it was necessary to be able to solve equations describing the system.

Mathematical modelling occupies an intermediate position between these two methods: there is no necessity to construct an actual physical model of the system. It is replaced by a mathematical model; *i.e.,* a description of the system in some algorithmic language. There is also no need to solve complex mathematical problems related to this description (*i.e.,* to solve systems of differential equations). The description of the system is simply entered into a computer that models the behaviour of the system (*i.e.,* provides precise descriptions of the system) under various conditions defined according to the purpose of the research assignment.

This method allows the scientist to produce a full description of complex systems, the individual parts of which are studied by different people or even in different sciences. One example is the human organism. Its individual parts (circulatory system, digestive system, nervous system, glandular system, etc.) are studied by different specialists, even though the parts are closely related to each other.

Modelling the human organism

To produce a mathematical model of the organism, first

of all it must be divided into individual parts and the state of each part described by some system of parameters. Among these parameters might be continuous quantities (for example, the percent content of sugar in the blood) or discrete quantities (for example, qualitatively differing levels of secretory function of the liver). The next step is description of the relationships between the separate parts. One of the most typical forms of relationship for this example is expressed by a sentence such as: When organ A shifts from state m_1 to state m_2, and organ B shifts from state n_1 to state n_2 after k days (weeks, months), organ C will change its state from l_1 to l_2 with probability p. The specialists must describe all of the relations of this type with which they are familiar (called simple relationships) to form the required mathematical model. Operations with this model in the computer permit the establishment of the complex relationships (*i.e.*, the influence of organs on other organs not directly, but through still other organs). This type of model can be used to study versions of the development of various diseases, various methods of treatment, etc.

Mathematical modelling can in principle be performed in any universal data processor, including the human brain. The brain, however, being a very complex system itself, is comparatively poorly suited for such routine but laborious work as the modelling of complex systems. Consequently, only the appearance of the computer provided a qualitative jump and made it possible to perform effective study of really complex systems in various areas of knowledge. Actually, the concern is with coarse qualitative criteria, differentiating simple systems from complex; if the structure and behaviour of a system can be studied by a single man in a reasonable time, the system is called simple. If the efforts of many persons and the use of special technical equipment (computers) are required to draw the whole picture, the system is called complex.

Cybernetic systems. *Self-teaching mechanisms.* Complex tasks arise in so-called self-teaching systems, in which attempts to achieve a certain final goal lead to changes in the methods of its attainment, and the setting of various intermediate goals. One of the simplest examples of this type of system is the so-called Shannon's mouse. It is a moving automaton placed in a maze. The final purpose of the automaton is to find the "food" placed at some point in the maze.

Shannon's mouse

At first, the mouse uses a simple trial and error algorithm, bumping into the walls until an exit is found. During the process of this type of search, it "learns" and memorizes the plan of the maze. After once reaching the goal (by chance), the mouse will use an entirely different, much more economical, and seemingly "intelligent" algorithm for reaching the goal in a subsequent experiment, based on "knowledge" of the layout of the maze.

Self-teaching is one type of self-improvement of control systems. In this case, the improvement occurs without changing the structure of the system. It is also possible, however, that the upper levels of a system may change the structure of lower levels in order to improve their functioning. This type of self-improvement is naturally called self-development.

Methods have been designed for the quantitative definition of self-improvement. In this case, in pure theory, the boundary between self-teaching and self-development disappears. In practice, however, the difference between these types of self-improvement is usually rather clear.

Biocybernetics. Many control systems in biology are constructed according to the principles of self-teaching and self-development. For example, in all probability, only a small portion of the organization of the human brain is determined genetically. Everything else is produced as a result of the effects of many, rather effective, mechanisms of self-improvement.

Highly complex examples of self-development can be found in the processes of biological evolution, as well as various social processes.

For systems of the high degree of complexity characteristic of living organisms and human society, an ever greater role is played not just by control processes, but by cognition processes. These systems have highly perfected systems of sensors and effectors (for example, the eye and the hand), capable of recognizing complex patterns and performing widely varied actions. The central portion of the system has a multilevel structure; the upper links of the structure develop abstract concepts and recognize the deeper regularities existing both in the system itself and in its surroundings.

By the last quarter of the 20th century, cybernetics had achieved significant success in the solution of the problem of pattern recognition, automation (by computer) of the processes of logical conclusion, formation of new concepts, and other problems. This success has allowed the creation of universal robots controlled by computers and imitating rather complex forms of conscious behaviour. (Ed.)

Control theory

GENERAL BACKGROUND

As long as human culture has existed, control has always meant some kind of power over man's environment. Cuneiform fragments suggest that the control of irrigation systems in Mesopotamia was a well-developed art at least by the 20th century BC. There were some ingenious control devices in the Greco-Roman culture, the details of which have been preserved. Methods for the automatic operation of windmills go back at least to the Middle Ages. Large-scale implementation of the idea of control, however, was impossible without a high-level of technological sophistication, and it is probably no accident that the principles of modern control started evolving only in the 19th century, concurrently with the Industrial Revolution. A serious scientific study of this field began only after World War II and is now a major aspect of what has come to be called the second industrial revolution.

Although control is sometimes equated with the notion of feedback control (which involves the transmission and return of information)—an isolated engineering invention, not a scientific discipline—modern usage tends to favour a rather wide meaning for the term; for instance, control and regulation of machines, muscular coordination and metabolism in biological organisms, prosthetic devices; also, broad aspects of coordinated activity in the social sphere such as optimization of business operations, control of economic activity by government policies, and even control of political decisions by democratic processes. Scientifically speaking, modern control should be viewed as that branch of system theory concerned with changing the behaviour of a given complex system by external actions. (For aspects of system theory related to information, see below.) If physics is the science of understanding the physical environment, then control should be viewed as the science of modifying that environment, in the physical, biological, or even social sense.

Examples of feedback control

Much more than even physics, control is a mathematically-oriented science. Control principles are always expressed in mathematical form and are potentially applicable to any concrete situation. At the same time, it must be emphasized that success in the use of the abstract principles of control depends in roughly equal measure on the status of basic scientific knowledge in the specific field of application, be it engineering, physics, astronomy, biology, medicine, econometrics, or any of the social sciences. This fact should be kept in mind to avoid confusion between the basic ideas of control (for instance, controllability) and certain spectacular applications of the moment in a narrow area (for instance, manned lunar travel).

EXAMPLES OF MODERN CONTROL SYSTEMS

To clarify the critical distinction between control principles and their embodiment in a real machine or system, the following common examples of control may be helpful. There are several broad classes of control systems, of which some are mentioned below.

Machines that cannot function without (feedback) control. Many of the basic devices of contemporary technology must be manufactured in such a way that they cannot be used for the intended task without modification by means of control external to the device. In other words, control is introduced after the device has been built; the same

effect cannot be brought about (in practice and sometimes even in theory) by an intrinsic modification of the characteristics of the device. The best known examples are the vacuum-tube or transistor amplifiers for high-fidelity sound systems. Vacuum tubes or transistors, when used alone, introduce intolerable distortion, but when they are placed inside a feedback control system any desired degree of fidelity can be achieved. A famous classical case is that of powered flight. Early pioneers failed, not because of their ignorance of the laws of aerodynamics, but because they did not realize the need for control and were unaware of the basic principles of stabilizing an inherently unstable device by means of control. Jet aircraft cannot be operated without automatic control to aid the pilot, and control is equally critical for helicopters. The accuracy of inertial navigation equipment (the modern space compass) cannot be improved indefinitely because of basic mechanical limitations, but these limitations can be reduced by several orders of magnitude by computer-directed statistical filtering, which is a variant of feedback control.

The role of computers in feedback control

Control of machines. In many cases, the operation of a machine to perform a task can be directed by a human (manual control), but it may be much more convenient to connect the machine directly to the measuring instrument (automatic control); *e.g.*, a thermostat (temperature-operated switch) may be used to turn on or off a refrigerator, oven, air-conditioning unit, or heating system. The dimming of automobile headlights, the setting of the diaphragm of a camera, the correct exposure for colour prints may be accomplished automatically by connecting a photocell or other light-responsive device directly to the machine in question. Related examples are the remote control of position (servomechanisms), speed control of motors (governors). It is emphasized that in such case a machine could function by itself, but a more useful system is obtained by letting the measuring device communicate with the machine in either a feedforward or feedback fashion.

Control of large systems. More advanced and more critical applications of control concern large and complex systems the very existence of which depends on coordinated operation using numerous individual control devices (usually directed by a computer). The launch of a spaceship, the 24-hour operation of a power plant, oil refinery, or chemical factory, the control of air traffic near a large airport, are well-known manifestations of this technological trend. An essential aspect of these systems is the fact that human participation in the control task, although theoretically possible, would be wholly impractical; it is the feasibility of applying automatic control that has given birth to these systems.

Biocontrol. The advancement of technology (artificial biology) and the deeper understanding of the processes of biology (natural technology) has given reason to hope that the two can be combined; man-made devices should be substituted for some natural functions. Examples are the artificial heart or kidney, nerve-controlled prosthetics, and control of brain functions by external electrical stimuli. Although definitely no longer in the science-fiction stage, progress in solving such problems has been slow not only because of the need for highly advanced technology but also because of the lack of fundamental knowledge about the details of control principles employed in the biological world.

Robots. On the most advanced level, the task of control science is the creation of robots. This is a collective term for devices exhibiting animal-like purposeful behaviour under the general command of (but without direct help from) man. Industrial manufacturing robots are already fairly common, but real breakthroughs in this field cannot be anticipated until there are fundamental scientific advances with regard to problems related to pattern recognition and the mathematical structuring of brain processes.

PRINCIPLES OF CONTROL

The scientific formulation of a control problem must be based on two kinds of information: (A) the behaviour of the system (*e.g.*, industrial plant) must be described in a mathematically precise way; (B) the purpose of control (criterion) and the environment (disturbances) must be specified, again in a mathematically precise way.

Information of type A means that the effect of any potential control action applied to the system is precisely known under all possible environmental circumstances. The choice of one or a few appropriate control actions, among the many possibilities that may be available, is then based on information of type B; and this choice, as stated before, is called optimization.

Basic task

The task of control theory is to study the mathematical quantification of these two basic problems and then to deduce applied-mathematical methods whereby a concrete answer to optimization can be obtained. Control theory does not deal with physical reality but only with its mathematical description (mathematical models). The knowledge embodied in control theory is always expressed with respect to certain classes of models, for instance, linear systems with constant coefficients, which will be treated in detail below. Thus control theory is applicable to any concrete situation (*e.g.*, physics, biology, economics) whenever that situation can be described, with high precision, by a model that belongs to a class for which the theory has already been developed. The limitations of the theory are not logical but depend only on the agreement between available models and the actual behaviour of the system to be controlled. Similar comments can be made about the mathematical representation of the criteria and disturbances.

Once the appropriate control action has been deduced by mathematical methods from the information mentioned above, the implementation of control becomes a technological task, which is best treated under the various specialized fields of engineering. The detailed manner in which a chemical plant is controlled may be quite different from that of an automobile factory, but the essential principles will be the same. Hence further discussion of the solution of the control problem will be limited here to the mathematical level.

To obtain a solution in this sense, it is convenient (but not absolutely necessary) to describe the system to be controlled, which is called the plant, in terms of its internal dynamical state. By this is meant a list of numbers (called the state vector) that expresses in quantitative form the effect of all external influences on the plant before the present moment, so that the future evolution of the plant can be exactly given from the knowledge of the present state and the future inputs. This situation implies, in an intuitively obvious way, that the control action at a given time can be specified as some function of the state at that time. Such a function of the state, which determines the control action that is to be taken at any instant, is called a control law. This is a more general concept than the earlier idea of feedback; in fact, a control law can incorporate both the feedback and feedforward methods of control.

In developing models to represent the control problem, it is unrealistic to assume that every component of the state vector can be measured exactly and instantaneously. Consequently in most cases the control problem has to be broadened to include the further problem of state determination, which may be viewed as the central task in statistical prediction and filtering theory. Thus, in principle, any control problem can be solved in two steps: (1) Building an optimal filter (so-called Kalman filter) to determine the best estimate of the present state vector; (2) determining an optimal control law and mechanizing it by substituting into it the estimate of the state vector obtained in step 1.

In practice, the two steps are implemented by a single unit of hardware, called the controller, which may be viewed as a special-purpose computer. The theoretical formulation given here can be shown to include all other previous methods as a special case; the only difference is in the engineering details of the controller.

The mathematical solution of a control problem may not always exist. The determination of rigorous existence conditions, beginning in the late 1950s, has had an important effect on the evolution of modern control, equally from the theoretical and the applied point of view. Most important is controllability; it expresses the fact that some

Controllability

kind of control is possible. If this condition is satisfied, methods of optimization can pick out the right kind of control using information of type B.

CONTROL OF LINEAR SYSTEMS

Systems with constant coefficients. The preceding considerations may be illustrated much more directly by taking the special case of linear systems with constant coefficients. The point of view and most of the results described below originated after 1950, but the beginnings of the theory are more than 100 years old dating from Maxwell's work published in 1868. In fact, this class of models is the only one for which a reasonably complete mathematical theory of control exists so far.

One may denote by $y(t)$ the number giving the value of the output of the plant at time t; similarly, the number $u(t)$ denotes the value of the input to the plant at the same time. A classical way to describe the behaviour of the plant (information of type A) is to assume that the input and output are related by a linear differential equation with constant coefficients $\alpha_1, \cdots, \alpha_n$ (see 25), which is called the open-loop equation of the plant. In this case the state vector is the list $(y(0), \cdots, y^{(n-1)}(0))$ consisting of the output at time $t=0$ and the first $n-1$ derivatives with respect to time of the function $y(t)$ at $t=0$. Under these circumstances, the control law is a linear combination of the state variables, with coefficients $\beta_1, \cdots, \beta_n$ (see 26).

The fact that this control law is linear in the state is an assumption needed to assure the manageability of mathematical machinery for the problem. It is desirable to assume also that the coefficients $\beta_1, \cdots, \beta_n$ are all constants so that the same formula can be used at any time. [A minus sign in the control law (26) expresses the classical idea of negative feedback; in the general theory, this has become merely a notational convention.]

Substitution of the control law into the linear differential equation of the plant (25) leads to the closed-loop differential equation for the controlled output of the plant (see 27). Here the coefficients γi are the differences $\alpha_i - \beta_i$. So the effect of control is to replace the coefficients α_i of the open-loop equation with the γ_i of the closed-loop equation. Since the β_i in the control law are arbitrary, it follows that arbitrary changes can be effected in the dynamical behaviour of the plant by means of control. In particular, it is possible to make a plant stable by means of the control law because the criterion for stability is given by the positiveness of the Routh–Hurwitz determinants (see 28). These conditions can of course always be met since the γ_i can be arranged to have any values.

The criterion for stability

(25) $$\frac{d^n y}{dt^n} + \alpha_1 \frac{d^{n-1} y}{dt^{n-1}} + \cdots + \alpha_n y = u(t)$$

(26) $$-u(0) = \beta_n y(0) + \beta_{n-1} \dot{y}(0) + \cdots + \beta_1 y^{(n-1)}(0)$$

(27) $$\frac{d^n y}{dt^n} + \gamma_1 \frac{d^{n-1} y}{dt^{n-1}} + \cdots + \gamma_n y = 0$$

(28) $$\gamma_1 > 0, \quad \begin{vmatrix} 1 & \gamma_2 \\ \gamma_1 & \gamma_3 \end{vmatrix} > 0, \cdots$$

(29) $$\frac{dx}{dt} = Fx + gu(t), \qquad y(t) = \langle h, x(t) \rangle$$

(30) $$F = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ -\alpha_n & -\alpha_{n-1} & -\alpha_{n-2} & \cdots & -\alpha_1 \end{bmatrix}, \quad g = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ 1 \end{bmatrix}$$

(31) $$\det[g \quad Fg \quad F^2 g \quad \cdots \quad F^{n-1} g] \neq 0$$

It is necessary to comment on this seemingly overly general (but indeed correct) result from two points of view First the result assumes that the plant description (see 25) is exactly true and that the state list $(y(t), \cdots, y^{(n-1)}(t))$ is known at every time t. These assumptions are never exactly correct, but they can be approximated in many cases with sufficient accuracy by restricting the output $y(t)$ to small deviations about a given steady-state value and by constructing a statistical filter to estimate the derivatives of the function $y(t)$ with respect to t. Second, the model (see 25) does not represent the most general linear system with constant coefficients; in fact, this model, because of its special form, implicitly possesses the property of controllability that may not hold for the general model in the linear, constant coefficient class.

The most general system of this class (with one input and one output) is given by a pair of state-vector equations (see 29); the symbol x denotes the state vector (a list of n arbitrary numbers), F is an $n \times n$ matrix, g is an n-vector, h is an n-vector, and $\langle , \rangle$ is the inner product; the symbols $u(t)$ and $y(t)$ have the same meaning as above. The equations of the plant discussed above (see 25) can always be transformed into the general case (see 29), but not necessarily vice versa.

Transforming the plant equations discussed above to the general form gives the canonical matrices F, g (shown in 30). If these expressions are substituted into the controllability criterion (see 31) for the general system (29), it is seen that the special model for the plant that was used above (25) is always controllable. Thus, it is clear that the true significance of the celebrated controllability concept can be grasped only after the control problem has been formulated in the state-vector style. In the classical development of control theory, too much emphasis was placed on special models for the plant (25); the advances since 1950 to a large extent resulted from a recognition that the most general model corresponding to the basic restrictive hypotheses of linearity and constant coefficients is given by the state-vector equations (29); this recognition eventually led to the discovery that for the general class there is a nontrivial controllability condition (31).

The explicit form of the controllability condition is of great practical and philosophical importance. It is a generic condition: it is satisfied by almost all systems of the linear, constant-coefficient class. Because the state-vector equations (29) accurately represent most physical systems making only small deviations about their steady-state behaviour, it follows further that in the natural world control in the small is, in principle, almost always possible. This fact of nature is the theoretical basis of practically all the presently existing control technology. On the other hand, little is known at present, in the scientific sense, about the ultimate limitations of control when the models in question are not linear. In particular, it is not known under what conditions control is possible in the large; that is, for arbitrary deviations from existing conditions. This lack of scientific knowledge should be kept in mind in assessing often exaggerated claims by economists and sociologists in regard to a possible improvement in human society by governmental control.

Control limitations in nonlinear modelling

Optimal control. To reduce the arbitrariness of the control law (see 26), it is necessary to use information of type B. This is done, conceptually, in the following way. It is assumed (after a change of variables, if necessary) that the state $x=0$ represents the ideal condition of the plant to be controlled. Any given state $x(0)$ not equal to zero at time $t=0$ is regarded as an undesirable deviation, attributed to unavoidable disturbances acting on the system. The task of control is to restore the ideal condition $x(t)=0$ in the best possible way. The specification of a suitable optimality criterion is frequently dictated by mathematical expediency.

The long mathematical history of the calculus of variations has suggested that the most natural criterion to consider is the minimization of an integral from zero to infinity, which measures the deviation of $x(t)$ from zero as

well as the cost of control. For example, the coefficients in the control law may be explicitly determined by minimizing a quadratic integral (see 32) by the choice of the function $u(t)$ with respect to solutions of the open-loop differential equation (see 25). The applied-mathematical technique for solution is well developed and requires the solution of a nonlinear differential equation of the Riccati type, which can be conveniently carried out by digital computer. It can be shown also that this general procedure includes older techniques based on semi-quantitative engineering analyses—for instance, the idea of maximizing the loop gain.

In many respects, optimization theory represents a satisfactory solution of the control problem (in some cases even outside the linear case), but basic difficulties remain in regard to the appropriate choice of the criterion represented by the quadratic forms in the performance index (see 32).

$$(32) \qquad \int_0^\infty [Q(x(t)) + R(u(t))]\,dt \qquad (Q, R \text{ are quadratic forms})$$

Optimal filtering and state estimation. At least in the linear case, the solution of the state-estimation problem can be accomplished by methods similar to that for optimal control; this follows from the well-known Duality Principle, which states that the dual equations (matrices and vectors replaced by their transposes) of control optimization are the equations for state estimation. The optimal filter is then a dynamical system similar to (25), in which the principal task is to determine a dual control law similar to (26).

In practice, the state estimation problem is much more important than the control problem; approximately 90–95 percent of practical controllers are devoted to state estimation, only about 5 percent being required for the implementation of the control law.

NONLINEAR CONTROL SYSTEMS

Whenever the controlled variable $y(t)$ is allowed to have large deviations from the steady state, the linear constant-coefficient model will cease to represent the plant accurately because of intrinsic nonlinearities involved in the description of most natural dynamical phenomena. Equally important are intentionally introduced nonlinearities that result from reasons of economy, simplicity and reliability of engineering or from ignorance of the fact that the savings achieved by nonlinear control devices may be negated by factors resulting from the greater intrinsic difficulty of control. A good example of unavoidable nonlinearities in the object to be controlled is a space vehicle whose rockets, at the current state of technology, can be controlled only by turning them on or off. Continuously variable control action (which is desirable) in this case is technologically impossible; the high cost of rocket engines, however, justifies extremely sophisticated computer-based technology to achieve optimal control. The purpose of the computer is then simply one of switching the rockets on or off; the difficulty of the control lies in the extreme precision with which the time must be determined when this takes place.

Example of a nonlinearity

When the basic mathematical data can be stated in the conventional optimization framework, effective methods are available for solving the optimal control problem even in the nonlinear case. The methods for doing this are an extension of the classical methods of the calculus of variations. These methods, however, have yielded little theoretical insight, and their straightforward application becomes prohibitively expensive for large-scale systems.

Even less satisfactory is the status of the nonlinear optimal filtering and state estimation problem, which, as has been noted, is a critical part of the general solution of the control problem. Because of nonlinearity, the Duality Principle no longer applies; even the formulation of the problem is controversial.

Nonlinear systems represent not a special case but simply everything that is not subjected to the special assumption of linearity; in a sense "nonlinear" is synonymous with "unknown." Scientific progress will undoubtedly occur by singling out special classes of systems subject to restrictive structural assumptions other than linearity. (R.E.K.)

BIBLIOGRAPHY

Linear and nonlinear programming. The history, theory, and applications of linear programming may be found in GEORGE B. DANTZIG, *Linear Programming and Extensions* (1963, reissued 1974). G. HADLEY, *Linear Programming* (1962), is also informative. L.R. FORD, JR., and D.R. FULKERSON, *Flows in Networks* (1962), stands as the classic work on the subject. An alternate source is T.C. HU, *Integer Programming and Network Flows* (1969). JAMES K. STRAYER, *Linear Programming and Its Applications* (1989), addresses problems and applications. G.V. SHENOY, *Linear Programming* (1989), focuses on the field from the point of view of management and is accessible to readers without a background in mathematics. HOWARD KARLOFF, *Linear Programming* (1991), provides an introduction from the perspective of theoretical computer science. Another useful work is AMI ARBEL, *Exploring Interior-Point Linear Programming: Algorithms and Software* (1993). One of the pathbreaking books on linear and nonlinear programming is G. ZOUTENDIJK, *Methods of Feasible Directions* (1960). LEON S. LASDON, *Optimization Theory for Large Systems* (1970), deals with linear and nonlinear programming problems. F.H. CLARKE, *Optimization and Nonsmooth Analysis* (1990), serves as a good exposition of optimization in terms of a general theory. G.B. DANTZIG and B.C. EAVES (eds.), *Studies in Optimization* (1974), is a volume of survey papers by leading experts. OLVI L. MANGASARIAN, *Nonlinear Programming* (1969, reprinted 1982), deals exclusively with theory; while WILLARD I. ZANGWILL, *Nonlinear Programming* (1969), primarily discusses algorithms. MOKHTAR S. BAZARAA, HANIF D. SHERALI, and C.M. SHETTY, *Nonlinear Programming: Theory and Algorithms,* 2nd ed. (1993), is an introductory textbook. STEPHEN A. VAVASIS, *Nonlinear Optimization* (1991), focuses on complexity issues. H.TH. JONGEN, P. JONKER, and F. TWILT, *Nonlinear Optimization in IR^n,* 2 vol. (1983–86), offers an extensive and advanced approach to the subject from the standpoint of differential geometry. From various conferences have come J. ABADIE (ed.), *Nonlinear Programming* (1967), and *Integer and Nonlinear Programming* (1970); GEORGE B. DANTZIG and ARTHUR F. VEINOTT, JR. (eds.), *Mathematics of the Decision Sciences,* 2 vol. (1968); and JEAN-PAUL PENOT (ed.), *New Methods in Optimization and Their Industrial Uses* (1989)—all feature papers on a wide range of subjects.

Cybernetics. THOMAS M. COVER and JOY A. THOMAS, *Elements of Information Theory* (1991), provides an introduction to information theory. NORBERT WIENER, *Cybernetics,* 2nd ed. (1961), presents a very general discussion. VIKTOR M. GLUSHKOV, *Introduction to Cybernetics,* trans. from Russian (1966), is also of interest. Additional sources include STAFFORD BEER, *Cybernetics and Management,* 2nd ed. (1967); JIŘÍ KLÍR and MIROSLAV VALACH, *Cybernetic Modelling,* trans. from Czech (1967); and CONSTANTIN VIRGIL NEGOITA, *Cybernetics and Applied Systems* (1992).

Control theory. A good book on modern control and system theory is RICHARD E. BELLMAN, *Dynamic Programming* (1957, reissued 1972). More mathematical treatments include L.S. PONTRYAGIN (L.S. PONTRIAGIN) *et al., The Mathematical Theory of Optimal Processes* (1962, reprinted 1986); and E.B. LEE and L. MARKUS, *Foundations of Optimal Control Theory* (1967, reprinted 1986). Control theory in the wider context of system theory is treated in R.E. KALMAN, P.L. FALB, and M.A. ARBIB, *Topics in Mathematical System Theory* (1969), especially chapter 2. ROGER W. BROCKETT, *Finite Dimensional Linear Systems* (1970), surveys the fundamental problems of description and optimization of linear constant-coefficient systems. A historical overview of feedback devices may be found in OTTO MAYR, *The Origins of Feedback Control* (1970; originally published in German, 1969). D. STANLEY-JONES and K. STANLEY-JONES, *The Kybernetics of Natural Systems* (1960), covers information on biocontrol. LESLIE M. HOCKING, *Optimal Control* (1991), gives an accessible introduction, including various applications. ATLE SEIERSTAD and KNUT SYDSÆTER, *Optimal Control Theory with Economic Applications* (1987), presents a mathematically rigorous development of optimal control theory from an economic perspective. E.J. MCSHANE, "The Calculus of Variations from the Beginning Through Optimal Control Theory," *SIAM Journal on Control and Optimization,* 27(5):916–939 (1989), surveys the field with emphasis on recent developments, especially the work of Pontryagin. H.T. BANKS (ed.), *Control and Estimation in Distributed Parameter Systems* (1992), is another informative work. (G.B.D./R.W.Co./R.E.K./Ed.)

Ōsaka-Kōbe Metropolitan Area

The Ōsaka-Kōbe metropolitan area is the second largest urban and industrial agglomeration in Japan. It is located on Ōsaka Bay in west-central Honshu at the eastern end of the Inland Sea. The cities of Ōsaka and Kōbe are at the centre of what is called by geographers the Hanshin Industrial Zone; as a result of the expansion of the urban area along the Inland Sea and northeast toward the city of Kyōto, the region is now included in the larger Keihanshin Industrial Zone. Neither of these zones is a political entity, but the larger of the two corresponds to the Kansai, one of Japan's traditional cultural areas. The Kansai, a region of ancient cities to the west (*sai*) of the mountain barrier (*kan*) near Mount Fuji, is the birthplace of the earliest Japanese state. It is an area of historically dense population that until well into the 20th century was the most industrialized and economically advanced part of Japan.

Ōsaka is the capital of Ōsaka urban prefecture (*fu*), an administrative division that includes the city of Ōsaka, a number of smaller cities, and large rural areas. Kōbe is the capital and largest city of Hyōgo prefecture (*ken*) and one of Japan's chief ports. There are many satellite industrial and residential cities around the two central cities.

This article is divided into the following sections:

Physical and human geography 227
- The landscape 227
  - The city site
  - Climate
  - The city layout
- The people 228
- The economy 229
  - Industry
  - Commerce
  - Transportation
- Administration and social conditions 229
  - Government
  - Public utilities
  - Health
  - Education
- Cultural life 229

History 229
- Ōsaka 229
  - Ancient and medieval periods
  - Early modern and modern periods
- Kōbe 230

Bibliography 230

Physical and human geography

THE LANDSCAPE

The city site. The city of Ōsaka is situated on the delta of the Yodo River. To the east of the central city, Ōsaka Castle, originally built by Toyotomi Hideyoshi, stands on a northern extension of the upland (about 65 feet [20 metres] above sea level) that rises in the southern part of Ōsaka urban prefecture to more than 3,000 feet. The metropolitan area spreads over the deltas of the Yodo, Yamato, and other rivers and into their diluvial uplands. The area is bounded by the Ikoma Mountains in the east, the Izumi Mountains in the south, and the Rokkō Mountains in the northwest. The southwestern boundary of Ōsaka Bay is formed by Awaji Island. On the northwestern shore of the bay is Kōbe, above which rises the granite peak of Mount Rokkō (3,058 feet). The region is geologically unstable.

Earthquakes

Although earthquakes occur only infrequently, they can be highly destructive; notable severe quakes include one that struck the area in 1596 and another that devastated Kōbe and neighbouring cities in 1995.

The coastline between the two cities has been altered by reclamation for port facilities and industries. Along the coast and in the uplands are the best residential areas of Kōbe and the cities of Ashiya, Nishinomiya, Takarazuka, Ikeda, and Itami. On the delta of the Kanzaki River, just west of Ōsaka, is the city of Amagasaki, a centre of heavy industry. To the north of Ōsaka are the cities of Toyonaka, Suita, and Ibaraki. Above them, on Senri Hill, are new towns developed since the 1960s. Northeast of Ōsaka, along the Yodo River, are the industrial and residential cities of Takatsuki, Moriguchi, Neyagawa, and Hirakata. To the east of Ōsaka are the cities of Kadoma, Higashiōsaka, and Yao. To the southeast are Fujiidera, Tondabayashi, Matsubara, and others, most of them old historical towns. To the southwest, on the coastal plain, are Sakai, Izumi-Ōtsu, Kaizuka, Kishiwada, and Izumi-Sano, some of them industrial and others residential. Urbanization extends to Nara, 25 miles (40 kilometres) east of Ōsaka, and to Kyōto, 25 miles northeast. A dense network of railways and roads winds throughout the area.

Climate. The region has a temperate climate. The annual mean temperature is about 60° F (16° C), and the annual rainfall averages 54 inches (1,370 millimetres). The temperature in August is often 86° F (30° C) or more, with no breeze from the sea at night. The January mean is about 40° F (4° C), and snow falls several times in winter. The rainy seasons are in June–July and September–October. In September the region is usually struck by one or two typhoons. The greatest typhoon disaster in the area's history occurred in 1934, when 3,000 people were killed. During the rainy season of June–July 1938, huge landslides from the Rokkō Mountains buried wide areas of Kōbe, and floods took the lives of 870 persons.

The citizens of Ōsaka once took pride in its smoky atmosphere as a mark of industrial progress, calling it the

Lawrence Smith—Photo Researchers



Ōsaka Castle.

"Capital of Smoke" ("Kemuri no miyako"); but by the mid-1970s the city's smog and air pollution were seen as harmful, and since then there has been noticeable improvement in the region's air quality. Other environmental problems receiving attention include water pollution and subsidence (sinking) of the earth in the Amagasaki region.

The city layout. *Street patterns.* The streets of central Ōsaka are laid out on a grid plan, but the rest of the city is a patchwork of planned grids and rambling streets. The north-south axis is Midō-suji ("Midō Street"), connecting Ōsaka railway station in the north and Namba station in the south. The east-west axis is Chūō Ōdōri ("Central Boulevard"), running from the Central Pier of the Port of Ōsaka in the west to the foot of the Ikoma Mountains in the east. Parallel to Midō-suji is the narrow Shinsaibashi-suji, the central shopping district. Dotombori, at the south end of Shinsaibashi-suji, is a crowded theatre and restaurant area.

Central business district

The central business district is the northern part of the downtown area. Nakanoshima, an island formed by arms of the Yodo River, contains City Hall, the Central Civic Hall, the Bank of Japan, and the headquarters of the Asahi Press and several large businesses. Until World War II the traditional commercial centres were Semba and Shimanouchi streets, where old-style white-walled shops with family quarters behind them had existed for centuries.

There are two large *sakariba* (amusement districts) in the city. Kita ("The North") is located just south of Ōsaka railway station, where the city's highest-priced land is found. Kita has a complex of high-rise office buildings and a large underground shopping centre. Minami ("The South") has many theatres and restaurants. Ōsaka's industrial areas are on the lower Yodo delta and in the eastern and northeastern parts of the city.

The street pattern of Kōbe is governed by its location between the mountains and the shore. Main streets run roughly east and west, crossed by short north-south streets and occasional longer streets going up into the hills. The central shopping street, Motomachi, runs between the Sannomiya and Kōbe railway stations. The central business district is near the harbour.

Settlement patterns. The Ōsaka area has been settled and built upon since prehistoric times; kitchen middens and pottery shards have been found dating to at least 7000 BC. There are many ancient burial mounds from the Tumulus period (*c.* AD 250–*c.* 550), and the contents of some of these have helped document the presence of settlers from the Korean peninsula. The ancient villages of the rice farmers were on the marshy plains, while the palaces, shrines, and temples were located on higher ground. Medieval settlements were in the uplands. Some modern residential areas are on the sites of several former settlements. Thus, Tezukayama, a residential development in Ōsaka south of the castle, is built over a number of ancient mounds.

The central part of Ōsaka is now primarily commercial; since 1920 there has been a migration from the city to the suburbs, helped along by private railway companies that have made suburban building land available along their rights-of-way. The Hankyū Electric Railway was particularly instrumental in developing the city of Toyonaka northwest of Ōsaka. Two of the large postwar housing developments are Senri New Town and Senboku New Town, started in 1961 and 1965, respectively.

About two-thirds of Ōsaka urban prefecture's dwellings are apartment houses. Much of Ōsaka and nearly all of Kōbe were destroyed during World War II, and inner-city areas are now occupied mostly by Western-style multistoried buildings. Traditional architectural styles can still be seen, however, in the Tekijuku, a school of foreign studies (18th century), and the Kōnoike (1708; rebuilt 1853) in Ōsaka, as well as in some residential houses in Kōbe. Examples of early Western-influenced architecture survive on Nakanoshima, including the Bank of Japan building (1903) and the Ōsaka Prefectural Library (1904). The redevelopment of the central business districts of both cities since the 1960s has produced many large, modern, and architecturally innovative buildings: notable examples include the office and hotel complex in front of Ōsaka railway station; the Semba Centre Building, which although only four floors in height extends about three-fifths of a mile along Chūō Ōdōri and is constructed under an elevated expressway and over a subway; and the convention centre built on the man-made Port Island in Kōbe Harbour.

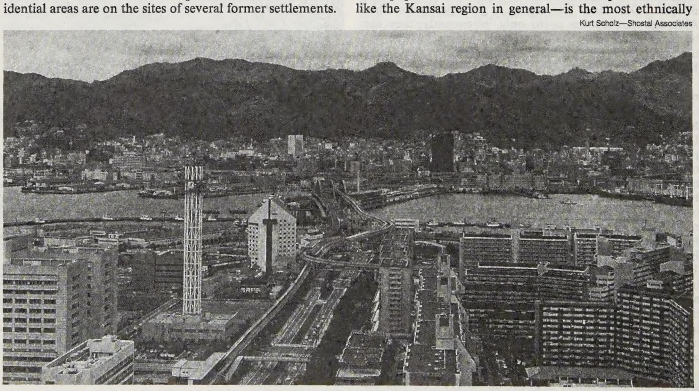
THE PEOPLE

Demographic trends

Demographic trends in the Ōsaka-Kōbe metropolitan region parallel two major national trends: sustained urban population growth through rural-urban migration and suburbanization. Thus, the region's population has been swelled for more than a century by a continuous stream of people moving in from rural areas; but the population of the city of Ōsaka—after reaching a peak of about 3,150,000 in the mid-1960s—has declined, as people have migrated from the city to the suburbs. In addition, the city's highest density is not in the centre but in the peripheral wards, because the population decrease has been greatest in the central wards. By contrast, the population of the city of Kōbe has increased steadily, although, like Ōsaka, there has been a loss of population in the inner city and high increases in peripheral wards.

The population of the Ōsaka-Kōbe metropolitan area—like the Kansai region in general—is the most ethnically

Kurt Scholz—Shostal Associates



Central Kōbe and Rokkō Mountains seen from Port Island.

diverse of Japan. Included are the country's largest concentrations of ethnic Koreans, most of whom are the Japan-born descendants of Koreans who migrated to Japan during the period (1910–45) when Korea was a Japanese colony and who are classed as resident aliens; Okinawans, who legally are Japanese citizens but who often are treated as internal aliens; and *burakumin*, the term being a euphemism for descendants of an outcaste group that was once legally, though not genetically, distinct from the general Japanese population. All three groups are subject to discrimination in education, employment, marriage, and housing. In Kōbe there are also sizable communities of Chinese, Indians, and Westerners, whose presence contributes an international flavour to the city's culture. (S.K./R.P.T.)

THE ECONOMY

Industry. Ōsaka was once known as the Manchester of the Orient because of its great textile industry. In the post-war era, such industries as machinery, iron and steel, fabricated metals, and chemicals became important. While these industries have continued to play a role in the city's economy, various light industries and such high-technology industries as computers and software have become significant. Between Ōsaka and Kōbe are several other industrial cities, the largest of which is Amagasaki. The major industries in Kōbe are shipbuilding and steel.

Commerce. In the past Ōsaka merchants greeted one another in the mornings with the query, "Are you making money?" Contemporary merchants and executives continue to greet each other by asking about the state of their businesses. The salutations reflect the traditional importance of commercial enterprise in Ōsaka. The city holds a substanial share of the nation's wholesale trade.

Ōsaka is Japan's second largest financial centre; it is headquarters for some of the world's largest banks, and it has one of its major stock exchanges. Together with Kōbe it is a leading port for foreign trade.

Railways

Transportation. Ōsaka is an important junction point of the national railway network, handling traffic between the Tokyo region to the northeast and regions farther west and south. Following the privatization of the formerly government-controlled Japanese National Railways (JNR) in 1987, Ōsaka became the headquarters of the West Japan Railway Company (JR Nishi Nihon), which operates passenger service in western Honshu. The region is also served by Shinkansen "bullet" express trains and the national freight rail system, both of which also are now private entities. In addition, other privately owned railroads provide suburban commuter and regional interurban passenger service between Ōsaka and Kōbe and their suburbs and to other major cities in central Honshu, particularly Kyōto and Nagoya. Rail lines also run inland from Kōbe to rural areas of Hyōgo prefecture north of the city. Ōsaka's subway system, started before World War II, underwent expansion in the late 1960s, the early 1980s, and the late 1990s, and construction on a new subway line began in 2000. Kōbe's first subway line was opened in the mid-1980s, and a second line went into operation in 2001.

Because of heavy automobile traffic, Ōsaka's main streets are one-way. A network of surface and elevated expressways runs through the central parts of Ōsaka and Kōbe, linking the two cities together as well as joining them to the national expressway system. Ordinary highways also span the whole region. Kōbe and Ōsaka are both international and domestic ports; passenger ships, freighters, and car ferries sail to the islands of Shikoku and Kyushu and to various ports of the Inland Sea. International air service is provided by Kansai International Airport, built on a man-made island in Ōsaka Bay and opened in 1994, while domestic flights use the older facility located a few miles northwest of the city at Itami.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The city of Ōsaka is the capital of Ōsaka urban prefecture, which consists of 31 cities and 13 towns and villages. It is also the centre of the Kinki region (*chihō*), which consists of the seven prefectures of Ōsaka, Kyōto, Hyōgo, Nara, Wakayama, Shiga, and Mie. Various prefectural and regional institutions have their main offices in Ōsaka. Both Ōsaka and Kōbe are divided administratively into wards (*ku*), each of which has an appointed mayor. The wards elect representatives to respective city councils headed by an elected mayor.

Public utilities. Ōsaka's main source of water is the Yodo River; Kōbe has several reservoirs on the slope of the Rokkō Mountains. Sewage services are adequate in the central urban areas. Electricity is available everywhere, and gas is available in most city areas.

Health. Medical care in Ōsaka centres on the hospitals of Ōsaka University and of the Ōsaka City University and on several other medical institutions. Other hospitals and health centres are distributed throughout the region.

Education. In Ōsaka and Hyōgo prefectures there are more than 100 universities and junior colleges. Ōsaka University, Ōsaka University of Foreign Studies, Ōsaka University of Education, Kōbe University, and Hyōgo University of Education are national universities. Public municipal institutions include Ōsaka City University and the University of Ōsaka Prefecture. Kansai University in Ōsaka and Kwansei Gakuin University in Nishinomiya are the oldest and largest private universities in the area.

CULTURAL LIFE

Ōsaka and Kyōto have long been leading centres of culture. Ōsaka is famous for its cuisine and is called the "kitchen of the land." Some noted examples of its cuisine include *oshizushi* (slices of raw fish spread over a base of rice, compressed on all sides in a wooden mold to form a cake, which is then cut into squares), *takoyaki* (fried dumplings with octopus filling), and *okonomiyaki* (pancakes with a filling of fish, shrimp, or vegetables). Ōsaka, strongly influenced by its merchant cultural traditions, has a more bustling, democratic tone than Kyōto and is more informal and direct in cultural and social expression than Tokyo. Ōsaka is the birthplace of *bunraku* (Japanese puppet theatre) and the modern home of *manzai*, a traditional form of two-person standup comic dialogue that traces its origins to the Nara period (710–794) and contrasts with *rakugo*, the sitdown comic monologues of Tokyo. Traditional and modern Japanese drama and music are performed at theatres and halls in the metropolitan area, as are Western music, operas, and plays. There are numerous museums, galleries, and libraries. Ōsaka is home to two national newspapers, the *Mainichi* and the *Asahi* (originally founded in Ōsaka, it now also has a Tokyo headquarters).

Parks

Green space in the city of Ōsaka is scarce, but recreational opportunities abound. The important parks include Nakanoshima, Ōsaka Castle, Tsurumi Ryokuchi, Nagai, and Tennoji, the latter with a zoo and botanical gardens. The suburbs have many historical sites and large recreation areas. Besides the spacious man-made Hattori Ryokuchi and Meiji no Mori Minoo parks, there are the recreational areas of the Kii Peninsula on the Pacific, the beaches of the Inland Sea, the historical cities of Nara and Kyōto, and scenic Lake Biwa, near Kyōto. At Kōbe, Mount Rokkō (in Seto-naikai National Park) can be ascended by road or by cable car; there is a golf course at the top and ponds for swimming. There are four professional baseball teams in the metropolitan area, and the national high school baseball championships are played each summer in Nishinomiya. The town of Takarazuka, northwest of Ōsaka, has been developed as an amusement centre; it houses the Girls Opera and Dancing Theatre. In 1970 the Japan World Exposition (Expo 70) was held near Senri New Town; Expo Memorial Park now holds the National Museum of Ethnology, the National Museum of Art, and a recreation area.

History

ŌSAKA

Ancient and medieval periods. The plain of Ōsaka was settled in Paleolithic times and by about AD 300 was a political centre. Among the many ancient burial mounds in the Ōsaka area is that ascribed to the semilegendary emperor Nintoku; the largest tomb of the Tumulus period,

the 5th-century structure is surrounded by three moats and occupies some 80 acres (32 hectares). Ancient Naniwa—in what is now Ōsaka—was the site of palace or capital complexes intermittently from the early 5th to the mid-7th century, but in 710 it lost its position to Nara, the first "permanent" national capital.

When Kyōto became the imperial capital in 794, land and water routes between Ōsaka and Kyōto were improved. The reclamation of the delta of the Yodo River allowed the building of new settlements, including Watanabe, which became a provincial capital and port during the Middle Ages. South of Ōsaka, on the eastern shore of the bay, is Sakai, which had emerged as a port town by the 14th century. There is evidence that, like some medieval European towns, it was self-governing in the 15th and 16th centuries, run by its leading merchants until they capitulated to the warlord Oda Nobunaga in 1569. These merchants grew wealthy from Sakai's lucrative domestic and foreign trade; under their patronage Sakai became a centre of the arts after Kyōto was devastated in the Ōnin War (1467–77). Sakai was also a centre of Christian proselytizing in the late 16th and early 17th centuries; the accounts of Jesuit missionaries tell of the wealth and cosmopolitan flavour of the city at that time.

In 1496—in the midst of a century of civil war—Rennyo, chief priest of the militant True Pure Land (Jōdo Shin) sect of Buddhism, selected a site near the mouth of the Yodo River for a fortress temple. Completed in 1532, this structure, the Ishiyama Hongan Temple, became the nucleus of a major town that was destroyed in 1580 by Nobunaga, after a siege of many years. Nobunaga's successor, Toyotomi Hideyoshi, built a great castle on the site with massive stone walls and broad moats; the castle town that developed around it was the origin of present-day Ōsaka.

Ōsaka's height of influence

From this base Hideyoshi brought the whole of Japan under his control, and Ōsaka was the seat of national power until his death in 1598.

Early modern and modern periods. The castle and town were badly damaged and depopulated during Tokugawa Ieyasu's siege of 1614–15, in which he eliminated Hideyoshi's heir and consolidated his power as shogun. Succeeding shoguns rebuilt the castle and town, and during the rest of the Tokugawa period (1603–1867) Ōsaka was a directly administered shogunal city. Unlike other towns of the period, Ōsaka was not a political centre and therefore was not dominated by the samurai (warrior) class. Instead, it became the country's main commercial city; feudal lords from throughout Japan established warehouses for their tax rice along the city's canals, and rice was traded actively. Many other goods were traded in Ōsaka—which had some 380 wholesale houses by 1679—and the city became an expanding commercial and manufacturing centre. These activities stimulated the rapid monetization of the regional economy.

As it grew more prosperous, Ōsaka became a centre of the cultural renaissance of the Genroku period (late 17th–early 18th century). Dramatic forms such as bunraku (puppet theatre) and kabuki prospered, and new genres of prose fiction arose, the styles and themes of which catered to the tastes of urban commoners and marked a shift in cultural arbitration away from the samurai class. During the 18th century, however, Ōsaka's position as cultural leader was lost to Edo (now Tokyo), but the city remained an educational centre, with schools in classical studies and in medicine. In the mid-19th century, when Japan was still closed to most Westerners, the Dutch language and Western science were studied by the Japanese in Ōsaka.

Ōsaka remained preeminent both as a port and as a centre of industry until World War II. Much of the city was destroyed by aerial bombardment during the war, however, including the castle (which subsequently was re-created). Postwar economic growth became focused largely in the Tokyo-Yokohama metropolitan area. The communist revolution in China deprived Ōsaka of its important China trade until the early 1970s, while the increasing economic role of the national government tended to encourage industrial location in the Tokyo-Yokohama area.

KŌBE

The history of Kōbe is as old as that of Ōsaka. In ancient times the name Kōbe was applied to a small fishing village separated by the Minato River from the town of Hyōgo, the chief port of the area. Hyōgo, also known as Ōwada and Muko, was an important port for trade with China and Korea as early as the 8th century. For many centuries it continued to be Japan's chief port for foreign trade, prospering especially during the 15th and 16th centuries, and the government maintained patrol boats there to control piracy in the Inland Sea. Briefly during the 12th century Taira Kiyomori made it the capital instead of Kyōto.

During the Tokugawa period, Hyōgo served as the outer port of Ōsaka until in 1868 it was reopened to foreign trade. Soon it was outstripped and absorbed by Kōbe, which has a deeper harbour. The combined ports have been called the port of Kōbe since the establishment of the Kōbe customhouse in 1872. Hyōgo and Kōbe were incorporated as the city of Kōbe in 1889. The many foreigners who settled there in the 19th century gave it an international and cosmopolitan atmosphere.

Incorporation of Kōbe

The size of the city increased in the late 19th and early 20th centuries through the absorption of adjacent communities. During World War II, air raids destroyed much of the city. It was rebuilt quickly after the war, its size again increasing by annexation. Kōbe has become one of the largest cities in Japan. Its port facilities, which have undergone tremendous expansion since the war, have been combined administratively with those of Ōsaka since the early 1970s. The earthquake that struck the region in January 1995 destroyed large areas of Kōbe and several of its suburbs, caused damage on nearby Awaji Island, killed some 5,500 people, and left thousands more homeless. Kōbe's port facilities and transportation systems were also severely damaged. Relief and reconstruction efforts began immediately. Local and national government agencies, Japan's Self-Defense Force (SDF), other Japanese cities, and other nations all joined in the effort to alleviate the suffering of survivors, rebuild the city, and to return life in the city to normal. In the cold conditions of winter, public buildings such as the city hall and schools were quickly transformed into shelters, the SDF set up tent cities in parks and school yards, and government authorities erected large prefabricated structures to house many families. In the rebuilding process, roads were widened and parks expanded. The rebuilding of larger structures incorporated the latest anti-earthquake technology, and many older wooden homes were replaced with concrete structures. Although reconstruction efforts progressed rapidly, many people continued to live in shelters and tents for several months or longer. Almost all of the reconstruction work was completed after 5 years.

BIBLIOGRAPHY. Two publications by the Association of Japanese Geographers are useful: *Japanese Cities: A Geographical Approach* (1970), for the academic study of postwar urban Japan; and *Geography of Japan* (1980), especially ch. 12–18, which contains scholarly analyses of Japanese urban development. WILLIAM B. HAUSER, *Economic Institutional Change in Tokugawa Japan: Ōsaka and the Kinai Cotton Trade* (1974), analyzes Ōsaka's premodern economic role; and JAMES L. MCCLAIN and OSAMU WAKITA (eds.) *Osaka: The Merchant's Capital of Early Modern Japan* (1999), describes the city's early modern history. *Osaka and Its Technology* (semiannual) includes essays on urban development and public works. A novel by JUNICHIRŌ TANIZAKI, *The Makioka Sisters* (1957, reissued 1983; originally published in Japanese, 3 vol., 1949), provides an excellent if romanticized view of life in the Ōsaka-Kōbe region before World War II. PAT TUCKER SPIER (ed.), *The River Without Bridges: An Encounter with the Japanese Buraku* (1986), discusses the civil rights of the *burakumin* in Ōsaka. The 1995 Kōbe earthquake is depicted in T.R. REID, "Kobe Wakes to a Nightmare," *National Geographic*, 188 (1): 112–136 (July 1995). JOHN FREDERICK ASHBURNE, *The Best of Kansai: Kyoto, Osaka, and Kobe* (1998), is a good guidebook for the city. (S.K./R.P.T./Ed.)

Pacific Islands

The Pacific Islands is an expression commonly accepted as including all of those islands in the Pacific Ocean that are collectively referred to as Melanesia, Micronesia, and Polynesia, also sometimes known as Oceania. This usage rules out the Australian island continent, the Asia-related Indonesian, Philippine, and Japanese archipelagoes, and the Ryukyu, Bonin-Volcano, and Kuril island arcs that project seaward from Japan. Neither does the term encompass the Aleutian chain connecting Kamchatka and Alaska nor such isolated islands of the Pacific Ocean as Juan Fernández off the coast of South America.

Although the Pacific Ocean makes up nearly one-third of the Earth's surface, the Pacific Islands discussed in this article add up to a little less than 500,000 square miles (1,300,000 square kilometres) of land area. New Guinea, the largest island in the world after Greenland, represents 70 percent of this total, and New Zealand accounts for 20 percent. The remaining 10 percent of the land area of the Pacific is divided among more than 10,000 scattered islands.

The Pacific Islands lie mainly in the area bounded by latitudes 23° N and 27° S and longitudes 130° E and 125° W. Exceptions to this are New Zealand, which lies in the southern temperate zone, and Easter Island, which stands in isolation at longitude 109° W, almost halfway to South America. (Australia and New Zealand are treated in separate articles, and Hawaii is treated in the article UNITED STATES OF AMERICA.)

The article is divided into the following major sections:

The region 231
  Physical and human geography 231
  History and cultural development 240
Traditional cultures of the Pacific Islands 245
  General considerations 245
  Melanesian cultures 247
  Micronesian cultures 253
  Polynesian cultures 258
Major political units of Melanesia 264
  Fiji 264
  New Caledonia 266
  Papua New Guinea 270
  Solomon Islands 274
  Vanuatu 275
Major political units of Micronesia 277
  Guam 277
  Kiribati 278
  Marshall Islands 279
  Micronesia 280
  Nauru 281
  Northern Mariana Islands 283
  Palau 285
  Tuvalu 287
Major political units of Polynesia 288
  American Samoa 288
  Cook Islands 289
  Easter Island 290
  French Polynesia 292
  Samoa 294
  Tonga 296
  Wallis and Futuna 298
Bibliography 299

THE REGION

For convenient reference, the Pacific Islands are customarily divided into three ethnogeographic groupings. The great arc of islands located north and east of Australia and south of the Equator is called Melanesia (from the Greek words *melas,* "black," and *nēsos,* "island") after the predominantly dark-skinned peoples of New Guinea, the Bismarcks, Solomon Islands, Vanuatu (formerly the New Hebrides), New Caledonia, and Fiji. North of the Equator and east of the Philippines is another island arc that ranges from Palau (Belau) and the Marianas in the west through the Carolines and Marshalls all the way to Kiribati (formerly the Gilbert Islands). This is Micronesia, so named because of the smaller size of these islands and atolls. In the eastern Pacific, and largely enclosed within a huge triangle formed by Hawaii in the north, New Zealand to the south, and Easter Island far to the east, are the "many" (*poly-*) islands of Polynesia. Other components of this widely scattered collection are Samoa, Tonga, French Polynesia (including the Society, Tuamotu, and Marquesas Islands), and the Cook Islands. In this, the last section of the Pacific Ocean to be inhabited, the islanders share a cultural tradition that relates them closely to many Fijians. Fiji, indeed, is actually a transitional territory between Melanesia and Polynesia.

Since the 16th century, the Western world has shown an interest in the Pacific Islands that has been expressed in the activities of explorers, scientists, artists and writers, missionaries, commercial entrepreneurs, and imperialistic statesmen. The variety of the Pacific's environments, both physical and biotic, continues to be a laboratory for experimenting in social and cultural adaptation. Though insularity has often dominated this process, its effect has been offset by the opportunities for human contact and exchange in many directions across the ocean's expanse. Since the 20th century, the islands and their inhabitants have continued to attract international interest, although for new reasons, such as their strategic significance in the relationships of the world powers in Europe, Asia, and America. Attention has also centred on the problems created for Pacific islanders by nature's limitation of land and resources in the face of expanding populations and rising standards of living.

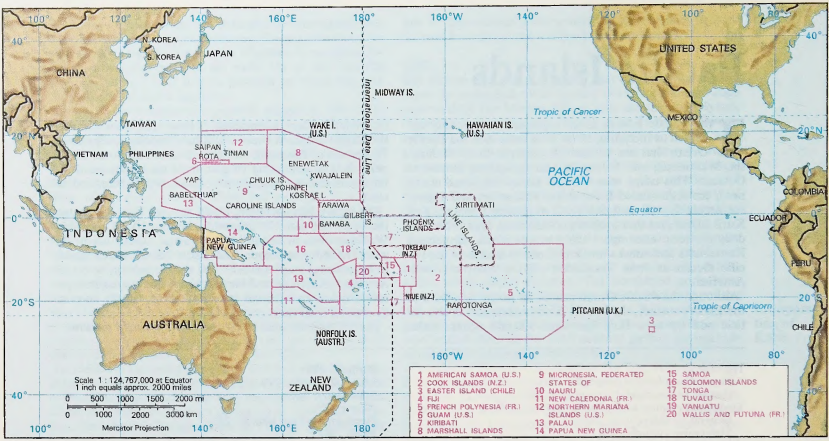
Physical and human geography

THE LAND

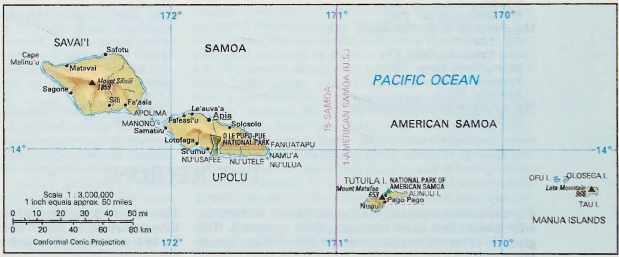
The island ecosystem. To know what it is like to live on a Pacific island, the intermixture of physical and biological characteristics of the particular island must be considered. Each of the myriad ecological systems in the Pacific is a unique complex of living organisms and their nonliving environment. Each is a functional system of interacting components that tends toward an equilibrium that is never quite achieved. The limited size of most Pacific islands makes it probable that almost any change, whether by human action or by some natural agency, will have repercussions elsewhere within the ecosystem. The landform, climate, soils, vegetation, and animal life all are elements to which people who live on an island must relate, for they, too, occupy a niche in the total ecological scheme.

Continental and oceanic islands

Relief. The islands may be classified as either continental or oceanic. The former are associated with the ancient continental platforms of Asia and Australia, now partially submerged. Oceanic islands, located eastward in the deeper Pacific basin, are differentiated as high volcanic-based islands or low coral islands and atolls. A coral island may be single, or two or more coral islets may be

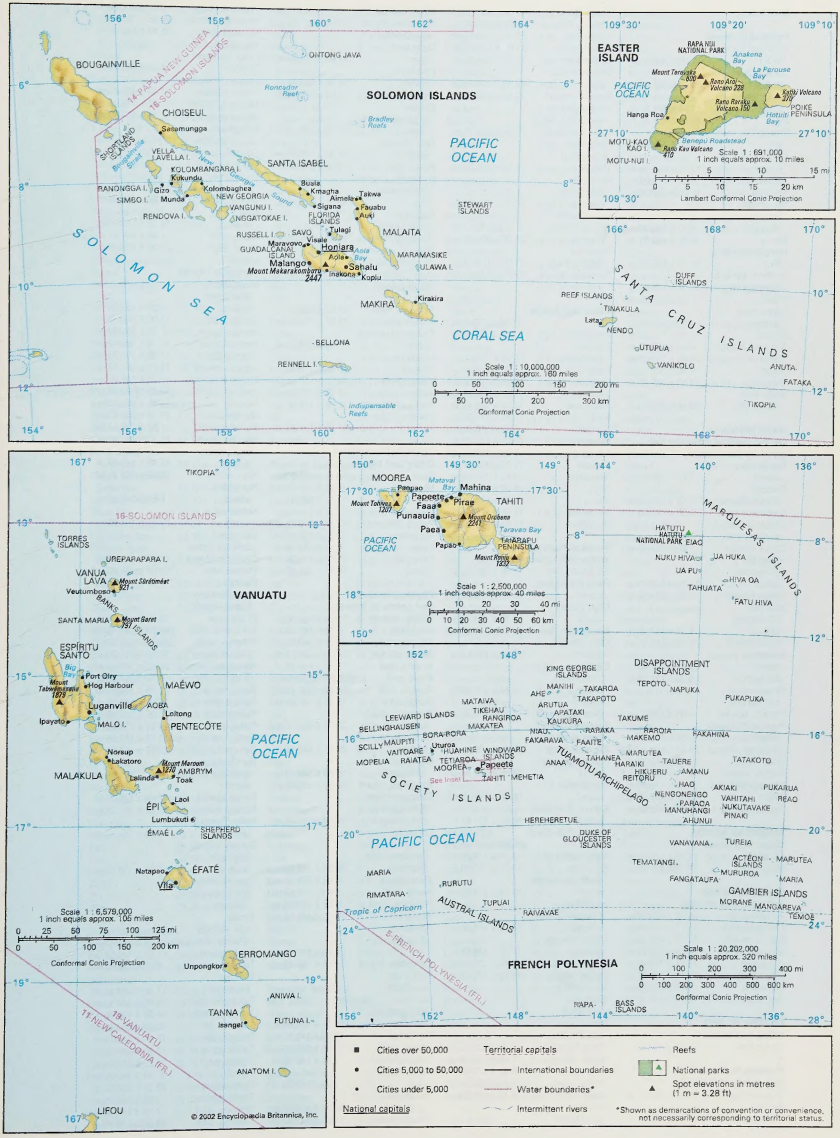

N. KOREA
S. KOREA
JAPAN
CHINA
TAIWAN
PHILIPPINES
VIETNAM
INDONESIA
AUSTRALIA
NEW ZEALAND
PAPUA NEW GUINEA
CAROLINE ISLANDS
MARSHALL ISLANDS
GILBERT IS.
PHOENIX ISLANDS
LINE ISLANDS
TOKELAU (N.Z.)
NIUE (N.Z.)
RAROTONGA
PITCAIRN (U.K.)
NORFOLK IS. (AUSTR.)
MIDWAY IS.
WAKE I. (U.S.)
HAWAIIAN IS. (U.S.)
UNITED STATES
MEXICO
COLOMBIA
ECUADOR
PERU
CHILE
PACIFIC OCEAN
Tropic of Cancer
Equator
Tropic of Capricorn
International Date Line
Scale 1 : 124,767,000 at Equator
1 inch equals approx. 2000 miles
Mercator Projection
1 AMERICAN SAMOA (U.S.)
2 COOK ISLANDS (N.Z.)
3 EASTER ISLAND (CHILE)
4 FIJI
5 FRENCH POLYNESIA (FR.)
6 GUAM (U.S.)
7 KIRIBATI
8 MARSHALL ISLANDS
9 MICRONESIA, FEDERATED STATES OF
10 NAURU
11 NEW CALEDONIA (FR.)
12 NORTHERN MARIANA ISLANDS (U.S.)
13 PALAU
14 PAPUA NEW GUINEA
15 SAMOA
16 SOLOMON ISLANDS
17 TONGA
18 TUVALU
19 VANUATU
20 WALLIS AND FUTUNA (FR.)


GUAM
PHILIPPINE SEA
PACIFIC OCEAN
Ritidian Point
Pati Point
Yigo
Dededo
Tumon Bay
Tamuning
Agana
Barrigada
Apra Harbor
Asan
Piti
Agat
Mangilao
Pago Bay
Umatac
Facpi Point
Talofofo Bay
Inarajan Bay
Ylig River
Fena Valley
Reservoir
GUAM TERRITORIAL SEASHORE
Mount Lamlam
Mount Bolanos
Mount Sasalaguan
COCOS I.
Scale 1 : 1,390,000
1 inch equals approx. 20 miles
Mercator Projection


SAVAI'I
SAMOA
PACIFIC OCEAN
AMERICAN SAMOA
Safotu
Cape Mulinu'u
Matavai
Mount Silisili
Sagone
Asau
Salelologa
Fagamalo
Lalomalava
Siufaga
Fagalii
Apolima
Manono
Samoa
Lotofaga
Fatausi'u
Lalomanu
Solosolo
Si'ufaga
Nu'usafee
Nu'utele
Nu'ulua
Tafatafa
Tula
Leone
Pago Pago
Tutuila I.
Aunu'u I.
Ofu I.
Olosega I.
Tau I.
Manua Islands
Scale 1 : 3,000,000
1 inch equals approx. 50 miles
Conformal Conic Projection


NEW CALEDONIA
PACIFIC OCEAN
CORAL SEA
LOYALTY ISLANDS
HUON I.
D'Entrecasteaux Reefs
Vils
EFATE
ERROMANGO
Unpongkor
Isangel
TANNA
POTT I.
Wala
BELEP IS.
ART I.
Astrolabe Reefs
BEAUTEMPS-BEAUPRÉ ATOLL
Koumac
Mount Panié
1629
OUVÉA
Fayaoué
Chépénéhé
Wé
LIFOU
Poindimié
Koné
Poya
Houaïlou
Bourail
Thio
Tadine
La Roche
MARÉ
IGA I.
Boulouparis
Mont Humboldt
1618
Dumbéa
Païta
Yaté
Nouméa
Mont-Dore
Gadji
ISLE OF PINES
Vao
WALPOLE I.
Durand Reefs
18°E/VANUATU
11°NEW CALEDONIA (FR.)
Scale 1 : 7,634,000
1 inch equals approx. 120 miles
Conformal Conic Projection
Tropic of Capricorn


VANUA LEVU
YASAWA GROUP
VITI LEVU
KORO SEA
LAU GROUP
FIJI
PACIFIC OCEAN
CIKOBIA
Labasa
RABI
Savusavu
TAVEUNI
OAMEA
Nabouwalu
EXPLORING ISLES
Rakiraki
KORO
Ba
Tavua
Vatukoula
Lautoka
Mount Tatawiri
1323
Levuka
MAGO
TUVUCA I.
MALOLO I.
OVALAU
NAIRAI
CICIA I.
Korovou
Nadi
Wunidawa
NAYAU I.
Namoli
Nausori
LAU
LAKEBA I.
Suva
SIGATOKA SAND DUNES NATIONAL PARK
Navua
BEQA
VATULELE
MOALA I.
TOTOYA I.
KABARA I.
OGEA LEVU
KADAVU I.
MATUKU I.
FULAGA I.
VATOA I.
ONO-I-LAU
Scale 1 : 7,194,000
1 inch equals approx. 114 miles
Conformal Conic Projection


SOLOMON ISLANDS
PACIFIC OCEAN
SOLOMON SEA
CORAL SEA
SANTA CRUZ ISLANDS
BOUGAINVILLE
CHOISEUL
SANTA ISABEL
MALAITA
MAKIRA
BELLONA
RENNELL I.
ONTONG JAVA
STEWART ISLANDS
DUFF ISLANDS
REEF ISLANDS
TINAKULA
NENDO
UTUPUA
VANIKOLO
ANUTA
FATAKA
TIKOPIA
Scale 1 : 10,000,000
1 inch equals approx. 160 miles
Conformal Conic Projection
EASTER ISLAND
RAPA NUI NATIONAL PARK
Hanga Roa
Scale 1 : 691,000
1 inch equals approx. 10 miles
Lambert Conformal Conic Projection
VANUATU
TORRES ISLANDS
UREPARAPARA I.
VANUA LAVA
SANTA MARIA
ESPÍRITU SANTO
MAÉWO
PENTECÔTE
MALAKULA
ÉPI
ÉFATÉ
ERROMANGO
TANNA
ANATOM I.
LIFOU
Scale 1 : 6,579,000
1 inch equals approx. 105 miles
Conformal Conic Projection
© 2002 Encyclopædia Britannica, Inc.
MOOREA
TAHITI
Scale 1 : 2,500,000
1 inch equals approx. 40 miles
Conformal Conic Projection
MARQUESAS ISLANDS
DISAPPOINTMENT ISLANDS
SOCIETY ISLANDS
TUAMOTU ARCHIPELAGO
AUSTRAL ISLANDS
GAMBIER ISLANDS
FRENCH POLYNESIA
Scale 1 : 20,202,000
1 inch equals approx. 320 miles
Conformal Conic Projection
Cities over 50,000
Cities 5,000 to 50,000
Cities under 5,000
National capitals
Territorial capitals
International boundaries
Water boundaries*
Intermittent rivers
Reefs
National parks
Spot elevations in metres (1 m = 3.28 ft)
*Shown as demarcations of convention or convenience, not necessarily corresponding to territorial status.

MAP INDEX

Political divisions
American Samoa (U.S.) . . . . . 14 20 s 170 00 w
Cook Islands (N.Z.) . . . . . 20 00 s 158 00 w
Easter Island (Chile) . . . . . 27 09 s 109 23 w
Fiji . . . . . 18 00 s 178 00 e
French Polynesia (Fr.) . . . . . 15 00 s 140 00 w
Guam (U.S.), *island* . . . . . 13 27 n 144 44 e
Kiribati . . . . . 5 00 s 170 00 w
Marshall Islands . . 10 00 n 167 00 e
Micronesia, Federated States of . . . . . 5 00 n 152 00 e
Nauru . . . . . 0 32 s 166 55 e
New Caledonia (Fr.) . . . . . 21 30 s 165 30 e
Northern Mariana Islands (U.S.) . . . 16 00 n 145 30 e
Palau . . . . . 7 30 n 134 30 e
Papua New Guinea . . . . . 6 00 s 147 00 e
Samoa . . . . . 13 35 s 172 20 w
Solomon Islands . . 8 00 s 159 00 e
Tonga . . . . . 20 00 s 175 00 w
Tuvalu . . . . . 8 00 s 178 00 e
Vanuatu . . . . . 16 00 s 167 00 e
Wallis and Futuna (Fr.) . . . . 13 18 s 176 10 w

Major islands (within political divisions not given on inset maps)
Babelthuap, *island* . . . . . 7 29 n 134 34 e
Banaba (Ocean Island) . . * 0 53 s 169 32 e
Caroline Islands . . 8 00 n 147 00 e
Christmas Island, see Kiritimati
Chuuk Islands . . . 7 25 n 151 43 e
Enewetak (Eniwetok), *island* . . . . . 11 25 n 162 20 e
Gilbert Islands . . . 0 00 174 00 e
Kiritimati (Christmas Island) . . . . . 1 52 n 157 25 w
Kosrae Island . . . . 5 18 n 162 59 e
Kwajalein, *island* . . . . . 8 43 n 167 43 e
Line Islands . . . . . 0 00 157 00 w
Ocean Island, see Banaba
Phoenix Islands . . 4 00 s 172 00 w
Pohnpei (Ponape), *island* . . . . . 6 54 n 158 13 e
Rarotonga, *island* . 21 14 s 159 46 w
Rota, *island* . . . . . 14 09 n 145 12 e
Saipan, *island* . . . 15 11 n 145 44 e
Tarawa, *atoll* . . . . . 1 25 n 173 00 e
Tinian, *island* . . . . 15 00 n 145 37 e
Yap, *island* . . . . . 9 31 n 138 06 e

AMERICAN SAMOA
Cities and towns
Nuuuli . . . . . 14 19 s 170 42 w
Pago Pago . . . . . 14 16 s 170 42 w

Physical features and points of interest
Aunu'a Island . . . . 14 17 s 170 33 w
Lata Mountain . . . . 14 14 s 169 29 w
Manua Islands . . . . 14 13 s 169 35 w
Matafao, Mount . . . 14 18 s 170 42 w
Ofu Island . . . . . 14 11 s 169 40 w
Olosega Island . . . . 14 10 s 169 37 w
Tau Island . . . . . 14 14 s 169 29 w
Tutuila Island . . . . 14 18 s 170 43 w

EASTER ISLAND
Cities and towns
Hanga Roa . . . . . 27 09 s 109 26 w

Physical features and points of interest
Anakena Bay . . . . 27 04 s 109 20 w
Benepú Roadstead . . . . . 27 10 s 109 25 w
Easter Island (Isla de Pascua) . . 27 07 s 109 22 w
Hotuiti Bay . . . . . 27 07 s 109 17 w
Katiki Volcano . . . . 27 06 s 109 16 w
La Perouse Bay . . . 27 04 s 109 18 w
Motu-Kao Kao Island . . . . . 27 12 s 109 28 w
Motu-Nui Island . . . 27 12 s 109 28 w
Pascua, Isla de, see Easter Island
Poike Peninsula . . . 27 07 s 109 15 w
Rano Aroi (Rana Roi) Volcano . . . . . 27 05 s 109 23 w
Rano Kau (Rana Kao) Volcano . . . . . 27 11 s 109 27 w
Rano Raraku (Rana Raraku) Volcano . . . . . 27 07 s 109 18 w
Rapa Nui National Park . . . . 27 03 s 109 23 w
Terevaka, Mount . . 27 05 s 109 23 w

FIJI
Cities and towns
Ba . . . . . 17 33 s 177 41 e
Korovou . . . . . 17 57 s 178 36 e
Labasa (Lambasa) . . . . 16 26 s 179 23 e
Lami . . . . . 18 07 s 178 25 e
Lautoka . . . . . 17 37 s 177 28 e
Levuka . . . . . 17 41 s 178 50 e
Nabouwalu . . . . . 16 59 s 178 42 e
Nadi . . . . . 17 48 s 177 26 e
Namosi . . . . . 18 02 s 178 06 e
Nausori . . . . . 18 02 s 178 32 e
Navua . . . . . 18 10 s 178 15 e
Rakiraki . . . . . 17 22 s 178 09 e
Savusavu . . . . . 16 45 s 179 20 e
Suva . . . . . 18 08 s 178 25 e
Tavua . . . . . 17 27 s 177 51 e
Vatukoula . . . . . 17 31 s 177 51 e
Vunidawa . . . . . 17 49 s 178 19 e

Physical features and points of interest
Ba (Mba), *river* . . . 17 29 s 177 40 e
Beqa (Mbengga), *island* . . . . . 18 23 s 178 08 e
Bligh Water, *sound* . . . . . 17 00 s 178 00 e
Bua (Mbua) Bay . . . 16 49 s 178 35 e
Cicia (Thithia), *island* . . . . . 17 45 s 179 18 w
Cikobia (Thikombia), *island* . . . . . 15 44 s 179 55 w
Dreketi (Ndreketi), *river* . . 16 33 s 178 52 e
Eastern Group, see Lau Group
Exploring Isles . . . 17 15 s 178 52 w
Fulaga (Fulanga) Island . . . . . 19 06 s 178 34 w
Gau (Ngau), *island* . . . . . 18 02 s 179 18 e
Kabara (Kambara) Island . . . . . 18 57 s 178 57 w
Kadavu (Kandavu) Island . . . . . 19 00 s 178 11 e
Koro, *island* . . . . . 17 19 s 179 23 e
Koro Sea . . . . . 18 00 s 179 50 e
Lakeba (Lakemba) Island . . . . . 18 13 s 178 47 w
Lau Group (Eastern Group), *islands* . . . . . 18 20 s 178 30 w
Mago (Mango) Island . . . . . 17 27 s 179 09 w
Mafolo Island . . . . 17 45 s 177 10 e
Matuku Island . . . . 19 10 s 179 46 e
Mba, see Ba
Mbengga, see Beqa
Mbúa, see Bua Bay
Moala Island . . . . 18 36 s 179 53 e
Nairai, *island* . . . . 17 49 s 179 24 e
Natewa Bay . . . . . 16 35 s 179 40 e
Navua, *river* . . . . 18 15 s 178 09 e
Nayau Island . . . . 17 58 s 179 03 w
Ndreketi, see Dreketi
Ngau, see Gau
Nggamea Island, see Qamea
Ogea Levu, *island* . . . . . 19 08 s 178 34 w
Ono-i-Lau, *islands* . . . . . 20 39 s 178 42 w
Ovalau, *island* . . . 17 40 s 178 48 e
Qamea (Nggamea), *island* . . . . . 16 46 s 179 46 w
Rabi (Rambi), *island* . . . . . 16 30 s 179 59 w
Rewa, *river* . . . . . 18 08 s 178 32 e
Sigatoka (Singatoka), *river* . . . . . 18 11 s 177 31 e
Sigatoka Sand Dunes National Park . . . 18 08 s 177 30 e
Taveuni, *island* . . 16 51 s 179 58 w
Thikombia Island, see Cikobia
Thithia Island, see Cicia
Tomanivi, Mount . . 17 37 s 178 01 e
Totoya Island . . . . 18 57 s 179 50 w
Tuvutha (Tuvuca) Island . . . . . 17 40 s 178 48 w
Vanua Levu, *island* . . . . . 16 33 s 179 15 e
Vatoa island . . . . 19 50 s 178 13 w
Vatulele, *island* . . 18 33 s 177 38 e
Viti Levu, *island* . . 18 00 s 178 00 e
Yasawa Group, *islands* . . . . . 17 00 s 177 23 e

FRENCH POLYNESIA
Cities and towns
Faaa . . . . . 17 33 s 149 36 w
Mahina . . . . . 17 30 s 149 27 w
Paea . . . . . 17 41 s 149 35 w
Papara . . . . . 17 31 s 149 49 w
Papao (Papara) . . . 17 45 s 149 33 w
Papeete . . . . . 17 32 s 149 34 w
Pirae . . . . . 17 31 s 149 32 w
Punaauia . . . . . 17 38 s 149 36 w
Uturoa . . . . . 16 44 s 151 26 w

Physical features and points of interest
Actéon Islands . . . 21 20 s 136 30 w
Ahe, *atoll* . . . . . 14 30 s 146 18 w
Ahunui, *atoll* . . . . 19 30 s 140 25 w
Akiaki, *island* . . . . 18 30 s 139 12 w
Amanu, *atoll* . . . . 17 48 s 140 46 w
Anaa, *atoll* . . . . . 17 25 s 145 30 w
Apataki, *atoll* . . . . 15 26 s 146 20 w
Arutua, *atoll* . . . . 15 18 s 146 44 w
Austral (Tubuai) Islands . . . . . 23 00 s 150 00 w
Bass (Marotiri) Islands . . . . . 27 55 s 143 26 w
Bellingshausen (One), *atoll* . . . . 15 48 s 154 30 w
Bora-Bora, *island* . 16 30 s 151 45 w
Disappointment (Désappointement) Islands . . . 14 10 s 141 20 w
Duke of Gloucester (Duc de Gloucester) Islands . . . . . 20 38 s 143 20 w
Elao, *island* . . . . . 8 00 s 140 40 w
Faaite, *atoll* . . . . 16 45 s 145 14 w
Fakahina, *atoll* . . . 15 59 s 140 08 w
Fakarava, *atoll* . . . 16 20 s 145 37 w
Fangataufa (Fagataufa), *atoll* . . . . . 22 14 s 138 45 w
Fatu Hiva, *island* . 10 28 s 138 38 w
Fenua Ura, see Scilly
Gambier (Mangareva) Islands . . . . . 23 09 s 134 58 w
Hao, *atoll* . . . . . 18 15 s 140 54 w
Haraiki, *atoll* . . . . 17 28 s 143 27 w
Hatutu, *island* . . . 7 56 s 140 32 w
Hatutu National Park . . . . . 7 56 s 140 32 w
Herehereitue, *atoll* . . . . . 19 54 s 144 58 w
Hikueru, *atoll* . . . . 17 36 s 142 37 w
Hiva Oa, *island* . . . 9 45 s 139 00 w
Huahine, *islands* . . 16 45 s 151 00 w
Kaukura, *atoll* . . . . 15 45 s 146 42 w
King George (Roi Georges) Islands . . . . . 14 32 s 145 08 w
Leeward (Sous le Vent) Islands . . . 16 38 s 151 30 w
Makatea, *island* . . . 15 50 s 148 15 w
Makemo, *atoll* . . . . 16 35 s 143 40 w
Mangareva, see Gambier Islands
Mangareva (Magareva), *island* . . . . . 23 07 s 134 57 w
Manihi, *atoll* . . . . 14 24 s 145 56 w
Manuhangi, *atoll* . . . . . 19 12 s 141 16 w
Maria, *atoll* . . . . . 21 48 s 154 41 w
Maria, *atoll* . . . . . 22 00 s 136 10 w
Marotiri, see Bass Islands
Marquesas (Marquises) Islands . . . . . 9 00 s 139 30 w
Marutea, *atoll* . . . . 17 00 s 143 10 w
Marutea, *atoll* . . . . 21 30 s 135 34 w
Mataiva, *atoll* . . . . 14 53 s 148 40 w
Mataivai Bay . . . . . 17 30 s 149 30 w
Maupiti, *island* . . . 16 27 s 152 15 w
Mehetia, *island* . . . 17 52 s 148 03 w
Moorea, *island* . . . . 17 32 s 149 50 w
Mopelia, *atoll* . . . . 16 50 s 153 55 w
Morane, *atoll* . . . . 23 10 s 137 07 w
Mururoa, *atoll* . . . . 21 52 s 138 55 w
Napuka, *island* . . . . 14 12 s 141 15 w
Nengonengo, *atoll* . . . . . 18 47 s 141 48 w
Niau, *island* . . . . . 16 09 s 146 21 w
Nuku Hiva, *island* . . . . . 8 54 s 140 06 w
Nukutavake, *island* . . . . . 19 17 s 138 50 w
One, see Bellingshausen
Orohena, Mount . . 17 37 s 149 28 w
Paraoa, *atoll* . . . . 19 09 s 140 43 w
Paumotu, see Tuamotu Archipelago
Pinaki, *atoll* . . . . 19 22 s 138 44 w
Pukapuka, *atoll* . . . 14 49 s 138 48 w
Pukarua, *atoll* . . . . 18 20 s 137 02 w
Raiatea, *island* . . . 16 50 s 151 25 w
Rakahanga, *island* . . 23 52 s 147 40 w
Rangiroa, *atoll* . . . 15 10 s 147 35 w
Rapa, *island* . . . . . 27 36 s 144 20 w
Raraka, *atoll* . . . . 16 10 s 144 54 w
Reao, *atoll* . . . . . 18 31 s 136 23 w
Reitoru, *atoll* . . . . 17 52 s 143 05 w
Rimatara, *island* . . 22 38 s 152 51 w
Roi Georges, see King George Islands
Roniu, Mount . . . . 17 49 s 149 12 w
Rurutu, *island* . . . . 22 26 s 151 20 w
Scilly (Fenua Ura), *atoll* . . . . 16 30 s 154 40 w
Society Islands (Îles de la Société) . . . . . 17 00 s 150 00 w
Sous le Vent, see Leeward Islands
Tahaa, see Vaitoare
Tahanea, *atoll* . . . . 16 52 s 144 45 w
Tahiti, *island* . . . . 17 37 s 149 27 w
Tahuata, *island* . . . 9 57 s 139 05 w
Taiarapu Peninsula . . . . . 17 47 s 149 14 w
Takapoto, *atoll* . . . 14 38 s 145 12 w
Takaroa, *atoll* . . . . 14 28 s 144 58 w
Takume, *atoll* . . . . 15 49 s 142 12 w
Taravao Bay . . . . . 17 43 s 149 17 w
Tatakoto, *atoll* . . . 17 20 s 138 23 w
Tauere, *atoll* . . . . 17 22 s 141 30 w
Tematangi, *atoll* . . 21 41 s 140 40 w
Temoe, *island* . . . . 23 20 s 134 29 w
Tepoto, *island* . . . . 14 06 s 141 24 w
Tetiaroa, *atoll* . . . . 17 05 s 149 32 w
Tikehau, *atoll* . . . . 15 00 s 148 10 w

Tohivea, Mount . . . 17 33 s 149 49 w
Tuamotu (Paumotu) Archipelago . . . . 19 00 s 142 00 w
Tubuai, see Tupuai
Tubuai Islands, see Austral Islands
Tupuai, *island* . . . 23 23 s 149 28 w
Tureia, *atoll* . . . . 20 50 s 138 32 w
Ua Huka, *island* . . 8 54 s 139 33 w
Ua Pu, *island* . . . . 9 23 s 140 03 w
Vahitahi, *atoll* . . . 18 44 s 138 52 w
Vaitoare (Tahaa), *island* . . . . . 16 41 s 151 27 w
Vanavana, *atoll* . . 20 47 s 139 09 w
Windward (Vent) Islands . . . . . 17 30 s 149 30 w

GUAM

Cities and towns

Agana . . . . . 13 28 n 144 45 e
Agana Heights . . . 13 27 n 144 44 e
Agat . . . . . 13 23 n 144 39 e
Barrigada . . . . . 13 28 n 144 48 e
Dededo . . . . . 13 31 n 144 50 e
Mangilao . . . . . 13 27 n 144 49 e
Tamuning . . . . . 13 29 n 144 47 e
Toto . . . . . 13 28 n 144 47 e
Yigo . . . . . 13 32 n 144 53 e

Physical features and points of interest

Apra Harbor . . . . 13 27 n 144 39 e
Bolanos, Mount . . 13 18 n 144 41 e
Cocos Island . . . . 13 14 n 144 39 e
Fena Valley Reservoir . . . . . 13 21 n 144 42 e
Guam Territorial Seashore . . . . . 13 18 n 144 40 e
Lamlam, Mount . . 13 20 n 144 40 e
Pago Bay . . . . . 13 25 n 144 48 e
Pati Point, *cape* . . 13 36 n 144 57 e
Philippine Sea . . . 13 35 n 144 50 e
Ritidian Point, *cape* . . . . . 13 39 n 144 52 e
Sasalaguan, Mount . . . . . 13 17 n 144 41 e
Tumon Bay . . . . . 13 31 n 144 48 e
Ugum, *river* . . . . . 13 20 n 144 45 e

NEW CALEDONIA

Cities and towns

Bouloupari . . . . . 21 52 s 166 03 e
Bourail . . . . . 21 34 s 165 29 e
Chépénéhé . . . . . 20 47 s 167 09 e
Dumbéa . . . . . 22 09 s 166 27 e
Fayaoué . . . . . 20 39 s 166 32 e
Gadji . . . . . 22 33 s 167 26 e
Houaïlou . . . . . 21 17 s 165 37 e
Koné . . . . . 21 04 s 164 52 e
Koumac . . . . . 20 34 s 164 17 e
La Roche . . . . . 21 28 s 168 02 e
Mont-Dore . . . . . 22 17 s 166 35 e
Nouméa (Numea) . . . . . 22 16 s 166 27 e
Paita . . . . . 22 08 s 166 22 e
Poindimié . . . . . 20 56 s 165 20 e
Poya . . . . . 21 21 s 165 09 e
Tadine . . . . . 21 33 s 167 53 e
Thio . . . . . 21 37 s 166 13 e
Vao . . . . . 22 40 s 167 29 e
Wala . . . . . 19 43 s 163 38 e
Wé . . . . . 20 54 s 167 16 e
Yaté . . . . . 22 09 s 166 54 e

Physical features and points of interest

Art Island . . . . . 19 43 s 163 39 e
Astrolabe Reefs . . 19 49 s 165 35 e
Beautemps-Beaupré Atoll . . . 20 20 s 166 10 e
Bélep Islands . . . 19 45 s 163 40 e
Coral Sea . . . . . 21 30 s 166 00 e
D'Entrecasteaux Reefs . . . . . 18 20 s 163 00 e
Diahot, *river* . . . . 20 17 s 164 18 e
Durand Reef . . . . 22 02 s 168 39 e
Humboldt, Mount . . . . . 21 53 s 166 25 e
Huon Island . . . . 18 01 s 162 57 e
Kunié, see Pines, Isle of
Lifou, *island* . . . . 20 53 s 167 13 e
Loyalty (Loyauté) Islands . . . . . 21 00 s 167 00 e
Maré, *island* . . . . 21 30 s 168 00 e
New Caledonia (Nouvelle-Calédonie), *island* . . . . . 21 30 s 165 30 e
Ouvéa (Uea, or Uvea), *island* . . . 20 35 s 166 35 e
Panié, Mount . . . . 20 36 s 164 46 e
Pines (Kunié), Isle of . . . . . 22 37 s 167 30 e
Pott Island . . . . . 19 35 s 163 36 e
Tiga Island . . . . . 21 08 s 167 49 e
Uea, see Ouvéa
Uvea, see Ouvéa
Walpole Island . . . 22 37 s 168 57 e

SAMOA

Cities and towns

Apia . . . . . 13 50 s 171 44 w
Asau, see Matavai
Fa'aala . . . . . 13 45 s 172 16 w
Faleasi'u . . . . . 13 48 s 171 54 w
Le'auva'a . . . . . 13 48 s 171 51 w
Lotofaga . . . . . 13 59 s 171 50 w
Matavai (Asau) . . . 13 28 s 172 35 w
Safotu . . . . . 13 27 s 172 24 w
Sagone . . . . . 13 39 s 172 35 w
Samatau . . . . . 13 54 s 172 02 w
Sili . . . . . 13 43 s 172 21 w
Si'umu . . . . . 14 01 s 171 47 w
Solosolo . . . . . 13 51 s 171 36 w

Physical features and points of interest

Apolima, *island* . . 13 49 s 172 07 w
Fanuatapu, *island* . . . . . 13 59 s 171 20 w
Manono, *island* . . 13 50 s 172 05 w
Mulinu'u, Cape . . . 13 31 s 172 49 w
Namu'a, *island* . . 14 00 s 171 22 w
Nu'ulua, *island* . . . 14 03 s 171 22 w
Nu'usafee, *island* . . . . . 14 02 s 171 41 w
Nu'utele, *island* . . 14 02 s 171 22 w
O Le Pupu-Pue National Park . . . 14 05 s 171 50 w
Savai'i, *island* . . . 13 35 s 172 25 w
Silisili, Mount . . . . 13 35 s 172 27 w
Upolu, *island* . . . . 13 55 s 171 45 w

SOLOMON ISLANDS

Cities and towns

Aimela . . . . . 8 19 s 160 42 e
Aola . . . . . 9 32 s 160 29 e
Auki . . . . . 8 46 s 160 42 e
Buala . . . . . 8 08 s 159 35 e
Fausabu . . . . . 8 34 s 160 43 e
Gizo . . . . . 8 06 s 156 51 e
Honiara . . . . . 9 26 s 159 57 e
Inakona . . . . . 9 49 s 160 02 e
Kirakira . . . . . 10 27 s 161 55 e
Kmagha . . . . . 8 13 s 159 42 e
Kolombaghea . . . . 8 01 s 157 36 e
Kopiu . . . . . 9 53 s 160 45 e
Kukundu . . . . . 8 01 s 156 57 e
Lata . . . . . 10 44 s 165 54 e
Malango . . . . . 9 41 s 159 43 e
Maravovo . . . . . 9 17 s 159 38 e
Munda . . . . . 8 19 s 157 15 e
Nolavlu . . . . . 9 44 s 160 31 e
Sasamungga . . . . 7 02 s 156 47 e
Sigana . . . . . 8 30 s 159 52 e
Takwa . . . . . 8 22 s 160 48 e
Tulagi (Tulaghi) . . 9 06 s 160 09 e
Visale . . . . . 9 15 s 159 42 e

Physical features and points of interest

Anuta, *island* . . . . 11 38 s 169 50 e
Aola Bay . . . . . 9 31 s 160 29 e
Bellona, *island* . . . 11 17 s 159 47 e
Bougainville Strait . . . . . 6 40 s 156 10 e
Bradley Reefs . . . 6 52 s 160 48 e
Choiseul, *island* . . . 7 05 s 157 00 e
Coral Sea . . . . . 12 00 s 164 00 e
Duff Islands . . . . . 9 50 s 167 10 e
Fataka, *island* . . . 11 55 s 170 12 e
Florida Islands . . . 9 00 s 160 10 e
Guadalcanal Island . . . . . 9 32 s 160 12 e
Indispensable Reefs . . . . . 12 40 s 160 25 e
Kolombangara Island . . . . . 8 00 s 157 05 e
Makerakomburu, Mount . . . . . 9 43 s 160 02 e
Makira (San Cristobal), *island* . . . . . 10 36 s 161 45 e
Malaita (Mala Mara), *island* . . . 9 00 s 161 00 e
Maramasike (Small Malaita), *island* . . . . . 9 32 s 161 27 e
Mu Nggava, see Renneli Island
Nendo, *island* . . . 10 45 s 165 54 e
New Georgia, *island* . . . . . 8 15 s 157 30 e
New Georgia Sound . . . . . 8 00 s 158 10 e
Nggatokae Island . . . . . 8 46 s 158 11 e
Ontong Java, *atoll* . . . . . 5 20 s 159 30 e
Rannongga Island . . . . . 8 05 s 156 34 e
Reef Islands . . . . 10 15 s 166 10 e
Rendova Island . . . 8 32 s 157 20 e
Rennell (Mu Nggava) Island . . . . . 11 40 s 160 10 e
Roncador Reef, *atoll* . . . . . 6 13 s 159 22 e
Russell Islands . . . 9 04 s 159 12 e
San Cristobal Island, see Makira
Santa Cruz Islands . . . . . 11 00 s 166 15 e
Santa Isabel (Ysabel), *island* . . . . . 8 00 s 159 00 e
Savo, *island* . . . . 9 08 s 159 49 e
Shortland Islands . . . . . 6 55 s 155 53 e
Sikaiana Atoll, see Stewart Islands
Simbo Island . . . . 8 17 s 156 33 e
Small Malaita, see Maramasike
Solomon Sea . . . . 8 00 s 155 00 e
Stewart Islands (Sikaiana Atoll) . . 8 20 s 162 40 e
Tikopia, *island* . . . 12 19 s 168 49 e
Tinakula, *island* . . 10 24 s 165 45 e
Ulawa Island . . . . 9 46 s 161 57 e
Utupua, *island* . . . 11 16 s 168 29 e
Vanguru Island . . . 8 38 s 158 00 e
Vanikolo, *island* . . 11 39 s 166 54 e
Vella Lavella Island . . . . . 7 45 s 156 40 e
Ysabel, see Santa Isabel

VANUATU

Cities and towns

Hog Harbour . . . . 15 09 s 167 06 e
Ipayato . . . . . 15 38 s 166 52 e
Isangel . . . . . 19 33 s 169 16 e
Lakatoro . . . . . 16 07 s 167 25 e
Lalinda . . . . . 16 21 s 168 03 e
Laol . . . . . 16 41 s 168 16 e
Lolteng . . . . . 15 33 s 168 09 e
Luganville . . . . . 15 32 s 167 10 e
Lumbukuti . . . . . 16 55 s 168 32 e
Natapao . . . . . 17 37 s 168 13 e
Norsup . . . . . 16 04 s 167 23 e
Port Olry . . . . . 15 03 s 167 04 e
Port-Vila, see Vila
Toak . . . . . 16 21 s 168 17 e
Unpongkor . . . . . 18 49 s 169 01 e
Veuturnboso . . . . 13 54 s 167 27 e
Vila (Port-Vila) . . . 17 44 s 168 18 e

Physical features and points of interest

Ambae, see Aoba
Ambrym (Ambrim), *island* . . . . . 16 15 s 168 07 e
Anatom (Aneityoum) Island . . . . . 20 12 s 169 49 e
Aniwa Island . . . . 19 15 s 169 36 e
Aoba (Ambae, or Oba), *island* . . . . 15 23 s 167 50 e
Aragh, see Pentecôte
Aurora, see Maéwo Island
Banks Islands . . . . 14 00 s 167 30 e
Big Bay . . . . . 15 03 s 166 54 e
Éfaté (Sandwich, or Vaté), *island* . . . . . 17 40 s 168 25 e
Émaé Island . . . . . 17 04 s 168 22 e
Épi (Tasiko), *island* . . . . . 16 45 s 168 15 e
Erromango (Eromango), *island* . . . . . 18 48 s 169 05 e
Espiritu Santo (Marina, or Santo), *island* . . . 15 20 s 166 53 e
Futuna (Foutouna) Island . . . . . 19 32 s 170 12 e
Garet, Mount, *volcano* . . . . . 14 17 s 167 30 e
Gaua, see Santa Maria
Maéwo (Aurora) Island . . . . . 15 09 s 168 07 e
Malakula (Malékoula), *island* . . . . . 16 15 s 167 30 e
Malo (St. Bartholomew) Island . . . . . 15 41 s 167 10 e
Marina, see Espiritu Santo
Maroum, Mount, *crater* . . . . . 16 15 s 168 07 e
Oba, see Aoba
Parapara Island, see Ureparapara
Pentecôte (Aragh, or Pentecost, or Raga), *island* . . . 15 45 s 168 10 e
St. Bartholomew, see Malo Island
Sandwich Island, see Éfaté
Santa Maria (Gaua), *island* . . . . . 14 17 s 167 30 e
Santo, see Espiritu Santo
Shepherd Islands . . . . . 16 55 s 168 35 e
Sûrétiméat, Mount . . . . . 13 47 s 167 28 e
Tabwémasana, Mount . . . . . 15 22 s 166 45 e
Tanna (Tana), *island* . . . . . 19 30 s 169 20 e
Tasiko Island, see Épi
Torres Islands . . . 13 15 s 166 37 e
Ureparapara (Parapara), *island* . . . . . 13 32 s 167 20 e
Vanua Lava (Vanoua Lava), *island* . . . . . 13 50 s 167 28 e
Vaté, see Éfaté

part of an atoll if connected by a reef ringing a lagoon. The "high–low" distinction is misleading as the two types occur in many combinations, and some coral islands have been elevated considerably by changes in the ocean level.

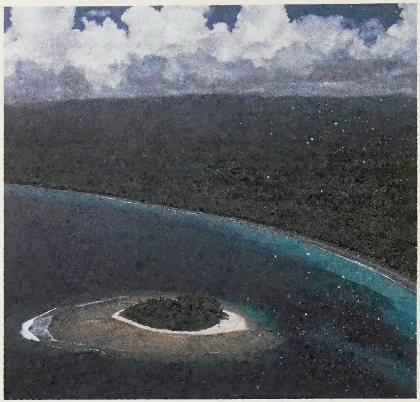
Continental islands. The islands of the broad western Pacific margin, formed mainly of metamorphosed rocks, sediments, and andesitic volcanic material, are separated from the basaltic volcanic islands of the central and eastern Pacific by deep ocean trenches along the eastern borders of Japan, the Marianas, New Guinea, Solomon Islands, Fiji, and New Zealand, a demarcation that is commonly called the Andesite Line. These continental islands, faulted and folded in mountainous arcs, tend to be higher and larger than those farther east and have rich soils that support almost every kind of vegetation. New Guinea, 1,500 miles long and with a maximum width of nearly 500 miles, is a good example. Its snowcapped mountains rise to about 16,400 feet (5,000 metres), its interior is dissected by high plateaus and extensive river systems, and its slopes and coastal margins contain dense forests and vast swamps.

High oceanic islands. Extensive volcanic mountain ranges in the central and eastern Pacific rise abruptly from the ocean deep, their cores of dense black basalt built up from lava flows from the fractured seafloor. Where summits stand in high relief above sea level, they represent most of the islands that constitute Polynesia and Micronesia. Small to intermediate in land area, nowhere do they match the extent of continental islands. Hawaii's snow-topped Mauna Kea reaches 13,796 feet, though most oceanic islands have peaks of somewhat less than 5,000 feet. Topography is extremely rugged, with sharp ridges, deep canyons, high cliffs, and waterfalls abounding. Human communities occupy the more congenial lower slopes, floodplains, and wide strands. The islands, rich in iron and magnesium oxides, are densely forested but lack the mineral wealth of continental islands.

Low coral islands. Most Pacific islands are coral formations, although all rest on volcanic or other cores. In the shallow waters of the tropics, both continental and oceanic islands attract coral growth in the form of fringing reefs, partially submerged platforms of consolidated limestone, with coral organisms at the ocean edge feeding on materials carried in by waves and currents. Coral-building polyps and algae secrete calcium carbonate from seawater, forming skeletal frameworks that adhere to land surfaces or to the rock remains of coralline ancestors. Many islands have been gradually submerged through a combination of sinking, caused by geologic action, and flooding, caused by the melting of ice caps. As islands were flooded, coral growth continued outward, producing barrier reefs farther from shore and separated from it by a lagoon.

Coral atolls

A coral atoll results when still further flooding reduces an island to a submarine condition. The usually irregular reef continues to build up in the warm shallows. It encircles a clear-surfaced lagoon of moderate depth and in time supports a number of islets, known locally as motu, built up from reef debris to 20 or 30 feet above sea level. Atolls exist in all shapes and sizes. Kwajalein Atoll in the Marshalls is the world's largest, being about 80 miles long and 20 miles wide, with a lagoon area of 655 square miles. Openings, or passages, which commonly occur on the leeward side of Pacific atolls, permit access to the lagoon by ocean shipping. The only source of fresh water is rain.



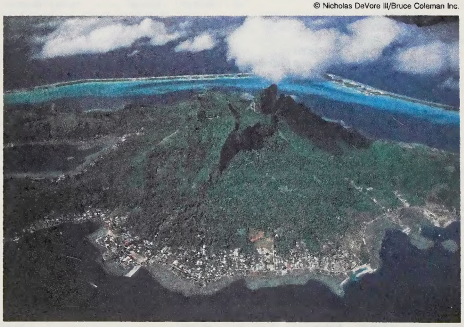
Small offshore coral atoll at the eastern end of Guadalcanal Island, Solomon Islands.

© Michael McCoy/Australasian Nature Transparencies

Successive elevations of an island above sea level by geologic action have created a variety of "raised" coral formations. The northern half of Guam, for example, is a coralline limestone plateau rising to 850 feet, while the mountains in the southern half of the island, formed by volcanic activity, reach elevations up to 1,300 feet. Nauru and Banaba (Ocean Island) are raised coral islands that stand at elevations of about 210 and 265 feet, respectively. They have deeper soil and a more adequate water supply than atoll islets, as well as surface deposits of phosphate rock (derived from guano) that have been mined commercially.

Soils. Pacific island soils develop through the action of

© Nicholas DeVore III/Bruce Coleman Inc.



Volcanic peaks of Bora-Bora, a high volcanic island surrounded by a lagoon, Society Islands, French Polynesia.

temperature, rainfall, and organic matter on the original rock materials. This process is further influenced by factors of time and land relief. Coral island soils are relatively infertile and deficient in organic materials. The mineral-bearing soils of the continental islands are more complex and richer than those of the volcanic-based high islands. The most productive soils on high islands occur in the lower valley slopes, alluvial floodplains, and deltas and in some instances are further enriched by volcanic ash deposits. Tropical temperatures and heavy rainfall have leached nutrients from the soils in many areas, producing lateritic soils, which become relatively infertile after two or three years of cultivation. Fertilizers must then be added, or else the land must be abandoned to allow it to recover by natural processes while other land is cleared and planted in its place.

Climate. The climate is tropical and heavily influenced by the sea. Humidity and temperature tend to be high and are generally uniform throughout the year. The region's varied precipitation patterns are affected by elevation and larger systems of air circulation.

Across the eastern and central Pacific, air currents trend westward as they flow toward the Equator, forming the northeast and southeast trade winds. These brisk winds bring light to moderate showers interspersed with brief but often heavy downpours or clear skies. The windward sides of high islands are cloudier and wetter than the drier leeward coasts. Seasonal shifts in wind direction frequently presage stormy weather. Where the trade winds meet near the Equator lie the doldrums, a region that often experiences little or no wind and has cloudy skies and high humidity.

Monsoon winds

The trade winds merge or give way to the monsoon winds in the far western Pacific, where the alternate cooling and heating of continental Asia produces a seasonal reversal of winds. From about November to March, the northwest monsoon from Asia brings heavy downpours to the northerly slopes of the western Carolines, New Guinea, and the Solomons. In summer the southeast monsoon reverses the process.

Typhoons occur frequently in western Micronesia from July to November and are active south of the Equator from Australia to the Society Islands four to six months later. These cyclonic storms are characterized by gale-force winds, torrential rains, and high waves, all of which can cause extensive damage to crops and buildings, especially on low-lying coral islands. Atolls in the equatorial region, however, have experienced droughts lasting as long as two or three years. The meteorologic and oceanographic phenomenon known as El Niño periodically causes major shifts in the region's weather patterns and can seriously disrupt life on the islands. The El Niño events of 1982–83 and 1997–98, for example, were associated with harmful droughts in the western Pacific and frequent cyclonic storms in the eastern Pacific. La Niña events are contrasting phenomena that tend to reduce precipitation in the eastern Pacific and increase the risk of cyclones in the western Pacific.

Plant and animal life. Each of the myriad ecological systems in the Pacific is a unique complex of living organisms and their environment. Because of the limited size of most Pacific islands, almost any change, whether by human action or by some natural agency, will probably have repercussions elsewhere in the ecosystem.

Most Pacific island flora is probably related to species from Indonesia and New Guinea. The variety of plant genera in the western Pacific is far greater than that in the eastern Pacific, providing evidence that seeds and fruits have been carried progressively eastward by ocean currents, birds, winds, and island voyagers. Plants have adapted to local differences in moisture, soil, salinity, and temperature, resulting in several endemic species. However, many plants have been introduced to the region from other parts of the world since the late 19th century.

Seacoast, or strand, vegetation is the most widespread type of vegetation in the Pacific. It includes shrubs, herbs, woody vines, and trees that have a high tolerance for the salt spray that is borne by ocean winds. Mangrove thickets proliferate in brackish swamps.

Numerous species are found on the high islands, where primary forests still survive in river valleys, on intermediate slopes, and on lowland plains. Rainforests are also common and consist of myriad huge trees overlooking smaller trees and shade-tolerant ferns, vines, and shrubs. Secondary forests and grasslands have grown up where virgin forests were destroyed by fire and by shifting cultivation. Grasslands are also associated with areas of poor soil and limited precipitation, as on leeward slopes.

Bats, rats, and, in New Guinea, wallabies, flying squirrels, and spiny anteaters were the only mammals to precede humans into the Pacific, after which pigs, dogs, poultry, goats, deer, and cattle were introduced. Most islands have some snakes and lizards, but crocodiles are restricted to the west. Seabirds, such as terns, frigates, albatross, petrels, and boobies, as well as migratory ducks, plovers, and curlews, are found nearly everywhere. Though oceanic islands support only a few land birds, New Guinea has unusual species such as the cockatoo, hornbill, bird of paradise, and cassowary. Insects, the most numerous of island fauna, include centipedes, cockroaches, lice, houseflies, and malaria-carrying *Anopheles* mosquitoes, which exist in Melanesia from coastal New Guinea to Vanuatu. Scorpions are also found.

Unusual bird species

Marine life is abundant in the Pacific Islands. Reefs and lagoons provide shelter for lobsters, shrimps, clams, oysters, snails, eels, octopuses, turtles, and numerous species of fish. Skipjack and yellowfin tuna, swordfish, marlin, and many other large fishes are found in deeper waters beyond the reefs. Sharks are the dominant predators. Groups of humpback and other whales and porpoises are also seen frequently at sea.

THE PEOPLE

Humans in the Pacific have had to adapt to island environments just as other species have done. The technologies and organizational systems introduced into the Pacific by migrants were established in habitats that varied from receptive to hostile. The earliest settlers were food-gathering peoples who probably did not greatly disrupt their surroundings. Their successors, practiced in horticulture and skilled in sea transport, were more able to fashion and control the local landscape according to their own customs. Later, Western practices and advanced technology led to increasingly larger settlements, which have threatened the inherently vulnerable island ecosystems.

Aboriginal groupings. Pacific islanders tend to identify themselves by their home island or their mother tongue, saying, for example, that they are from Nauru or that they speak Fijian. Occasionally, they may invoke another, and larger, identity, claiming to be Polynesian, Micronesian, or Melanesian. However, as marks of ethnic, linguistic, or cultural affiliation these latter designations may be misleading. Melanesians are sometimes thought to resemble indigenous Australians in appearance, and Micronesians are said to have more Asian physical characteristics; Polynesians are thought to exhibit characteristics of both. A great deal of intermarriage has taken place throughout the Pacific since the first immigrants arrived in the southwestern islands. The linguistic pattern is also complex. Some generalizations about cultural practices in the three regions may nevertheless be made, although many characteristics overlap among the traditionally ascribed areas, and there are exceptions to every statement.

Generalizations about the traditional regions

Polynesia. Polynesians are the most homogeneous in speech, custom, and physical appearance, although western Polynesians (Samoans and Tongans) are moderately distinct from the rest. They have long been accomplished farmers and fishers and have directed much energy to nonmaterial pursuits. Epic mythology, copious genealogies, sophisticated social etiquettes, hereditary aristocracies, and elaborated religious formality, with varying degrees of emphasis, characterized society in pre-European Polynesia. A kinship system developed that recognized the worth of both maternal and paternal family ties. Secular leaders, regarded as lineal descendants of deified ancestors, served gods and humans alike. Artistic creativity was encouraged and rewarded in wood and stone sculpture, featherwork, tapa (bark cloth), and tattooing, and artists

were accorded privileges commensurate with those granted warriors, navigators, herbalists, and seers.

Micronesia. Eight or 10 cultural-linguistic areas in Micronesia attest to its greater heterogeneity. Palau, Yap, and the Mariana Islands, in western Micronesia, suggest affinities with Melanesia, Indonesia, and the Philippines. Island farmsteads and hamlets are customarily located near the

Wolfgang Kaehler/Corbis



Islanders netting fish near Ifalik, Caroline Islands, Micronesia.

shore, reflecting interests in fishing, canoeing, and interisland trade. Except in Kiribati (the Gilbert, Phoenix, and Line Islands), matrilineal clans and lineages influence property inheritance, succession to traditional titles, and intracommunity competition. Among the more common traditional art forms are mat work, shell ornaments, loom weaving, tattooing, and functionally crafted wood and shell artifacts.

Melanesia. Melanesia is a region of unending contrast, where "beach" populations benefit from coastal trading and cultural exchanges, whereas more traditional and isolated "bush" populations exist in the interiors of the larger islands. Polynesian influence touches Fiji and a few outlying islands to the northwest. Numerous distinct indigenous groups (largely ethnic Papuans) are spread throughout the interior of New Guinea.

Melanesians depend heavily on agriculture, including pig raising. Communities are usually organized around patrilineal groups. In most Melanesian societies, leadership depends on the local "Big Man," who, aided by his many relatives, gains support within his own village and enhances his influence in others nearby by hosting more and grander feasts than his rivals and amassing wealth through ceremonial exchanges of valuable goods. Opportunities for upward mobility are better for the sons of an established "Big Man," assuming they can prove themselves. Although males dominate most cultural activities, women are influential in certain religious and exchange systems. Art forms associated with these activities include dance masks, sculptured figures, body scarification, and carved mortuary standards.

For a detailed treatment of these cultural areas, see below *Traditional cultures of the Pacific Islands.*

Fluctuation of island societies

Interaction with Western societies. During five centuries of contact with Europeans, followed by Americans and Asians, island societies have fluctuated between change and disorganization on the one hand and stability and reintegration on the other. Pacific islanders, reacting to the novelty and authority of Western civilization, disrupted their relatively balanced ecosystems to serve new economies and, under pressure from religious missions and alien governments, to adopt new practices and beliefs. However, reasonably stable island societies emerged from that initial confusion and attempted to preserve their traditional ethos while fusing elements from both cultures. In many areas, European, American, and Asian residents have continued to maintain their own cultural identities in social enclaves, although there has also been considerable intermarriage among the various ethnic groups.

World War II (1939–45) stimulated many island communities to seek greater participation in Western society. Some joined mystical religious movements called cargo cults, principally in Melanesia, which blended traditional and Christian elements in systems aimed at material gain. Other islanders, including native peoples, attempted to advance themselves economically and politically by promoting nationalism. Most islanders were opposed to assimilation with Western societies and argued for more independent status within the structure of a Westernized "Pacific way."

Introduced diseases contributed to the decline of the population until about 1900, when more-advanced medical treatments and health education became established in the region. Numerous island populations increased nearly threefold during the next five decades. Many islanders subsequently migrated from overcrowded outlying islands to port towns and island capitals, often to seek jobs or further education.

Emigration to ports overseas

The populations of some island states are largely urbanized. However, many urban residents contend with poor housing and sanitation, underemployment, a scarcity of land for residential use, and a decline of the extended family structure. Alcoholism and crime are widespread. Meanwhile, communities in the outer islands are less able to function because of the exodus of skilled adults to towns. The urban centres link the hinterland communities by sea and air transport and by telecommunications networks; they also serve as a point of departure for such overseas destinations as Guam and Hawaii, where employment opportunities in the private sector are greater.

THE ECONOMY

Agriculture and natural resources. Agricultural production consists of both family enterprises and commercial plantation systems, which predominate on the larger islands. Coconut palms and textile screw pines (pandanus palms) flourish on coral islands. Breadfruit and banana trees and papaya plants may also be grown farther inland, as may marsh taro (*Cyrtosperma*), a root crop cultivated in pits dug to groundwater level and enriched with plant debris. Yams, taro (*Colocasia*), cassavas (manioc), sweet potatoes, breadfruit, bananas, coconuts, and papayas are staples in high-island economies. Perishable fruits, such as pineapples, bananas, and citrus fruits, require markets close at hand unless they are

Bob and Ira Spring



Open-pit nickel mine on mainland New Caledonia.

locally processed or can be assured of prompt delivery to overseas destinations. Sugar, exported mainly from Hawaii and Fiji, requires careful management, costly machinery, and specialized labour. Experiments in growing coffee, cacao, spices, and other cash crops have been undertaken to stimulate diversification and to minimize the hazards of a one-crop economy. Timber and wood by-products are processed commercially on some islands.

Marine resources, although almost unlimited, require skilled labour and capital facilities for commercial exploitation. Deep-sea fishing, mostly for tuna, is conducted throughout the Pacific, mainly by Korean, Taiwanese, Japanese, and American fishing fleets. Canneries employing islanders are operated in American Samoa, Fiji, and Solomon Islands. Local cooperatives have been successful in marketing fresh fish.

Fiji, Papua New Guinea, Irian Jaya (a province of Indonesia comprising the western half of the island of New Guinea and its offshore islands), and Solomon Islands have profited from either gold or gold and copper discoveries, and New Caledonia is rich in nickel ores. Oil reserves and their exploitation are restricted to Irian Jaya, although prospecting in Papua New Guinea and Solomon Islands is encouraging. Natural phosphates on Nauru continue to be mined, but those on Banaba were depleted in 1979. In 1979 the South Pacific Forum, an organization of independent and self-governing countries, established the Forum Fisheries Agency (FFA) to facilitate mutual cooperation and assistance in fisheries and policing the 200-nautical-mile Exclusive Economic Zone (EEZ) of each member state. This group of 16 entities in 1987 concluded a multilateral treaty with the United States, whereby the United States agreed to pay $60,000,000 in license fees and assistance to the fishing industry in return for permission for American fishing fleets to operate in the EEZs of the FFA member states. The latter are free to negotiate further agreements with other nations on the Pacific Rim.

Economic handicaps

Trade. Pacific islanders, as producers of agricultural, marine, and mineral commodities, face problems of market demand, labour supply, management skills, and transport that restrict them to an insignificant role in world export trade. Neither do the small, scattered populations present an attractive consumer market to overseas entrepreneurs. The combination of limited exportable products, heavy dependency on food imports, high cost of fuel imports, and overreliance on foreign aid makes each island state's economy extremely vulnerable.

More than half of the exports from the Pacific Islands are sent to Japan, the United States, Australia, New Zealand, and members of the European Economic Community (EEC). Imports are received mainly from the United States, Australia, New Zealand, France, Japan, and Singapore. Almost all island groups import far more, in dollar amounts, than they export. External financial aid for economic development is received primarily from Australia, New Zealand, the United States (for Guam, American Samoa, and the U.S. interests in Micronesia), France (for New Caledonia, Wallis and Futuna, and French Polynesia), and from international organizations (the Asian Development Bank, the United Nations Development Programme, and the EEC). Most grants-in-aid are made on a bilateral basis; few are negotiated on a multilateral or regional basis. Most employment opportunities for islanders are in government service agencies, except where mining or agricultural production contributes significantly to national income.

In the 1970s tourism opened up new employment opportunities and sources of revenue. By 1980 Guam, the Northern Marianas, Fiji, French Polynesia, and New Caledonia, among others, had followed Hawaii's earlier lead in attracting visitors by air and sea to enjoy the tropical scenery, handicrafts, and friendly hospitality of the Pacific Islands. Australians made up the largest group of tourists traveling to destinations south of the Equator, while Japanese made up the majority of vacationers to the north. However, the increasing ability of international airlines to bypass midway island stops, the devastation of hotel facilities by more frequent hurricanes, the fluctuating currency values in Pacific Rim countries, and political disorders in Fiji and New Caledonia slowed the growth of the tourist industry. In some cases island labour was mostly restricted to the menial services, and management posts and industry profits were more often enjoyed by expatriate personnel and foreign-based corporations. The degree of government and private involvement and investment in tourist accommodations and entertainment varied with each island state. Some island populations, indeed, were more concerned with safeguarding their traditional cultures than deriving income from their commercialization.

Response to commercialization

Transportation and communication. Most islanders, by resorting to some combination of road, canoe, motorboat, interisland freighter, local air service, or international airline, can travel within a week or two to such distant cities as Sydney, Auckland, San Francisco, Vancouver, Tokyo, Manila, Hong Kong, or Singapore. Because of infrequent or irregular travel schedules, inhabitants of more remote islands may need two months or more to make such a journey. More than a score of shipping and airline companies offer carriers to move outbound and inbound cargo and passengers through such island centres as Guam, Port Moresby, Honiara, Nouméa, Suva, Pago Pago, Papeete, and Honolulu. This readier access, in both directions, has developed largely because islanders have demanded better facilities for travel and commercial expansion.

Radio, radiotelephone, cable, and satellite services have greatly improved communications. Two-way radio transmission is used, even in distant islands, to inform central authorities of emergencies and other critical needs. Radio broadcasts and news publications in English, French, and indigenous languages (such as Samoan, Fijian, and Tok Pisin [Melanesian Pidgin]) keep the island public informed about local and world events. Commercial television or videotaped programs from the metropolitan nations are available in urbanized centres.

ADMINISTRATIVE INSTITUTIONS

By the mid-20th century, overpopulation in a region of fragmented land areas, widely scattered communities, poor communications, inadequate resources, and rising costs of living posed a fundamental dilemma. Political responsibility for the situation rested largely with the five metropolitan nations (Australia, France, New Zealand, the United Kingdom, and the United States) administering the possessions, protectorates, and trusteeships that perpetuated a colonial heritage that was no longer popular. The amelioration of social and economic conditions seemed to await changes in the political environment, while the question of the five nations' willingness to share their territorial interests with the emerging native elites remained undecided.

United Nations trust territories

Changing administrations. Two United Nations trust territories were the first to achieve sovereignty as independent nations—Western Samoa in 1962 and Nauru in 1968. The first continued to rely on New Zealand in foreign affairs, while Nauru ended its trusteeship ties with Australia, New Zealand, and the United Kingdom. New Zealand, a Commonwealth member with a self-conscious Maori Polynesian population surviving inside its borders, continued an active relationship with other Polynesian groups—the Cook Islands, Niue, and Tokelau. Cook Islanders in 1965 and Niueans in 1974 became self-governing in free association with New Zealand, from which they required only support in certain aspects of external affairs and financial aid.

In 1975 Australia, a Commonwealth country with territorial interests in Melanesia, relinquished its hold on Papua, which it had acquired from Great Britain in 1906, and its trusteeship in northeastern New Guinea, which the United Nations had granted in 1946. The fully independent nation of Papua New Guinea was thereby created. British Fiji and Tonga gained their independence from the United Kingdom in 1970. Other British colonies achieved freedom in subsequent years—the Solomon Islands in 1978; the Ellice Islands, renamed Tuvalu, also in 1978; the Gilbert Islands, which then became Kiribati, in 1979; and the New Hebrides, which Britain administered jointly with France until 1980, when the island group assumed nationhood as Vanuatu.

France is represented in the Pacific by three overseas

territories—French Polynesia, New Caledonia, and Wallis and Futuna. Each of these territories enjoys a degree of local autonomy and is represented in the French parliament by elected delegates. Popular movements toward greater self-government or independence are active in both French Polynesia and New Caledonia.

The United States is interested primarily in the islands north of the Equator. Hawaii, formerly a territory, became the 50th U.S. state in 1959. Other U.S. territories are Guam and American (eastern) Samoa, both of which have been under civilian administration since 1950–51, after a half-century of naval rule. The United Nations Security Council established the Trust Territory of the Pacific Islands (TTPI) in 1947, covering the Northern Mariana Islands, the Caroline Islands, and the Marshall Islands, and granted administration to the United States as a strategic-area trusteeship.

Micronesians in the TTPI have since negotiated in separate groups with the United States about their status. In 1975 the Northern Mariana Islands elected commonwealth status with the United States, and a formal constitution went into effect in 1978. Three other political entities emerged in the remaining islands—the Republic of Palau (Belau); the Federated States of Micronesia, composed of Kosrae, Pohnpei (formerly Ponape), Chuuk (formerly Truk), and Yap; and the Republic of the Marshall Islands. Constitutional governments were established in the Federated States of Micronesia and the Marshall Islands in 1979, and Palau followed suit in 1981. Each of these then negotiated a Compact of Free Association with the United States, which in 1986 proclaimed that the trusteeship was terminated for the Republic of the Marshall Islands, the Federated States of Micronesia, and the Commonwealth of the Northern Mariana Islands. In 1990 the UN Security Council dissolved the UN trusteeship for all three. Palau, the last territory in the TTPI, became independent in 1994 after formally approving the compact in a referendum.

Two Pacific island territories are politically peripheral to contemporary Oceania. Western New Guinea, which was formerly a part of the Dutch East Indies, became a province of Indonesia in 1963 and was called Irian Barat until 1973, when it was renamed Irian Jaya. Easter Island has been a dependency of Chile since 1888, and development of its indigenous population is strongly oriented toward the interests of Chile.

Regional cooperation. The average Pacific island state is too small and limited in human and natural resources to function well on the world stage and has sought the advantages of joint action through regional cooperation among governments, churches, educational institutions, businesses, workers' unions, and cultural organizations. Over 200 regional bodies operate in the Pacific Islands. Notable among them are the South Pacific Forum Fisheries Agency, the South Pacific Regional Environment Programme, and the University of the South Pacific.

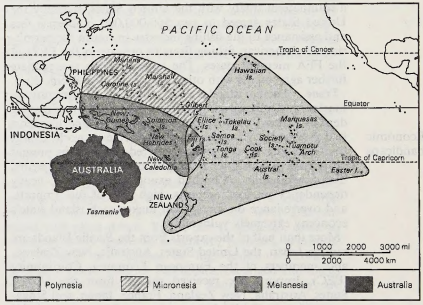
The South Pacific Commission

The first significant attempt at regional cooperation came in 1947, when the major powers that had dependencies in the region agreed to form the South Pacific Commission (SPC) to provide research and consultative services in health, social, and economic development to island governments. Members now include the five metropolitan powers—Australia, France, the United Kingdom, New Zealand, and the United States—and 21 island states and dependent territories.

Proposals to add a political dimension to the SPC's activities failed, and by the mid-1980s its image as a metropolitan body had seriously weakened its impact in the region. In part to fill this political gap, the South Pacific Forum was organized in 1971 by leaders of five independent and self-governing island countries and Australia and New Zealand; by the early 1990s the forum's members numbered 15. In 1973 the forum established the South Pacific Bureau for Economic Cooperation (SPEC) to promote regional alliances in trade, shipping and air services, telecommunications, and external aid. SPEC assumed secretariat duties for the forum in 1975, and in 1988 it was renamed the South Pacific Forum Secretariat. The SPEC has also provided liaison with other regional organizations, such as the EEC and the United Nations Economic and Social Commission for Asia and the Pacific. The forum has taken strong positions in support of the Law of the Sea and its 200-nautical-mile Exclusive Economic Zone concept and of the acceleration of the decolonization process in French Polynesia and New Caledonia, and in 1985 its members adopted the South Pacific Nuclear Free Zone Treaty. (L.E.M.)

History and cultural development

The prehistory of the Pacific Islands, the period before written materials begin, extends back at least 33,000 years, a date provided in 1985 by sites in the Bismarck Archipelago. It is probable that human settlement significantly antedates the limit of radiocarbon determination of 40,000 years. Migration extended into the 2nd millennium AD and resulted in the populating of every habitable island. Since the arrival of Europeans in the early 16th century, the prehistoric world has been transformed to varying extents, at first by contact with passing explorers and then, from the late 18th century, by the influence of more permanent visitors: castaways, beachcombers, missionaries, and traders. During the 19th and 20th centuries settlers, labourers and other immigrant communities (predominantly Indian and Chinese), and European administrators arrived. Missionaries, settlers, and immigrants still make up significant segments of the population on the islands today, even though European governments, with the exception of the French, have entirely withdrawn from the region.



The culture areas of the Pacific.

Both the prehistorical and historical periods present problems in the evidence of human activity. Archaeology has provided dates for the earliest settlement; however, until more sites are uncovered, sampling error makes certainty extremely difficult. Linguistics, using a chronology of sound changes, traces the time and place of dispersion of language groups; but a considerable number of the languages of Oceania are as yet unstudied and unclassified. Genetics, making use of the distribution of leukocyte antigens in present blood-group samples, has been able to establish, between contemporary human groups, connections that reveal past migrations. Unfortunately, systematic sampling has not yet been carried out.

The distortion of later Oceanic history

The period of Oceanic history for which documentary evidence is available also presents difficulties. The documents are chiefly of European origin and are, therefore, the products of people who may not accurately have recorded a culture different from their own—a culture they perceived and understood only imperfectly. This distortion can be corrected to some extent by using the findings of social anthropology and the oral traditions of the Pacific island people, but these sources are difficult because, by definition, they are contemporary. They may describe the past inaccurately because they serve contemporary purposes; they do not record the past for its own sake. But

the main historiographic problem of Oceania is its diversity. Some 10,000 islands scattered across 500,000 square miles of ocean, a variety of cultures, hundreds of mutually unintelligible languages, and diverse historical experiences make it hard to generalize.

PREHISTORIC TIMES AND THE PROLIFERATION OF CULTURE

The influence of physical geography. The physical diversity means that contact between islands has never been easy because of the distances involved; it also means that the physical environment is not only isolated but varied. The large continental islands of Melanesia—such as New Guinea, New Caledonia, and Viti Levu (Great Fiji)—have a different physical basis of life and a wide variation in climate and fertility; moreover, their rugged terrain has made for social isolation. The smaller volcanic high islands, such as Samoa and Tahiti, have greater homogeneity and are very fertile, supporting life well above the subsistence level. The coral atolls, the low islands of Oceania, support a narrower range of vegetation; they are more exposed to bad weather and support an existence closer to subsistence, except for the rich marine life in the lagoons.

Physical environment does not determine the kind of society that exists, but it does set limits to it. The large islands of Melanesia produce marked differences between people of the coast and those of the interior. Their long coastlines have acted as a filter for many different arrivals in Oceania. The valleys have perpetuated differences. Thus Melanesia is characterized by many small groups of people, divided from each other by language and custom. Political and social organization has been small in scale. The margin above subsistence has not been great enough to allow for elaborate ceremonial. The high volcanic islands of Polynesia offered no such barriers to social and political unity. Their fertility allowed the development of elaborate social, religious, and political ceremonial. The low islands also allowed homogeneity and wide social groupings, but their land resources offered no great margin for ceremonial life. These contrasts within Oceania were obvious to early European visitors, but they conceal a similarity: whether small, with leadership a matter of acquiring influence rather than hereditary position, or larger, with chiefs surrounded with awe and reverence, Oceanic societies all rested on the principle of reciprocity. Every gift or service had to be returned.

Origins of Oceanic peoples. Jules-Sébastien-César Dumont d'Urville, an early 19th-century French navigator and explorer, classified the islanders as Melanesian, Micronesian, and Polynesian. The apparent differences among the islanders were regarded as evidence of separate waves of ethnically different people out of Southeast Asia (a discredited variant theory traced the Polynesians to South America). Recent research suggests that the differences arose within the islands themselves, through the intermixture of an original settlement of non-Austronesian-language speakers from Southeast Asia with a later wave of Austronesian speakers. The earlier wave of settlement occurred in Melanesia at least 33,000 years ago and probably, since New Guinea and Australia were then linked by land, at dates contemporaneous with Australian dates of settlement, extending back to the limit of radiocarbon determination of 40,000 years. Secure dates in the interior of New Guinea approach 30,000 years ago.

The later Austronesian speakers, who are associated with the pottery called Lapita ware (from the site at which it was first discovered in New Caledonia), established themselves in the Bismarck Archipelago about 4,000 years ago and then spread to Fiji, Tonga, and Samoa, which have been regarded as the Polynesian homeland. They then dispersed to Micronesia and eastern Polynesia. New evidence, however, has led to disagreement among prehistorians about the Lapita complex: it may have arrived in Fiji with a later wave of seafaring immigrants. There is also disagreement about the speed with which the Lapita culture, distinctively linked with the Polynesians, moved from Southeast Asia through Melanesia into Fiji and thence to eastern Polynesia; the theory of rapid spread has been challenged on the basis of the length of time needed for a Polynesian homeland to develop in the Fiji-Samoa-Tonga region. Nevertheless, the known evidence of the Lapita culture outside the Bismarcks is dated after 3,500 years ago. In Fiji and western Polynesia the date is before 1000 BC; in eastern Polynesia it is about 500 BC. Allowing for new evidence to appear and for arguments as to its significance, there are still grounds for supposing that, from a Polynesian homeland region (Fiji-Samoa-Tonga), eastern Polynesia was settled by the Lapita culture bearers. It is possible that the Marquesas were settled as early as the 2nd century BC, rather than AD 300, a date at which settlements may have occurred in Hawaii. The Society Islands were occupied by at least the 9th century AD.

DEVELOPMENT OF DISTINCTIVE CULTURES

Oceanic societies, at the time of European discovery, were Neolithic. They had developed a technology based on stone, bone, and shell; they cultivated tubers and tree fruits, all of which were of Southeast Asian origin, with the exception of the sweet potato, which derives from South America and which culture had spread through most of Polynesia in pre-European times but only marginally into Melanesia. This Neolithic cultivation was associated with three domesticated animals: the pig, dog, and chicken, also of Asian ancestry. The coastal people had developed techniques of fishing and considerable skills as sailors; navigation between the closer islands was well-developed. Some skills were lost: pottery, for example, disappeared in the Samoas and the Marquesas shortly after initial settlement. And with the control of the environment that Oceanic technology offered, there is evidence of overpopulation, settlement spreading into less-favoured areas of the islands and being fortified, as in the Marquesas. (For a further discussion of the indigenous cultures, see below *Traditional cultures of the Pacific Islands*.)

THE PERIOD OF EUROPEAN EXPLORATION

The 16th and 17th centuries. The Oceanic world was not a static one, but changes were slow compared with those that attended European discovery. Vasco Núñez de Balboa was the first European to sight the Pacific, in 1513; seven years later Ferdinand Magellan rounded South America and sailed across the ocean, missing the main island groups but probably encountering Pukapuka in the Tuamotu Islands and Guam. After his death in the Philippines, his expedition encountered some of the Carolines. These northern islands were further explored by the Spaniards as they established a galley trade between Manila and Acapulco. The next major Spanish discoveries were made by Álvaro de Mendaña de Neira and Pedro Fernández de Quirós. In 1567 the former set out from Peru to discover the great southern continent that was believed to exist in the South Pacific. He found the Solomons but failed to find them again on his second journey, during which he died. In 1606 his chief pilot, Quirós, after finding part of the Tuamotu Archipelago, found the northern Cooks, Tikopia, and the New Hebrides. One of his companions, Luis Váez de Torres, found southeastern New Guinea and then the strait (later named for him) between that island and Australia, although the discovery was unknown to later sailors. These Spanish expeditions were motivated by the search for riches, by zeal to extend religion, and, in the case of Quirós, by an interest in discovery for its own sake. But with the voyage of Torres, the Spanish effort was ended.

Sighting by Balboa

Thereafter, the Dutch, who were already established in Indonesia, entered the Pacific. They too looked for a southern continent. In 1615–16 the Dutch navigator Jakob Le Maire came from the east through the Tuamotus to find Tonga, New Ireland, and New Hanover. In 1642 Abel Janszoon Tasman, sailing from Batavia (now Jakarta), the Dutch headquarters in the East Indies, saw New Zealand, Tonga, some of the Fijis, and New Britain. The Dutch were primarily interested in commerce; they found none. Tasman thought that New Zealand was part of the great southern continent. The effect of these visitors on Oceania was transitory. They stayed for periods of at most a few months. Their contacts with the islanders were those

of simple barter, but the demands they made upon food supplies often caused hostilities in which some European and many islanders' lives were lost, as on Guadalcanal and in the Marquesas during Mendaña's visits.

The 18th century. During the early 18th century the extent of Oceania was further revealed. The English pirate William Dampier visited New Hanover, New Britain, and New Ireland in command of a Royal Navy ship. Dampier was the forerunner of scientific exploration, and he proved that those islands were separated from each other and from Australia. In 1722 the Dutch admiral Jacob Roggeveen crossed the Pacific from east to west on a voyage of exploration that also had commercial objects. He found Easter Island, more of the Tuamotu Archipelago, the northern islands of the Society group, and some of the Samoan islands. These voyages were not essentially different from earlier ones, but they foreshadowed the scientific interest of the later 18th century. The execution of that interest was delayed by European wars. But in 1765 the English admiral John Byron (grandfather of the poet), who was sent by the British Admiralty in search of the supposed southern continent, found more of the Tuamotus and the southern Gilberts. In 1767 Samuel Wallis and Philip Carteret followed, but their ships were separated as they entered the Pacific. Wallis found Tahiti, more of the Tuamotus, and the Society Islands, while Carteret found Pitcairn and rediscovered the Solomons of Mendaña, although he did not so identify them. This was left to the French following Louis-Antoine de Bougainville's visit in 1768, during which he also found some of the New Hebrides and Rossel Island in the Louisiade Archipelago.

The explorers contributed greatly to Europe's knowledge of the Pacific. Dampier's *A New Voyage Round the World* and Bougainville's description of the "noble savage" in Tahiti were particularly influential. The interest their journeys created was in part responsible for the instructions given to the greatest of all 18th-century explorers of Oceania, James Cook. After three voyages he left others little to do but fill in occasional details of Oceania. Cook was sent in 1769 to observe the transit of Venus at Tahiti and then to search for the great southern continent. Cook found some of the Society Islands, but he also circumnavigated New Zealand, and he defined the limits of eastern Australia. During his second voyage (1772–75), he proved that there was no southern continent, but he also made further discoveries in Oceania: in the Tuamotus, the Cooks, the Marquesas, Fiji, New Caledonia, the New Hebrides, and Norfolk Island. His third voyage (1776–79) was mainly concerned with the North Pacific, during which he found some of the Tongan group, Christmas Island, and Hawaii. He had completed the main work of discovery with an exactitude thitherto unknown. Although his contacts with islanders were not in essence different from those of his predecessors, his relations with them were nevertheless more prolonged and more humane. And his exploration of eastern Australia, through the account of his naturalist, Joseph Banks, was of great importance in Oceania, for it led to European settlement close to the islands.

Importance of Cook's travels

EARLY EUROPEAN SETTLEMENT

Itinerants. With the establishment in 1788 of the Australian settlement, Oceania became a source of supply. In 1793 pigs from Tahiti were landed at Sydney, and until 1826 the trade, although subject to fluctuations due to the competition of other cargoes and Tahitian wares, was important. The competition among Europeans for sandalwood, pearl shell, and bêche-de-mer (sea cucumber), valuable cargoes that attracted ships from the Australian colony, further involved Oceania with the European world. In 1804 sandalwood was found in Fiji, and for the next 10 years it attracted European traders. The sealing industry drew seal hunters to New Zealand, and fur traders, in the 1790s, wintered in Hawaii. All of these contacts began to affect Oceanic societies because they were sustained and prolonged. Together with the castaways and beachcombers who had begun to live in the islands from the days of first European contact and who increased in numbers with commercial shipping, these European contacts began to transform Oceania. Castaways, such as HMS *Bounty* mutineers who went to Tahiti in 1789, began to alter the political state of the islands they lived in by supporting with their muskets the chiefs who befriended them.

The *Bounty* mutineers

Missionary activity. Such Europeans had an important effect, but they were dependent upon the people they lived with for their survival. Very different were the missionaries, who traveled to Oceania with the deliberate intention of changing Oceanic society. In 1797 the London Missionary Society (LMS), now the Congregational Council for World Mission, sent a party to Tahiti. After some vicissitudes the missionaries converted the chief Pomare II, who controlled the area of Matavai Bay, where Europeans had called since Wallis' discovery. The LMS failed in its first attempts in Tonga and the Marquesas, although it was more successful in Huahine, the Tuamotus, the Cook Islands, and later in Samoa. Other missionary societies followed. In 1822 the Methodists began to work in Tonga; in 1835 they went to Fiji. In the 1840s Roman Catholic missionaries began working in New Caledonia, and the Church of England began to penetrate into Oceania from New Zealand. These missionaries encountered societies in Polynesia that already had a problem of law and order from the influence of European beachcombers and traders. Its solution was to create missionary kingdoms, in the case of British missionaries, or the establishment of direct political control, in the case of the French.

In Tahiti, Hawaii, and Tonga, native chiefs whose power was established by their access to European arms and support not only became kings but took missionary advisers and missionary-designed codes of law. In 1819 Pomare II of Tahiti promulgated such a code. In Tonga, Taufa'ahau took the name of George in 1833, and in 1845, when he took the Tongan title of Tu'i Kanokupolu, he became "king" of Tonga; in 1862, under the influence of the Reverend Shirley Baker, he adopted a constitution. By attempting to enforce a scriptural code of law, these missionary kingdoms were an answer to the problems of European lawlessness in the islands. If the missionaries could not prevent the sale of arms, they could at least make sure that these passed into the hands of friendly chiefs. But the authority of these "kings" was challenged from two sides. By becoming Christian they had cut themselves off from the mana (a Polynesian religious concept sometimes described as an all-pervasive energy) that came from the old gods, and this produced nativist reactions. In Tahiti in 1830 there was a revolt against the new Christian order by supporters of the old ways; in Tonga in 1831 there was a similar reaction. In Samoa, where the holder of the Malietoa (district chief) title had embraced Christianity from Tahitian missionaries, there were heretical movements. If traditional beliefs thus resisted the chiefs and their missionary supporters, the European traders also resisted the political authority of the kings. Dissidents and heretics looked to these Europeans for leadership, and these Europeans looked to their own national governments for protection.

Missionary kingdoms

In Melanesia the story is somewhat different. In Fiji the missionaries who landed in 1835, accompanied by an envoy from George of Tonga, made no headway with the rising chief Cakobau, who was not converted until 1854, when his fortunes were at a low ebb and he needed Tongan support. Elsewhere in Melanesia, the absence of chiefs meant that missionary work had to be conducted with small groups of people and repeated every few miles. There was no wholesale conversion of the kind that had happened in Polynesia. The attempt of the LMS in the New Hebrides in the 1840s came to nothing. The Anglican Melanesian mission in the Solomons made slow progress in the 1850s. In New Guinea, mission work, divided into four spheres of influence in Papua, did not begin systematically until the 1870s. Micronesia was a backwater. The Spaniards had established missionaries in the Marianas in 1668, but the missionaries in the Carolines were killed in 1733. The main effort came from the Hawaiian Evangelical Mission in the 1850s. The general effect of mission activity was, nevertheless, the same as it was in Polynesia. It dissolved the old ties of society by attacking the supernatural sanctions that supported leadership and social mores. It altered the political structure of Oceanic soci-

eties. It incidentally introduced both European goods and the desire for them. The missionaries themselves acted as intermediaries between Oceanic societies and other Europeans—as political advisers, as agents, and as interpreters.

Growth of trading communities. Beachcombers and castaways preceded missionaries in many of the islands, but the growth of trading communities was in part the result of the missionaries' work in restraining native violence. Those traders were initially pork traders in Tahiti, but European captains followed valuable cargoes from island to island. When the supply of sandalwood was depleted in Fiji by 1813, the traders then found it in Hawaii in the 1820s, in the New Hebrides in 1825, and in New Caledonia in 1840. Pearl shell attracted traders to the Tuamotus in 1807. The sandalwood trade declined as supplies were exhausted. Pearling declined as the native inhabitants took reprisals for the atrocities that had accompanied both trades. The demands of the Oceanians also changed the character of trade. Once native polities were established, the demand for muskets fell off; under missionary influence, the demand for alcohol was limited. What the islanders now wanted was clothing and hardware. The exchange traders were not guilty of the cruelties that those looking for sandalwood or pearls were apt to perpetrate, and exchange trading encouraged the growth of resident agents in the islands, a development that met the needs of the whalers who came ashore to refit their vessels. After 1840 it also met the needs of the staple trade of the islands—coconut oil, used for soap and candles. Copra trading, from which the oil came, became the mainstay of European trade because even islands that had no other resources had coconut palms.

European trade centres

Such commerce promoted the growth of the port towns and of resident trading communities. Papeete in Tahiti, Apia in Samoa, and Levuka in Fiji became European centres, including not only respectable traders but also lawless men who might be escaped convicts from New South Wales (Australia) or others escaping from the rules of settled societies. These were frontier towns and could be regulated only with difficulty by native kings or by visiting European captains.

Establishment of plantation societies. The problem became more urgent with the advent of permanent European settlers. In Fiji, for example, after Cakobau's first offer to cede the islands to Great Britain in 1858, Europeans arrived to establish plantations, at first of coconuts, then, during the American Civil War, of cotton, and afterward of sugar. The development in Samoa was similar. But planters needed land on a much larger scale than traders, and they needed labour in much greater quantities to work the plantations. Both land sales and labour recruitment caused friction, for "ownership" was not an Oceanic concept; thus, land titles were disputed or resented, and the recruitment of labour often caused the breakup of traditional societies if too many males left their communities and the creation of immigrant labour communities if they did not. By 1870 there were 2,000 such permanent European residents in Fiji. Politically the settlers had an interest in stability, and economically they needed security of title to land and a supply of labour. Neither requirement was satisfied by the missionary kingdoms. Nor was it satisfied by native governments that were not guided by missionaries. In Tahiti, Tonga, Samoa, and Fiji no native authority was able to keep order in the novel circumstances created by European enterprise; in any case, the native kings were themselves open to challenge within their own societies. Pomare II encountered revolt in Tahiti, Samoan politics were always a matter of rivalry between chiefs, and Cakobau's government was threatened by the Tongan chief Ma'afu, who had established his own confederacy in the Lau Islands of the Fiji group.

COLONIAL RULE

Involvement of foreign governments. Eventually the unstable internal conditions in the Pacific began to draw in European governments, all of which acknowledged some responsibility for the protection of their nationals and their property. The French government was the first to intervene, after the expulsion of two Roman Catholic missionaries from Tahiti in 1836. In the same year two more were deported from Hawaii. In 1839 the Archbishop of Chalcedon suggested regular association between the Roman Catholic missions and the French navy, but the French government was also aware of the need for a good naval station for the fleet and for French commerce and for a place of penal settlement. Abel DuPetit-Thouars thus took possession of Tuahata and the southeast Marquesas in 1842 and in the same year persuaded the Tahitians to ask for a French protectorate, which was formally granted in 1843. In 1853 the presence of French missionaries in New Caledonia led to French annexation, possibly for fear of British action, certainly to establish a penal colony (to which convicts were transported until 1897). Other European nations intervened for different reasons. In 1857 August Unshelm, as agent for J.C. Godeffroy and Son, set up the company's depot at Apia, and Samoa became the greatest trading centre in the islands; and even when Godeffroy failed in 1879, the Deutsche Handel und Plantagengesellschaft (German Trading and Plantation Company) took over, and Samoa remained the favourite colony of the colonial party in German politics. British nationals had trading and plantation interests in the islands; to give some protection to these interests, the British government had appointed consuls to those islands governed by recognizable rulers, but their powers to maintain order were limited and, except for the visits of warships, unenforceable. The United States also appointed consuls.

Naval stations and penal colonies

The rivalry between these officers and between European entrepreneurs, and the involvement of both in the internal politics of Oceanic societies, merely emphasized to metropolitan governments the disordered condition of the islands. In Tahiti the problem was resolved by French annexation. In Samoa, after a tripartite supervision set up by the Samoa Act of 1889 came to grief in European rivalries and Samoan factionalism over chieftainships, an agreement of 1899 divided the Samoa group between Germany and the United States; Britain received compensation elsewhere. Britain's main concern was in fact with the activity of its nationals: in Fiji, where it accepted the offer of cession of 1874, it did so primarily because native authority had broken down. But Britain also had been concerned with the labour trade by which the Queensland (Australia) plantations took islanders, who were sometimes recruited under doubtful or brutal conditions. In the 1860s this trade flourished in the New Hebrides, and violence there led the missionaries to protest. Then the labour trade moved north to the Solomons, where again there was violence, including the murder of the Anglican bishop in the Santa Cruz group, from which five men had been taken by recruiters. The British solution was the Western Pacific Order in Council (1877), which empowered the governor of Fiji to exercise authority over British nationals and vessels in a wide area of the western Pacific. The problem still remained, however, of non-British nationals in islands that had neither native kings nor European governors, especially those of Melanesia.

The annexation of Melanesian islands

European government, like both mission and commercial enterprise, had been slower to penetrate Melanesia. Missionary activity did not begin in New Guinea until 1873. There was not much labour recruiting. The first main activity was the gold rush of 1877, but German traders had arrived on the northern coast in 1873, followed by the firm of Hernsheim & Co. (a general trader) in 1875. Such foreign interest produced a demand in the Australian colonies for annexation for reasons quite unconnected with the internal situation in the islands. German interests were marked in Micronesia, French in the New Hebrides. A number of groups in Australia also looked on New Guinea as a rich possession. But the British government, notwithstanding Queensland's abortive attempt at annexation in 1883, would not annex unless the Australian colonies paid the cost of administration, the same argument it was applying to New Zealand's interest in the Cook Islands. When the Australian colonies agreed to pay, the British government acted. Southeast New Guinea was declared a protectorate in 1884 and annexed four years later; the Cooks became a protectorate in 1888 and were annexed in 1901. Germany annexed northeast New

Guinea in 1884, including the Bismarck Archipelago; in 1886 it took possession of the northern Solomons (Buka and Bougainville). The British established a protectorate over the rest of the Solomons in 1893. In Micronesia the Germans, after an attempt to annex the Spanish possession of the Carolines in 1885, finally bought them from Spain with the Palaus and the Marianas (excepting Guam) in 1899. They had annexed the Marshalls in 1885 and, under a convention with Britain of 1886, the phosphate-rich island of Nauru. By that convention Britain's interest in the Gilberts was recognized, although no protectorate was declared until 1892. The Ellices were added to it, the group becoming the Gilbert and Ellice Islands. France declared a protectorate over Wallis and Futuna in 1887, and in the same year a convention set up a mixed British and French naval commission in the New Hebrides. Its authority was limited, and in 1906 a Condominium was agreed upon by which such difficult legal questions as land title were settled and a joint administration set up. The process of partition was completed by the United States, which took Guam in the Spanish-American War (1898), annexed the republic of Hawaii that same year, and obtained American (eastern) Samoa by agreement with Germany and Great Britain in 1899 and by deeds of cession (1900, 1904) from island chiefs.

Patterns of colonial administration. With the exception of Tonga, which remained an independent kingdom under British protection (from 1900) with a consul who was not to interfere in internal affairs, almost the whole of Oceania passed under the control of European powers and the United States between 1842 and the end of the century. Having acquired colonies, these powers governed Oceania with metropolitan institutions modified to a degree by local circumstances. Thus, Britain reproduced in the islands the pattern of crown colony government, which derived from its own political development: a governor who represented the king; an executive council of senior officials; and, where the European population justified it, a legislative council to advise the governor. Within this form of government, administration was adapted to local conditions. Thus, Sir Arthur Gordon, the first governor of Fiji, set up a system of native administration that incorporated the chiefs: the island was divided into provinces and districts that, on the information available to him, represented the old divisions of Fiji, and over each he tried to select the chief to take administrative office. Even in Melanesia, where chieftainship was not highly developed, the British attempted to appoint chiefs who were men of influence. The first administrator of British New Guinea was a former officer in Gordon's government, William MacGregor, who first tried appointed chiefs and then settled for village constables. The Australians, who took over British New Guinea in 1906 and rechristened it Papua, followed the British pattern. The first Australian governor, Sir Hubert Murray, although he introduced measures of native development, still preserved the British pattern of colonial government, as did New Zealand in the Cook Islands. Beneath the governor, district administration tried to incorporate both native leadership and the technical and professional direction of specialized departments (such as agriculture and health), as well as those departments that dealt with the questions raised by European settlement (such as land and labour).

The British system

Other nations had different patterns. The Germans, if only because domestic German politics made colonies an incident of European policy, tried to administer their colonies through commercial companies. In northeastern New Guinea, the German New Guinea Company was commissioned to administer the colony as a commercial enterprise. Only when it failed did the imperial government assume responsibility (1899). In the Marshalls, the German firms known as the Jaluit Gesellschaft became a chartered company under a government commissioner in 1885. In Western Samoa, in the first decade of the 20th century, the governor Wilhelm Solf attempted to control the importation of Chinese labour for the plantations and tried to enlist Samoan interest for the government, but he was subject to the pressure commercial interests were able to exert in Germany itself. In the French territories, colonial rule meant assimilation to French institutions. The governor was analogous to the prefect of a French *département*, assisted by an administrative council and from time to time by a general council drawn from French citizens. Where such a council existed, its powers were limited to an optional section of the budget, the rest of which was obligatory. In effect, the governor ruled by administrative decree. In the U.S. territories there was also a marked assimilation to U.S. metropolitan forms of government. When Hawaii was annexed in 1898, the president of the republic became a U.S. governor. When eastern Samoa was given to the United States under the convention of 1899, President William McKinley, in 1900, placed it under the authority of the Department of the Navy, the commanding officer of the station also becoming governor and administering the islands with the help of his technical officers and the advice of a Samoan fono, or legislature. These colonial governments were adapted to local circumstances. In the Polynesian islands and in Fiji, Britain and Germany attempted to incorporate the authority of the chiefs into their governments, both as advisers and as local officials in the districts, as did the United States in American Samoa. But in both Hawaii and Tahiti the old system of rank had broken down under the impact of missionaries, traders, and settlers, so that it could not be used for administrative purposes but had to be replaced by appointed local officials. In Melanesia, where there was in general no chiefly authority that could be used because influence was acquired by criteria that made it fleeting, the colonial powers had no choice but to use appointed local headmen. The Germans and the British used appointed headmen in New Guinea and the Solomons, and the system was supervised by a patrol of European officers with an escort of armed native police. The patrols were brief and infrequent, however, and their effect was limited.

The Germans in Samoa

Colonial rule after World War I. This pattern of colonial rule in Oceania was altered by the outbreak of World War I in 1914. An Australian force took German New Guinea, and a New Zealand force took German (Western) Samoa; Japan took the Carolines, the Marshalls, the Palaus, and the Marianas. At the end of the war these German territories, together with Nauru, were retained by the occupying powers as mandates under the League of Nations. The professed aim of the administrators was to help the people of these territories to stand on their own feet under the strenuous conditions of the modern world.

This stress on Oceanic interests was not new, but it became an international standard. Still, the first step was the establishment of government control and of law and order before any other measures could be taken. In New Guinea, in the Solomons, and in many parts of Melanesia, the interior was rarely known, let alone controlled. So the attention of colonial government was concentrated on opening the interior. The government-sponsored exploration in 1933 of the grass valleys of New Guinea, which had been known for a decade by missionaries and miners, presented the Australian administration with the problem of 750,000 new people. Australia's resources for administering New Guinea were spread thinly. Health and education were left to the missions. By contrast, Fiji was more developed. With the importation of labourers from India to work the sugar plantations from 1879, and with the reforms in the indenture system (which came after a commission of 1909), the large Indian population of Fiji received the attention of the government in education and health matters, but the increase in this population raised the difficult question of the Fijians' future. They played a minor part in the economic life of the colony, and the trend of official policy was to preserve them within their villages under a separate system of administration, which was reorganized in 1944 as the Fijian Administration—to be virtually a state within a state. But the government's resources were not enough to introduce welfare measures, nor was it able to promote development on any great scale. A good deal depended on the missions and other private organizations. In Samoa, where the old society had retained its organization, there was resistance to change. The New Zealand administration of Western Samoa had

The role of the missions

begun with the objective of promoting the welfare of the native race, which meant health, education, and better use of the land. By recognizing Samoan councils it tried to ensure Samoan support, but the policies broke down in execution. In American Samoa the U.S. Navy provided welfare services as part of its routine work but could do little more, because its principal concern was running the naval base. In French Polynesia native policy aimed at making the people French citizens. The welfare services were directed from Paris, but limited funding was available.

WORLD WAR II AND ITS AFTERMATH

The islands were affected directly by external events such as the Great Depression of the 1930s and the fluctuations in world markets for copra, sugar, and other products of Oceania. The principal achievements of the colonial powers were improvements in medical care, partly by controlling such endemic diseases as malaria in Melanesia, and a rough balance between European and indigenous interests. But welfare policies and island administration were both interrupted by World War II. The Japanese were established in the north of Oceania, where they treated their mandates as part of Japan itself. In 1941 they advanced into the rest of Oceania, reaching and controlling most of New Guinea and, at the peak of their advance, much of the Solomons. New Caledonia, the New Hebrides, Fiji, and the islands of Polynesia were not occupied, but the effects of the war made colonial government there secondary to military operations. After the war the Trusteeship Council of the United Nations replaced the mandates and all of the colonial powers accepted that independence or self-government was the aim of their rule. The Oceanians themselves had been exposed to more intensive European (and Japanese) influences, and their horizons had widened. The colonial powers felt a greater urgency to promote development and to make available greater resources to achieve it.

Beginnings of self-government. Colonial governments were reorganized to give indigenous people a part in government. In Western Samoa in 1947 the Legislative Council was given a Samoan majority and considerable powers. In American Samoa naval rule was replaced in 1951 by civilian control, and a legislature of two houses was set up, which by 1960 became a lawmaking body of Samoans. In French Polynesia and New Caledonia, elected assemblies were given considerable local autonomy in 1956; both territories chose to stay within the French Community in 1958. The trend toward a limited degree of internal autonomy and increased political participation by native residents continued in the 1960s in Fiji, the Cook Islands, and other Pacific dependencies.

Independence movements. Despite UN support for self-determination after 1945, the Pacific has not been completely decolonized, because of strategic and economic reasons. France granted French citizenship in its Pacific territories in 1946; however, it jailed nationalist leaders in Tahiti and New Caledonia during unrest in the 1950s. The United States, which had strategic interest in the North Pacific territories it acquired, consolidated military positions in the area following World War II. The Northern Marianas chose to become a commonwealth of the United States, whereas resource-poor Marshall Islands, Palau, and the Federated States of Micronesia all voted for free association with the United States, granting them self-government but continued economic and military ties. The Cook Islands and Niue chose free association with New Zealand, granting them local self-government and New Zealand citizenship and subsidies. However, most of the other Pacific Islands achieved independence by 1980. The newly independent territories represented a sizable addition to the ranks of microstates.

The speed of political development in the Pacific Islands was partly a matter of external pressure in the United Nations; but the colonial governments, with the exception of the French, were already moving toward self-government or independence. There were no mass nationalist movements, as in Africa and Asia, to whose demands colonial governments responded. The reaction to European rule usually took the form of nativistic movements or cargo cults in which rituals attempted to secure "cargo" diverted by Europeans. Occasionally, as with the Mau ("Strongly Held View") movement in Western Samoa in the 1920s and '30s, there was more overtly political action. In Fiji and Papua New Guinea, political parties formed. In the French territories of French Polynesia and New Caledonia, European-style political parties have demanded greater local autonomy and, as a minority, independence. Violent confrontations did break out in the 1980s between separatists and French supporters, followed by protests in the 1990s against French nuclear weapons testing in the region. As a result, accords were signed in 1988 and 1998 regarding a timetable for self-determination. However, mass nationalist movements were largely absent from the region, in part because colonial governments had given priority to Oceanic interests but mainly because of the nature of Oceanic societies, in which kinship ties and a preference for consensus as "correct" behaviour led to "the Melanesian way" or "the Pacific way" as a style of politics.

Absence of mass nationalist movements

—(F.J.W./Ed.)

For statistical data on the land and people of the Pacific Islands, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 967 and 977, and the *Index*.

TRADITIONAL CULTURES OF THE PACIFIC ISLANDS

General considerations

INDIGENOUS PEOPLES

Archaeological research has been undertaken in the Pacific Islands to trace the region's migration and settlement patterns. The region is vast, however, and understanding is both incomplete and uneven, despite collaboration among archaeologists, linguists, and physical anthropologists.

Earliest occupation

The area first penetrated by human immigrants was limited to New Guinea, Australia, and Tasmania, which are thought to have been connected in the Pleistocene Epoch (1,600,000 to 10,000 years ago) in what has been called the Greater Australian continent. Excavations in southeastern Australia reveal that people of modern type, *Homo sapiens sapiens,* had arrived there by 40,000 years ago. Migration eastward from New Guinea began by the end of the Pleistocene, but movement beyond the Melanesian region required sophisticated watercraft and did not occur until 4,000 to 3,000 years ago. Then, generations of voyagers ranged eastward into Polynesia (Fiji, Tonga, Samoa) and northward into eastern Micronesia, bearing with them distinctive physical, cultural, and linguistic characteristics.

Aboriginal groups. The reconstruction of the cultural history of Oceanian peoples has produced many theories and explanations. The opportunities for groups to develop in isolation and over generations to evolve adaptive responses to special conditions were numerous. Just as important to the ethnic diversity of the area is the hybridization produced by centuries of ebb and flow of migrating peoples. Such diversity is certainly the combined result of isolation or its lack, adaptation to differing environments, mating patterns, and natural selection.

Authorities disagree over whether early hominid forms known from fossil finds to have inhabited Indonesian, Melanesian, and Australian areas were progenitors of peoples found there in recent times. There is similarly a great diversity of opinion on just which groups may have been antecedent. Some specialists once proposed commonality of origin or at least some link between the indigenous peoples of sub-Saharan Africa and the darkly pigmented peoples of Melanesia. Blood-group research studies, however, showed no evidence for such claims. In general, notions of

a series of massive migrations of Asian peoples, and their interbreeding with antecedent peoples, seem to most specialists to be simplistic. Blood-group genetic research has contributed to a better understanding of certain populations, however, particularly the Negritos of the Philippines, who are thought by some researchers to be descended from late Pleistocene peoples who inhabited the islands of Southeast Asia and who are called Proto-Malays.

Main ethnic elements

The ethnic composition of Pacific peoples is a composite of many elements; Negrito peoples have survived or been reported in remote areas of Melanesia, Australia, and Tasmania in addition to the Philippines. The Australian Aboriginals may be described as a subcontinental isolate, although some genetic evidence points to ancient ties with New Guinea populations. Generalizations about the normative type of a particular culture area are anyhow misleading. There are general tendencies, of course, but there are also such ranges of characteristics within virtually all areas that a representative of one area could easily be assimilated in one or more other areas.

Recognizing the difficulties of generalization, then, it may still be noted that Tongans, Fijians, Samoans, and Hawaiians are often tall and inclined to corpulence. Pigmentation generally tends to be darker among Melanesian groups than is usual in Polynesia or Micronesia. In some of the Melanesian islands, such as the Solomons and New Britain, skin colour is virtually blue-black in some individuals. The ruggedness of Australian Aboriginal facial characters exceeds that of other regions of the Pacific Islands. The people of the New Guinea Highlands are of relatively short stature and have distinctive facial features.

Migration sequences. The influxes of population into Australia and Tasmania continued until at least the final phase of the Pleistocene (12,000 to 10,000 years ago). Tasmania was then cut off by the rise in sea level and the Bass Strait was formed. New Guinea, because of its size and location, continued to receive migrants well past the end of the last ice age. Migration to Micronesia and Polynesia appears not to have begun until 4,000 to 3,000 years ago: the islands of these areas were thus the last habitable portions of the Earth to be populated.

Chronological sequences based on pottery styles, types of fishhooks, and other dating techniques suggest that human presence in Micronesia was earlier than the oldest radioactive-carbon date yet found there of approximately 3,500 years ago. The evidence points to the entry of people into central Polynesia from eastern Melanesia. In due course, migrations radiating from both western and eastern central Polynesian island groups are thought to have settled the more distant reaches of Polynesia. The Marquesas were reached by people from Samoa as early as AD 300. Some authorities have suggested that settlement may have been even earlier. Hawaii was first reached midway to late in the 1st millennium AD by voyagers from the Marquesas, who were followed several centuries later by those from the Societies. The Societies also appear to have been the point of origin for the Polynesians who settled the Cook Islands, and people from either the Society or Marquesas islands reached New Zealand before AD 1000. Although the validity of the radioactive-carbon dates that have been recorded for Easter Island, indicating human presence by 400, has been questioned, the dates are in accord with the time of settlement of the Marquesas (whence Easter Island probably was reached).

The watercraft used by the ancient migrants from Asia were probably crude rafts or coracles; it is likely that the highly sophisticated sailing vessels so distinctive of the cultures of the islands developed over a long period of time. However simple the early craft, those that were used thousands of years later to explore and settle the far reaches of Micronesia and Polynesia were impressive indeed. These were great double-hulled vessels capable of carrying a hundred or more voyagers and their provisions for a voyage lasting a month or longer, as well as carrying domesticated animals and plants for propagation in a new homeland.

The issue of deliberate versus accidental voyages of discovery in the Pacific has prompted an impressive array of arguments from specialists on both sides. There is considerable evidence that some islands were, indeed, populated by seafarers who, because of a storm or mere random sailing, came unexpectedly upon their new home. There is also a great deal of evidence, however, of the navigational skills and abilities of Oceanian peoples and of certain deliberate voyages of discovery. It is also clear that return voyages were made to points of origin.

Language groups. The study of Oceanian languages sometimes requires generalizations that are no more than statements of probabilities. Many languages have yet to be studied, and some are extinct. Nevertheless, linguists offer the following characterization.

Two major language classifications

The languages of Polynesia, Micronesia, Melanesia, and Indonesia may be divided into two major groups: the Austronesian (formerly called Malayo-Polynesian) and the Papuan (or non-Austronesian). Virtually all of the latter group are found on New Guinea and islands to the east and west of it. Some 750 or more Papuan languages have been identified. The number of Austronesian languages probably exceeds 1,200. Their geographic distribution extends from Easter Island to Madagascar and from Hawaii and Taiwan to New Zealand. They also are spoken in New Guinea, except for the southern coastal region. There are even some Austronesian languages found in Southeast Asian mainland communities.

Most of the approximately 260 languages of Australia, although in some instances quite different from each other, are thought to be related to one another—more than 85 percent being closely related. The great majority of these were either extinct or nearing extinction at the turn of the 21st century. Common structural features strongly suggest a single original linguistic stock. Attempts to trace the Australian languages to those of other areas, even to contiguous ones, have proved fruitless; this probably indicates the great antiquity of the languages of Australia and the long isolation of the Aboriginal population (see also LANGUAGES OF THE WORLD: *Austronesian languages, Papuan languages,* and *Australian Aboriginal languages*).

Effect of European contact on population trends. The nature and number of the various island peoples at the time of the arrival of Europeans are uncertain. The period of European exploration was spread over several hundred years, beginning in the 16th century and becoming intense by the 18th and 19th. It was 1788, for example, before colonists from Europe settled in Australia. The number of indigenous peoples at that time is not known, and their widely dispersed and seminomadic hunting-and-gathering groups are estimated to have been organized into several hundred to nearly a thousand tribes. Tribal groups were generally associated with a given territory in which they eked out an existence from a relatively inhospitable land. Theirs was a simple yet highly functional material culture, which contrasted with a complex social system and a rich ceremonial life, featuring ancestor worship and totemic beliefs. They had rudimentary, impermanent shelters, went unclothed but for personal adornment, and foraged, hunted, and fought with simple but efficient implements and weapons. Some have estimated an Australian Aboriginal population of 300,000 at the time of European contact, but this is at best a guess. It is thought that the population had over time stabilized in relation to the ecology of the subcontinent and that, as a food-gathering and hunting culture, it was in balance with the environment. Like many other Oceanian peoples, the indigenous peoples of Australia suffered severe population loss as a result of contact with the Western world.

Precolonial Aboriginal culture

DIVERSITY OF PACIFIC CULTURE PATTERNS

The people and cultures of present-day Oceania differ greatly from those of the days when the areal classification into Polynesia, Melanesia, Micronesia, and Australia was devised. The Aboriginals of Australia are few in number, and many are of mixed ancestry. The Tasmanian Aboriginals fared even worse and became extinct in the late 19th century. Other Pacific peoples, such as those living in the Mariana Islands and on Easter Island, often are more Spanish and Chilean, respectively, than they are Micronesian or Polynesian. On some islands there are substantial populations of East Indian, Chinese, Japanese, or European-North American origin. With few exceptions, such as

in remote areas of New Guinea, the peoples of the Pacific Islands have been exposed intensively to Western civilization and have adapted their ways to it either by choice or by necessity. Only certain features of their traditional cultures remain. There are, in greatly varying numbers, native inhabitants of all regions who manifest the physical features commonly associated with their culture areas. In some places, however, the native population is in the minority. It is, therefore, no longer accurate to say that the people of the Hawaiian Islands possess a Polynesian culture or that those of New Caledonia are Melanesian culturally. Nevertheless, the fourfold typology of Oceanian peoples and their cultures continues to be used with reference to the contemporary inhabitants as well as to the traditional people of Oceania and their ways of life.

Social structure. Highly stratified social structures are typical of the aboriginal cultures of Polynesia and Micronesia. Hereditary chieftainship was closely related to supernatural beliefs that incorporated concepts of power (mana) and avoidance (tapu). Features of these two culture areas include an elaborate mythology, specialist craftsmen, distinctive artistic styles, pandemic and sometimes savage warfare, strong bonds of kinship and a related emphasis on genealogies, and, in places, strong trading or tributary relationships between island communities.

Mana and tapu

Melanesian societies are, because of their great diversity, less easily characterized. They were in general less concerned with social rank based on birth than with prestige gained through manipulation of resources. Root-crop agriculture (principally the cultivation of yams, sweet potatoes, and taro) was practiced as in many parts of Micronesia and Polynesia, and the resources of the sea were widely exploited. Ancestor worship and a rich ceremonial life were combined to produce a wealth of religious practices; they marked life-crisis situations and incorporated both venerative and propitiative behaviour. In many areas complicated and highly ritualistic trading relationships were developed, and there were usually distinct differences between coastal and inland societies. Personal adornment was often elaborate; warring raids, in which heads were taken and cannibalism was sometimes practiced, were common; stone tools were polished; and animism was featured in supernatural beliefs. Melanesia's domesticated animals were those found elsewhere in the Pacific region, but pigs were vastly more important in the pursuit of personal prestige.

Technological development. Most of the precontact people of the Pacific were at what is known in the West as the Neolithic stage of technological development; that is, they worked stone by grinding rather than by chipping and flaking as in the Paleolithic stage. Some areas also had various other cultural features of the Neolithic such as agriculture, animal husbandry, pottery, and watercraft.

Seashells were used generally both as ornaments and as cutting tools. Shark, snake, and lizard skins were used as heads for drums. Bird, whale, and human bone, as well as fish vertebrae, were used for ornamental purposes and for tools, implements, and weapons. Basalt, obsidian, jadeite, chalk, coral stone, and other lithic materials are irregularly distributed throughout the Pacific Islands. Where these materials were available, craftsmen produced a bounty of objects—adze blades, ornaments, images, and weapons—by skillfully chipping and grinding. Native woods of both hard and soft varieties were used to produce finely executed weapons, drums, masks, images, and building ornamentation. Great logs of pine were felled in New Zealand to produce flotillas of war vessels that could carry more than 100 men. Other islanders, without suitable timber for such canoes, scoured the beaches for driftwood logs or devised canoes from smaller trees, fastening planks together with cordage and caulking the seams with gum. Polynesian ironwood provided a source for some of the most intricately carved and heaviest of weapons. Ebony, bamboo, tree fern, and palm all added to the lumber supply of the carver. Feathers of colourful birds were woven into garments and effigies in some places. Headdresses of feathers appeared in many parts of the Pacific region, but only the Hawaiians devised beautiful feather helmets. The kiwi, the Hawaiian honeycreeper, the tropic bird, the parrot, and the cassowary are but a few of the birds that added their plumage to the artisan's kit. Fibres and barks were used abundantly, particularly orchid fibre, hibiscus, banana, flax, pandanus and ti leaves, and coconut-husk fibre. Pigments such as turmeric were obtained from plants. Earth and burned coral provided others. Pacific island craftsmen also used resins, gums, seaweed, spiderwebs, flying-fox fur, human hair, dog skin, kangaroo fur, boar tusks, vegetable seeds, shells, and teeth from humans, dogs, fish, and fruit bats (see also OCEANIC ARTS).



Pandanus growing in the ash plain at the base of Mount Yasur, an active volcano on Tanna, Vanuatu; at right, Lake Siwi.
© Gary & Robyn Wilson/Australasian Nature Transparencies

The environment in which the traditional island native lived provided the setting for his culture. If there was little suitable land for agriculture, he had only tree crops such as coconut or pandanus; yam growing could thus not become a factor in social prestige nor could tapa making become a fine art. Although environment did not absolutely determine the nature of the culture that evolved in a given location, it served as a limiting factor; the technology of a people was in direct consonance with the surrounding region. The peoples who lived on continental islands (that is, on islands geomorphologically connected with a continent) had at their disposal a variety of minerals, a diversified flora and fauna, and, usually, a good freshwater supply. By contrast, the resident of a small coral atoll was preoccupied with securing an adequate supply of potable water and had fewer species of plants and animals. The type of island, its age, composition, and elevation, as well as its location and climate, were vital factors in the development of a particular culture. Oceanian peoples had to adapt to a broad range of extremely diversified environments—tiny Micronesian atoll, humid New Guinea forest, or arid Australian desert. (R.W.Fo./Ed.)

Environmental limitations

Melanesian cultures

Melanesian is an old anthropological category that has been rendered somewhat obsolete by developments in prehistory and linguistics. It groups together the dark-skinned peoples of the southwestern Pacific and characterizes them culturally in terms of Neolithic root-crop economies, highly developed systems of exchange, generally small-scale and nonhierarchical polities, and diversity in social structure. As a culture area, Melanesia can be reexamined in the light of modern evidence.

The vast continental island of New Guinea and the islands that form an arc from its eastern end down toward the southeast—the Bismarck Archipelago, the Solomon Islands, the New Hebrides (the state of Vanuatu), and New

Caledonia—represent a meeting ground of two cultural traditions and populations. The earliest, for which the term Papuan provides the best label, is ancient: Papuans occupied the Sahul continent (now partly submerged) at least 40,000 years ago. As hunting-and-gathering peoples whose ways of life were adapted to tropical rain forest, they occupied the equatorial zone that, after sea levels rose at the end of the Pleistocene glacial period, became the vast island of New Guinea.

Modern descendants of these early populations speak languages belonging to a number of different families. They are categorized as Papuan (see LANGUAGES OF THE WORLD: *Papuan languages*).

Perhaps partly through indirect contact with developments in Southeast Asia, Papuan peoples developed one of the earliest agricultural complexes in the world (perhaps 9,000 years old, contemporaneous with the dawn of agriculture in the Middle East). Evidence indicates that they domesticated root crops and sugarcane and may have kept domestic pigs. By 5,000 years ago agricultural production in parts of the New Guinea Highlands was marked by systems of water control in agriculture and associated pig husbandry, both of which became intensified over subsequent millennia.

About 4,000 years ago seafaring peoples bearing a Southeast Asian cultural tradition must have been moving in areas north of New Guinea; by 3,500 years ago they had occupied parts of the islands of the Bismarck Archipelago. Their presence is marked by the appearance of the distinctive pottery and associated tools and ornaments of shell that define the Lapita culture. They apparently spoke a language of the Austronesian family (see LANGUAGES OF THE WORLD: *Austronesian languages*) related to languages of the Philippines and the Indonesian archipelago. This early language is labeled Proto-Oceanic: from it are descended the languages of central and eastern Micronesia and Polynesia; the languages of the Solomons, Vanuatu, and New Caledonia; and many of the languages of coastal eastern New Guinea, adjacent islands, and much of the Bismarck Archipelago.

The speakers of Proto-Oceanic, who had a maritime orientation and sophisticated seagoing technology, probably had a system of hereditary chiefs with political-religious authority and elaborate cosmogonies and complex religious systems that were not unlike those recorded in western Polynesia.

Earliest Melanesian settlement

The Bismarck Archipelago east of New Guinea was already occupied by speakers of Papuan languages (whose earliest settlement has been dated to 30,000 years ago). The dark-skinned, woolly-haired populations anthropologists have classed as Melanesian that now occupy the Bismarck Archipelago and the arcs of islands extending to the southeast (the Solomons, Vanuatu, New Caledonia) represent the mixing of cultural traditions and biological heritages of Papuan and Austronesian peoples. The mixing may have taken place largely within the zone of the Bismarcks prior to the settlement of the islands to the southeast (although the exact process and relative contributions of these historical populations is debated). A great deal of economic interchange took place between a Southeast Asian complex based on root- and tree-crop cultivation and on maritime technology and the already well-developed Papuan agricultural and technological systems. It is probable that an interchange of other cultural traditions, from social organization to religion, took place as well. Some Austronesian-speaking communities—perhaps ones that retained their maritime orientation—appear to have remained relatively isolated from intermarriage and cultural interchange.

Linguistically, in the interchange between Papuan and Austronesian peoples the latter were clearly dominant. Almost all languages spoken by dark-skinned peoples in the Pacific east of the Bismarcks are classed as Oceanic Austronesian, although some (especially those of the eastern tip of New Guinea and adjoining islands, but perhaps also those of the Santa Cruz Islands, southern Vanuatu, New Caledonia, and the Loyalty Islands) reflect extensive Papuan borrowings. The languages spoken by dark-skinned peoples of the southeastern Solomons and northern and central Vanuatu are closely related to the languages of Polynesia and central and eastern Micronesia.

The zone east of New Guinea can be characterized as island, or seaboard, Melanesia, to contrast it with the Papuan-speaking zone of the continental island of New Guinea itself. "Melanesian" has sometimes been used in a narrow sense to label the dark-skinned peoples speaking Austronesian languages, in contrast to "Papuans." More commonly, "Melanesian" has been used in a more inclusive sense, to label both.

Little evidence has been found to identify the earliest settlers of the zone of island Melanesia south and east of the Bismarcks. It seems likely that the Solomons chain was settled by Papuan-speaking populations following the early occupation of the Bismarcks. The islands to the southeast of the Solomons—the Santa Cruz group, Banks Islands, New Hebrides, New Caledonia, and the Loyalties—are separated by much larger gaps of open sea, and they may well have been unoccupied prior to the arrival of Austronesian-speaking peoples in the Bismarcks. The pioneer settlers of large zones of island Melanesia may have been speakers of Austronesian languages who represented the genetic and cultural fusion described above, and others, contemporaneous with them, who had remained relatively isolated from intermarriage and cultural interchange with Papuan speakers. In the period from 3,500 to 3,000 years ago the latter established coastal communities and associated trade systems in the southeastern Solomons, New Hebrides, New Caledonia, and Fiji. Long-distance trade, particularly of shell ornaments and obsidian, suggests that these widely spread communities characterized by the Lapita pottery tradition were linked politically. The settlement of eastern Micronesia by speakers of Oceanic languages, perhaps from the Solomons, apparently took place in this period. Fiji apparently was initially colonized by Lapita-making peoples and later was settled by dark-skinned, culturally Melanesian peoples after Fiji had been a springboard to the settlement of western Polynesia. The spread of this Lapita tradition from Fiji into Samoa and Tonga fits closely with the evidence of the initial breakup of a linguistically reconstructed Proto-Polynesian.

Long-distance trade

Although the prehistory of most of island Melanesia remains little-documented, some evidence suggests that the cultural, linguistic, and political fragmentation that prevailed at the time of European penetration, with a half-dozen languages and cultures often represented on a single island, is partly a product of devolution in the past 2,000 years. More hierarchical regional political systems and concomitant trade systems seem to have fragmented and devolved, with progressive involution, the replacement of trade by exchange, the disappearance of political hierarchy (except in some coastal zones), and the separation of language or dialect groups as results. The causes of these transformations, perhaps including internecine warfare and the spread of diseases (of malaria in particular), remain unclear. These changes were accelerated by European penetration, which further disrupted trade systems, intensified intercommunity warfare by supplying firearms, thinned populations by introducing diseases and indentured labour, and eroded traditional authority systems. The island Melanesia described by anthropologists reflects both long-term devolution and recent disruption.

Although the mix of Austronesian and Papuan cultural elements varies in different parts of island Melanesia, in many ways the classification of these peoples and their cultures along with Papuan-speakers as Melanesians—in contrast to the Micronesians and Polynesians—does violence to the ethnological, linguistic, and archaeological evidence. The Melanesians of northern Vanuatu and the southeastern Solomons speak languages very closely related to those of Polynesia and eastern Micronesia. Culturally, they are in many ways more closely related to these other Austronesian-speaking peoples than to the Papuans of interior New Guinea. The Melanesians' religious systems, for example, incorporate such concepts as mana and, in the Solomons, tapu.

A further complexity is that the region has emerged not only into the modern world but also into the community of nations: Papua New Guinea, Solomon Islands, Van-

uatu, and Fiji constitute separate nation-states; western New Guinea (Irian Jaya) is a province of Indonesia; and the indigenous peoples of New Caledonia and the Loyalty Islands are locked in a struggle for independence with French settlers and the French government.

SOCIAL AND POLITICAL LIFE

Settlement patterns. In many areas, local groups were not concentrated in villages but were dispersed through territories in scattered homesteads and hamlets. In many cases these were occupied for short periods until the groups moved on to follow cultivation cycles. Where communities were in danger of surprise attack, they tended to cluster more closely; in interior areas they were usually sited on ridges and peaks.

In parts of the Sepik plains of New Guinea, huge villages (some with populations of more than 1,000 people) represented the aggregation of descent-based local groups. In the agricultural heartland of northern Kiriwina, in the Trobriands (Massim), villages of up to 200 people were arrayed around a central dance ground; villages at least as large were packed together on coral platforms in the lagoons of northern Malaita (Solomons). In general, larger villages settled over a span of generations were characteristic of coastal environments, and smaller, shifting settlements were characteristic of interior areas.

Residential separation of the sexes was common. Men's clubhouses, a focus of ritual and military solidarity, were common in many areas of New Guinea and island Melanesia. The men's huge cult-houses of the Sepik River basin and on the southern Papuan coast are examples. In the montane, sparsely settled interior zone of New Guinea, men's longhouses were built as defenses against the threat of raiding and as centres of cult activities. Women and children occupied domestic dwellings; in some regions women were isolated during menstruation and childbirth.

Gender relations. In many parts of New Guinea and island Melanesia, male-female relationships were polarized. In New Guinea a zone of extreme polarization extended from the Papuan coast (Marind-anim, Asmat) along the southern face of the Highlands (Anga speakers, Papuan plateau) and the high central mountains (Mountain Ok peoples) down into the Sepik. Men's cult secrecy, ritualized male homosexuality, elaborate men's initiation rituals, and the celebration of warfare were accompanied by belief in the dangers emanating from women's bodies and by separation of the sexes. Peoples throughout this zone were preoccupied with growth and with the physical fluids and substances (menstrual blood, semen, vaginal fluids) that are agents of reproduction and growth, all regarded as sources of power and danger.

Extreme sexual polarization

These concerns were echoed in the eastern Highlands with a physical separation of the sexes, men's initiations and cult rituals, the use of sacred flutes, ritualized nose- or penis-bleeding (ostensibly in imitation of menstruation), and accompanying myths of an ancient female power that fell into the hands of men. In the Sepik basin complexes of what has been called "pseudo-procreative" ritual accompanied male cult activities.

In the central and western Highlands, where populations were dense and sweet potato and pig production were intensive, the risk of pollution by the female body and the array of male ritual activities were subordinated to the politics of extracting female labour, acquiring prestige and power through exchange, and mobilizing armed strength. Although cosmological beliefs reinforced the separation of male and female realms, women were actually a danger more as potential spies from enemy groups than as sources of polluting substances and essences.

In the Massim, in contrast, the powers of women, reproductive and productive, were represented both in the realities of social relations and in ideologies of descent and cosmic processes. In the Trobriands, women played prominent parts in some rituals. Matrilineages (in much of the Massim called *susu*, meaning "breast" or "breast milk") provided symbols of cosmic reproduction as well as physical and social reproduction.

While some polarization of the sexes emerges in accounts from the Solomons and Vanuatu, sexual segregation and concerns with pollution by females seem to have more to do with the preservation of symbolic boundaries than the inherent dangers of women's bodies and essences. In everyday domestic politics, women had relatively high status in many island Melanesian communities. Their complementary but central role was widely given cultural recognition. In many areas both paternal and maternal links provided bases of ritual status or local group affiliation, and women were accorded importance in ritual and as healers, elders, and ancestors.

Kinship and local groups. The societies of precolonial New Guinea and island Melanesia were characteristically organized in local groups based on kinship and descent and linked together by intermarriage. In the usual absence of centralized political institutions, such local groups were relatively autonomous political units. In most areas the local groups were relatively small, with between 20 and 100 members; in densely settled areas of the New Guinea Highlands, and parts of the Sepik plains, kinship- and descent-based local groups were considerably larger.

Land-holding

Although domestic groups or individuals held rights over gardens and cultivated trees, in transmission, alienability, and the distribution of rights, the dominant pattern was the holding of corporate title to land by a local group. That is, title to land was inherited and held collectively by (some of) the descendants of those who initially cleared it. Use rights might then be extended to others. In patrilineal systems, for example, rights might be given to individuals related to the land through mothers or grandmothers, or to outsiders. In coastal zones, corporate title might be held to reefs or fishing grounds.

In many areas the relationship between people and land was conceptualized in terms of chains of descent—in male line (patrilineal descent), female line (matrilineal descent), or some combination of male and female links (cognatic descent)—from founding ancestors. Patrilineal descent systems prevail in most of lowland New Guinea; matrilineal descent systems, in much of the Massim and the Bismarck Archipelago and in much of the Solomons and the Banks Islands. New Caledonia and northern Vanuatu have primarily patrilineal descent systems. Considerable variation is found within these areas, however.

Societies of the central and western Highlands of New Guinea have been described as segmentary patrilineal descent systems. While some of them emphasize chains of descent connecting the living to ancestors, most seem unconcerned with such connections. They use father–son links as the main mechanism of group recruitment but are open to the attachment of refugees and individuals connected through women. The segmentary structures (of phratries, or clans above the level of local groups) characteristically use brother–brother and father–son links to represent what were once in fact relatively unstable political alliances. With intergroup warfare, the sheer size of local groups was a key to survival, and groups crystallized and dispersed according to the tides of warfare.

In patrilineal (and Highlands filiation-based) systems, political leadership within local groups, expressed in warfare and exchange, provided an arena for competition among men. Where matrilineal descent systems had developed, arenas for men's assertion of and competition for leadership and status parallel to the descent-based local groups tended to develop: examples are the secret societies of New Britain and the Banks Islands and the overseas exchange systems of the Massim.

Matrilineal descent patterns

The matrilineally organized societies of island Melanesia represent several variations on the themes of descent and residence. In parts of the Bismarck Archipelago, matrilineal moieties were localized on the basis of uxorilocal postmarital residence (that is, residence with the wife's kin after marriage), with communities containing members of each moiety. In parts of the Massim, matrilineal clans or subclans were at least partly localized on the basis of avunculocal postmarital residence: after childhood in his father's settlement, a boy moved to his mother's brother's village, where he later lived with his wife. In other parts of the Massim, notably Dobu, postmarital residence alternated between the husband's matrilineage village and the wife's matrilineage village. In other areas (such as

parts of the Solomons), postmarital residence lay with the husband's kin, so that matrilineages were only indirectly localized by virtue of coresidence of several lineages in the same village. (Such a juxtaposition of two or more matrilineages or matriclans in the same village, or contiguous villages, is common in matrilineal descent systems in Melanesia; in patrilineal descent systems, local communities more commonly comprised members of a single descent group and in-marrying wives.)

Ties of intermarriage were important in creating and maintaining connections between descent- or kinship-based local groups. Marriages negotiated with enemies made at least temporary allies of them. Particularly where marriage entailed a transfer to her husband's group of rights to a wife's fertility as well as to her labour, these transfers of rights were validated by bridewealth in the form of valuables or pigs or both. (The custom is resilient in much of modern Melanesia despite Christianity and capitalist economic relationships.) Polygyny was relatively widespread, at least for prominent leaders, who could use the productive and reproductive labour to accumulate surplus pigs and root crops. In some areas, as in the Trobriands, polygynous marriages of high-ranking leaders were instruments of political alliance and of tributary relationships between descent-based local groups. (In the Trobriands, because a matrilineal subclan was obligated to cultivate and ceremonially present yams for the husbands of its absent female members, a leader with many wives became a centre of yam distribution.) Where polygyny was not practiced, leaders could use the financing of the bridewealth of followers to create a labour force.

The stability of marriage bonds varied widely from society to society. The dissolution of marriages was constrained, particularly in patrilineally structured societies, by rights acquired through bridewealth. Marriage tended to be less stable in some matrilineally organized societies.

"Complementary filiation"

Ties of kinship created through marriage alliances between local groups crosscut and were complementary to divisions based on unilineal descent. In the patrilineally organized societies of lowland New Guinea and island Melanesia, a person's connection to the mother's group and ancestors was often recognized in acts of kinship support, in ritual, and in the parts played by the groups in marriages, mortuary ceremonies, and other exchanges. In the matrilineally organized societies of island Melanesia, ties to the father's matrilineal relatives were similarly expressed. The complementary parts played by maternal and paternal subclans in Trobriand (Massim) mortuary rites were particularly complex. In some areas, secondary land rights were extended on the basis of such "complementary filiation" (so that a person could acquire interests in, or cultivate, land of the father's matrilineage or mother's patrilineage). Such secondary ties could be the basis for succession to corporate title when groups died out, dwindled, or were dispersed by warfare.

Throughout Melanesia, obligations toward kin constituted the ultimate moral imperative. Systems of exchange were constructed out of kinship obligation. Rights deriving from birth commonly had to be validated by prestations or the fulfillment of obligations. One case was the payment of a midwife by a father in matrilineally organized Mota (Banks Islands): unless he paid the fee, and thus validated his status as father, the person who paid in his stead acquired rights over the child.

Warfare and feuding. Large-scale armed confrontations between warriors were common in parts of New Guinea and some parts of island Melanesia. Ethnographers have often stressed the pageantry and ritual posturing involved and have inferred that few lives were lost; these accounts also often emphasize that territory was not gained or lost. But further evidence from the New Guinea Highlands and other parts of the island suggests that warfare, or in some areas clandestine raiding, had a high cost of human life (among the Mae Enga of the western Highlands, as many as 15 percent of male deaths occurred in war, and even today a resumption of armed combat takes a substantial toll of life). Moreover, it is now clear that victorious groups often did displace their enemies from territories, as in many parts of New Guinea (such as the Sepik and the southern and eastern Highlands). Homicidal raiding of neighbouring peoples was widespread and was associated with head-hunting in such regions as the southern New Guinea coast (Asmat, Marind-anim) and the western Solomons (Roviana, Vella Lavella) and with cannibalism in others (southern Massim, New Caledonia). The cultural orientations of many Papuan and island Melanesian peoples were shaped by a warrior ethic—an ethos of bravery and violence, vengeance and honour—and by the religious imperatives that promoted aggression.

Endemic blood feuding

In other areas, such as Malaita in the Solomons and much of Vanuatu, large-scale combat was rare, but blood feuding was endemic. Where deaths were attributed to sorcery, a common pattern in both Papuan- and Austronesian-speaking areas, accusations of sorcery commonly triggered vengeance murders. So, in some areas, did seduction, adultery, theft, or ritual insult. Chains of blood feuding led neighbouring villages or clans into vengeance killings; feuding was laid to rest only when murders were finally balanced out or death compensation was paid.

Political leadership. Both in Papuan-speaking New Guinea and Austronesian-speaking island Melanesia, leaders of local groups characteristically emerged on the basis of success in prestige economy, warfare, or both. This mode of leadership, based on status achieved through entrepreneurial success and the influence and obligations attendant to it, has been elevated in anthropological stereotypes of the "Big Man" as prototypical Melanesian leader.

Problems of anthropological stereotype

The stereotype needs critical examination. With warfare and feuding endemic in much of precolonial Melanesia, the importance of warrior leaders in many societies had waned or been eclipsed by leadership in competitive feasting and other exchange by the time anthropologists arrived. A second problem in the stereotype, at least for Austronesian-speaking Melanesia, is that it seems likely that early Oceanic-speaking peoples had hereditary chiefs. Thus, the emergence of entrepreneurial (or warrior) leaders to whom others deferred by virtue of their deeds, not their rank, may well have been a concomitant of the process of political devolution.

The most serious problem in the Big Man stereotype is that in many parts of island Melanesia, at the time of first European penetration, societies were led by hereditary chiefs. This was true of parts of Austronesian-speaking coastal New Guinea (*e.g.,* Mekeo, Motu), parts of the Solomons (*e.g.,* Buka, Shortlands, Small Malaita), parts of Vanuatu (Aneityum), and most of New Caledonia. In other areas leadership was based on conceptions of rank but with succession to leadership based on a complex relation between hereditary right and demonstrated ability (indeed, such a relation was important in much of Polynesia as well). Bronisław Malinowski's "Paramount Chief" in northern Kiriwina, in the Trobriands, has become recognized as a hybrid figure, not quite a chief in a classical sense but of an exalted rank that set him apart from the *mumi* of the Siuai of Bougainville or the *ngwane-inoto* of northern Malaita, on whom the Big Man stereotype was modeled. In much of precolonial island Melanesia, particularly in coastal areas, leaders seem to have fallen into this middle zone, with rank and achievement operating together to confer leadership and each reinforcing the other. The power of such local leaders depended partly on monopolies over trade or prestige exchange systems, or regional domination based on war, both of which were disrupted by European penetration. The stereotypical image of the Big Man is partly a result of the fact that ethnographic accounts were gathered in the colonial period and by observers operating in a tradition that discounted and disguised historical processes.

ECONOMIC SYSTEMS

Production and technology. The ancient root-crop cultivation systems of Papuan and Austronesian cultural traditions depended on swidden horticulture, a practice of shifting cultivation whereby rain-forest gardens are cleared, planted, harvested, and then left fallow for periods of up to a generation. Ground stone tools (and in some coral-island areas, shell tools) and fire were used to clear forest; digging sticks were used for cultivation.

The primary plant domesticates were taro and yams (*Dioscorea*), with plantains, sago, pandanus, leafy greens such as *Hibiscus manihot,* and sugarcane being other important domesticates. In swampy areas of New Guinea, sago production continues to support dense populations. Sweet potatoes, a New World domesticate that reached New Guinea through the Moluccas no earlier than the 16th century, allowed an intensification of production and dense settlement of higher altitudes (where *Pueraria*—a root vegetable—and taro had been cultivated earlier). The progressive expansion of intensified cultivation in the great highland valleys of New Guinea, accelerated in recent centuries, has resulted in the progressive transformation of montane forest hunting territories (also used for foraging and swidden cultivation) into *Imperata* grass, further accelerating reliance on pig husbandry and intensified production, particularly of sweet potatoes.

Results of intensified cultivation

In island Melanesia either yams or taro typically has been the major staple crop. The two have quite different productive regimens (as well as symbolic meanings), and focus on one or the other (which often is difficult to account for on ecological grounds) tends to structure community life. Yams are planted and harvested seasonally, and the tubers, if unblemished and dry, keep for several months. Hence, communities that focus on yam production tend to have an annual cycle and to emphasize communal labour and common enterprise. Taro corms rot quickly after harvest, and taro has no seasonal cycle; the shoots are replanted after the corms have been cut off, so that both harvest and planting are continuous throughout the year. Cultivation by individual families, and more flexible local group structures, are promoted by the regimen of taro production.

In New Guinea and parts of island Melanesia agriculture was intensified by water control, mounding and composting, or terracing. Terraced and irrigated taro fields were found in parts of the New Hebrides and New Caledonia; the mounded and composted gardens of parts of highland New Guinea, associated with water control, allowed intensification of production and continuous use of land.

Root- and tree-crop production was augmented by the raising of domestic pigs, fishing (on coasts or in interior rivers, lakes, and streams), and hunting of marsupials and birds (and, sometimes, gathering of insects and grubs). Gathering of wild vegetable foods, including tubers, greens, nuts (notably the canarium almond, which was extremely important in areas such as northern Malaita in the Solomons), and fruits, augmented diets or provided emergency rations. Food crops were complemented by a range of cultigens used for purposes other than food. The areca nut and accompanying betel pepper, for example, were widely chewed as a stimulant and were a crucial medium of sociality in large zones of Melanesia; kava (*Piper methysticum*) served a similar purpose in parts of Vanuatu and in Fiji; and plants such as ginger and *Cordyline fruticosa* were used in ritual, magic, and medicine.

Although swidden horticulture typified the region, many peoples of montane New Guinea relied heavily on hunting and gathering and had low-intensity food production systems. Competition for hunting territories was apparently a major factor in warfare and raiding, particularly among the peoples of the ecologically marginal zones of the southern regions of the Highlands.

Exploitation of the coastal zone

In island Melanesia, coastal zones offered rich environments for the exploitation of fish, shellfish, and sea turtles. As sources of shell used for valuables and ornaments, of salt and lime, and of marine food products, the coastal zones contributed goods that coastal dwellers could trade and barter with peoples of the interior. In some areas, trade of root crops for marine products was institutionalized. In northern Malaita, in the Solomons, the coastal dwellers, who lived on islets and coral platforms dredged from the lagoon floor, bartered fish for root vegetables at regular markets. Other Solomons coastal communities specialized in manufacturing and exporting shell beads, which were widely used as valuables. Similar arrangements occurred in the Admiralty Islands and other areas of island Melanesia.

In addition to manufacturing ground stone adzes and axes, peoples of New Guinea and the islands to the east commanded a broad range of Neolithic technology. Bags and nets for fishing, hunting, and trapping were made from bush fibres, and bark cloth was widely manufactured. Giant bamboo served a multitude of purposes, providing cooking vessels, water containers, torches, and carving knives. Canoes, ranging from small dugouts to elaborate composite seagoing canoes, were manufactured. Weapons included bows and arrows, spears, and clubs.

Although this technology was virtually universally distributed through the zone, particular peoples specialized in particular forms of production. Specialization in part followed the distribution of resources, such as clay for pottery, chert or greenstone for tool blades, or trees for canoes or weapons. But systems of trade and regional exchange seem to have depended on political and cultural imperatives as well as resources.

Trade and exchange systems. The regional trading systems of the islands around the eastern end of New Guinea were particularly elaborate. In the Massim—what is now Milne Bay province of Papua New Guinea (taking in the D'Entrecasteaux Islands, the Louisiades, and nearby islands)—the trade of pottery from the Amphletts, canoe timber and greenstone blades from Murua (Woodlark), carved platters and canoe prow boards, and other specialized products was complemented by a flow of yams and pigs from resource-rich areas to smaller, ecologically less-favoured islands. Some islanders, such as those from Tubetube in the southern Massim, did little producing and specialized instead as middlemen and traders. Similar interdependencies and specializations, with some communities acting as export traders and middlemen, occurred in the Vitiaz Straits between New Guinea and New Britain. Through chains of intermediary trading partnerships among neighbouring peoples, trade systems in New Guinea connected communities that were separated by scores of miles of rugged mountains. Such networks carried salt, shell, and other objects from coasts to interiors and forest products, such as blackpalm, from interiors to coasts. They connected specialist communities, such as those extracting salt, collecting ochre, or making blades, to communities far away.

Trade networks

Both Papuan-speaking and Austronesian-speaking zones of Melanesia had highly elaborated exchange systems, in which surpluses of pigs and root crops were used in prestige feasts and ceremonial valuables (usually shell beads or other shell objects, but also including dolphin and dog teeth and a range of other material objects) were exchanged. Elaborations of ceremonial exchange, in mortuary feasts and homicide-compensation payments, bridewealth presentations, and various forms of competitive feasting, were foci of community production, social cooperation, status rivalry, and political conflict. In some areas, at least, competitive exchanges were a surrogate for warfare (and in some instances they seem to have grown out of homicide compensation).

In New Guinea the most highly developed exchange systems were those of the western Highlands: the Enga *tee,* the Hagen *moka,* and other mass exchanges of pigs and shell. Cycles of pig production were orchestrated so that vast surpluses of sweet potatoes were required to feed expanded herds; mass pig kills, or the presentation of live pigs to the leaders of rival clans, were a focus of political rivalry and community productive labour.

In island Melanesia the best-known prestige exchange system was the kula of the Massim, documented by Malinowski. As he described it, the kula entailed the endless circulation of valuables through a network of communities on islands forming a giant ring. Through the network passed intrinsically useless valuables: shell-bead necklaces passed around the ring in a clockwise direction and were exchanged for arm shells, which moved counterclockwise. The exchanges were made between partners in neighbouring communities or, in the most exciting, dangerous, and prestigious exchanges, between neighbouring islands. Research in different parts of the kula zone has shown the exchange to have been considerably more complex than Malinowski was able to see from his vantage point in the Trobriands (as well as highly resilient in the face of Westernization and economic change). Early European

penetration did much to change the nature of exchanges and to stabilize a system that was highly unstable because of intercommunity warfare.

Complex systems of prestige feasting, often with a strong competitive element, have been described for many parts of island Melanesia, including Goodenough Island (in the Massim) and the Solomons (Bougainville, Malaita). In precolonial northern New Hebrides, status rivalry was played out through complex graded societies in which men moved to progressively higher grades by sponsoring feasts and presenting valuables and pigs.

Although strung shell beads and other valuables were best known as exchange tokens in the prestige economy, in some parts of Melanesia they served very much like money—as media of economic value and exchange. They were used to buy and sell pigs, fish, craft products, and even land. The *tambu* of the Tolai of New Britain are well known, and similar valuables were used on Malaita and in some other areas as generalized instruments of value, with standard "denominations" and standard "prices" for commodities.

RELIGION

The past in the present

In Austronesian-speaking Melanesia a strong orientation to ancestors and the past was common, but it was a past manifested in the present, with ancestral ghosts and other spirit beings as everyday participants in social life. (While beliefs in gods and powerful spirits were held in the region, such beings figured more in myths than in everyday life and practical religion.)

Human effort in the uncertain projects of war, gardening, and the pursuit of prestige was thought to succeed only when complemented by support from invisible beings and forces, which were manipulated by magical formulas and elicited through prayer and sacrifice. The presence and effects of invisible ghosts and spirits were manifested in dream, revealed in divination, and inferred from human success or failure, prosperity or disaster, health or death. In such a world, religion was not a separate sphere of the transcendental but a part of everyday life; and religion and magic were not clearly distinguishable. The most sacred rituals often entailed the performance of magic, and performance of magic for personal, manipulative ends might be accompanied by prayer or even sacrifice. Although ghosts and spirit beings were not usually moral agents in their intervention in human life, there were some well-documented exceptions (such as Manus, in the Admiralties). Spirit beings were viewed in parts of island Melanesia as policing the punctilious observance of rules regarding pollution or exogamy and the correct performance of ritual. The concepts of tapu (sacred, forbidden, off-limits) and mana (efficacy or potency), well known for Polynesia, are fairly widely distributed in Oceanic Austronesian-speaking island Melanesia: mana is found in many Solomons languages and some languages of northern Vanuatu, and tapu or tabu is found widely in the Solomons and the Austronesian languages of the New Guinea area.

Island Melanesian societies lacked full-time religious specialists, so those who acted as priests or officiants or as community magicians, intermediating with ghosts and spirits, were in everyday life indistinguishable from their fellows. Some forms of everyday magic—for gardening, fishing, attracting valuables or lovers—were widely known, although knowledge of magic often constituted a form of property. Other forms, for powers of fighting or theft, tended to be closely guarded, and malevolent magic for destructive ends was secretly held and generally used in clandestine fashion (although in a few cases, as among the Mekeo of coastal New Guinea, sorcerers acted as agents of chiefs). In many Melanesian societies sorcery was seen as the major cause of death or illness. Belief in witchcraft occurred in some areas.

Diversity in New Guinean religions

Religions of Papuan-speaking New Guinea remain diverse. In much of New Guinea, ghosts and other spirit beings are much less important participants in everyday life than among Austronesian speakers, although there are many exceptions. Thus, the vast *tambaran* houses of the Sepik, with their elaborate rituals and developed beliefs in spirit beings (as described for such peoples as Abelam and Plains Arapesh) contrast sharply with the relatively secular pragmatism of much of the Highlands. Among such montane peoples as Telefomin, Bimin Kuskusmin, and Baktamin, highly complex male initiatory cults progressively reveal cosmic secrets to initiates. In New Guinea as in island Melanesia, fear of sorcery is widespread; and, among such peoples as the Fore of the Highlands, accusations of sorcery are a major cause of hostility between groups and of blood feuding. Some highland peoples, such as Chimbu, Kuma, and Hewa, believe that witches—humans acting in the grip of forces or agencies beyond their conscious control—prey on the living, taking possession of them or draining their bodily substances.

In some areas, such as the Sepik and the lagoons of northern Malaita in the Solomons, where traditional religious systems and associated cult practices remain, the old religions are giving way under the combined pressures of Christianity and capitalist development. A striking phenomenon of the early colonial period was the emergence of cargo cults in coastal New Guinea and island Melanesia. The movements, such as the Vailala Madness of the Gulf Division and the cargo cults of the Rai coast, were based on the revelations of local prophets that the ancestors were withholding European material goods from indigenous peoples. Cult doctrines included the iconoclastic destruction of old ceremonial objects and the moral, social, and logistical preparation for the arrival of vast quantities of Western "cargo," expected to be delivered by ship or plane. Cargo cults were widespread in New Guinea, the Bismarcks, and parts of the Solomons and New Hebrides (the John Frum Movement of Tanna is well known). Some of the movements were highly political and explicitly anticolonial in character, with a spectrum connecting millenarian movements, such as the Vailala Madness, at one end and political movements with mystical overtones, such as the postwar Maasina Rule movement in the Solomons, at the other.

ART

Melanesian art is highly varied. In much of highland New Guinea the body itself becomes a focus for art, with face and body painting, wigs and headdresses, and elaborate costumes. In lowland New Guinea, ebullient art traditions like the painting and carvings of such Sepik peoples as Iatmul and Abelam have become widely known. The curvilinear art of the Massim, of which Trobriand prow boards and dancing shields are examples, has also attracted interest. The *malanggan* carvings of New Ireland are equally spectacular and well-known and are relatively well-documented anthropologically. The latter, in contrast to the Sepik and Massim carvings, are ephemeral art; the fretwork *malanggan*, like some of the fern bole carvings of Vanuatu, were created for ceremonies and abandoned or destroyed afterward.

The art of the Massim

In some cultural traditions of Melanesia, carvings and other art forms had strong religious significance. Masks were a focus of creativity in several regions; they were often used in elaborate ceremonials, with masked figures impersonating mythical beings or perpetuating or dramatizing cult secrets. Many peoples, however, as in the Solomons, the Huon Gulf, and the Admiralties, decorated virtually every object not immediately discarded, however utilitarian.

Music and dance have been less well-documented, partly because (until the era of tape recording and film) they were less easy to preserve than material objects, and partly because Christianization of much of the area led to the early abandonment of many forms of music and dance. Although it is difficult to generalize, dancing often was focused primarily on displaying the bodies and costumes of the dancers (and sometimes the collective strength of the group they represented). Complex dance forms have been recorded from some areas, however. Music ranges from dirges at wakes and love songs to highly complex forms such as the polyphonic panpipe music, with as many as eight contrapuntal voices, played by orchestras on Malaita in the Solomons. Also characteristic of the region are the genres of epic narrative, myth, folktale, and oratory, redolent with metaphor and mythic allusion. These traditions,

too, have often been lost because of the incursions of Christianity and the unavailability, until modern times, of effective recording devices.

EFFECTS OF EUROPEAN CONTACT

The indigenous peoples of much of the Melanesian region are now part of the world economic system and are subjected to pressures of Christianization and Westernization. In some areas such forces have operated for more than a century. In some interior areas, however, particularly in New Guinea, Western penetration came much later—in the 1930s or even after. But today the most remote regions have become accessible, and they have been transformed.

Papua New Guinea, Solomon Islands, and Vanuatu are now sovereign states and members of the United Nations; the indigenous Kanak peoples of New Caledonia have become a minority in a French overseas territory, battling for independence. The indigenous peoples of western New Guinea (Irian Jaya), since incorporation into Indonesia, have been subjected to massive disruption, dislocation, political repression, and the forced accommodation of large settler populations under national transmigration policies.

With Melanesians now serving as diplomats, businessmen, bishops, doctors, lawyers, and professors, many generalizations about Melanesia as a region marked by "primitive societies" have become anachronistic. Some general observations still seem possible, however. One is that classless societies have become class-stratified societies, with politicians, public servants, and entrepreneurs constituting an emerging elite. Moreover, at least in the English-speaking areas, the elites increasingly share a common (Westernized and consumerist) culture and common political and economic interests and ideologies that cut across boundaries not only between cultures and language groups but also between nations.

Newly class-stratified societies

Internally, the countries of modern Melanesia show increasing polarization between metropolitan centres and village hinterlands. Squatter settlements on urban peripheries and movement into towns are increasingly found, and both serve as links between the villages and urban life. The more remote villages are poor and have little access to the educational, medical, and economic services of the state. It is in the marginal areas that the traditional culture tends to be the most resilient.

Cash cropping and various modes of capitalist enterprise, as well as dependence on imported goods, have penetrated ever farther into the Melanesian village hinterlands, with some areas (notably parts of highland New Guinea) attaining a measure of prosperity by Western standards through the production of coffee or other high-value crops. Roads and airfields now connect once-isolated hinterlands to regional networks.

Among the new elite, cultural nationalist ideologies have tended to focus on "custom" and "the Melanesian way"; cultural revivalism has become a prominent theme. Art festivals, cultural centres, and ideologies of *kastom* have cast in a more positive light the traditional cultural elements, such as ceremonial exchange, dance and music, and oral traditions, that had long been suppressed by the more conservative and evangelistic forms of Christianity. The emphasis on traditional cultures as sources of identity has been expressed in the perpetuation or revival of old genres of exchange. In Papua New Guinea the kula exchange of arm shells and necklaces continues in the Massim, carried on through the medium of air travel and among politicians, professionals, and public servants as well as by villagers in canoes. Members of the new elite conspicuously pay bridewealth in shell valuables.

Christianity has been a powerful force of change within the region for many decades. Much of island Melanesia and coastal New Guinea has been Christianized for two or three generations or longer. In the colonial period, missions were a major force of education and local economic change. Many of the leaders of postcolonial Melanesia have come from mission schools and backgrounds, and some have been trained as ministers or evangelists. In terms of their orientations and their leaders, the Melanesian postcolonial states are among the most Christian nations on earth.

Different Christian denominations, and even individual missionaries, have in varying degrees been sympathetic to and knowledgeable about local languages and cultures. Missionary work, in conjunction with the imposition of colonial rule, has helped Melanesians to transcend and to liberate themselves from customary practices that exacted a heavy cost in violence, cruelty, and exploitation; but it has also destroyed much cultural richness, and the process continues in remote regions of Papua New Guinea.

(R.M.Ke.)

Micronesian cultures

Seven major high-island cultures can be distinguished in Micronesia: those of the Chamorros, most of whom live on the four southern islands of the Marianas; the Palauans; the Yapese; the Trukese, inhabiting about 12 high islands of varying size in the large Truk Lagoon; the Pohnpeians; the Kosraeans; and some of the inhabitants of the isolated island of Nauru, which is geologically a raised atoll (without exposed volcanic rock).

Diversity of the Micronesian peoples

The inhabitants of most of the low islands or atolls in Micronesia are culturally distinct from the high islanders, though in contact with them. In the east are found two culturally distinctive groups of atolls, the Marshalls and the Gilberts (politically constituted as the Republic of the Marshall Islands and the Republic of Kiribati, respectively), ranging from the northwest to the southeast over about 1,400 miles. The culture of Banaba (formerly Ocean Island), a raised atoll, is quite similar to that of the Gilberts and will be treated herein as a variant of Gilbertese culture. Three atolls within sailing distance of Pohnpei—Mokil, Pingelap, and Ngatik—show closer cultural relationships to the people of Pohnpei than to any other large population but are clearly distinct from them. The Hall Islands, atolls to the north of Truk, and the Mortlock (Nomoi) Islands, atolls to the south, are culturally closest to Truk. The remaining low islands to the west of Truk also show linguistic and cultural relationships to Truk, with the differences becoming more and more marked as distance increases. The low islands between Namonuito and Yap were once part of a ceremonial exchange system called the Yapese Empire by anthropologists. Linguistically and culturally, however, these low-island people were closer to the Trukese than to the Yapese.

The cultural diversity of Micronesia is indicated by the number of mutually unintelligible languages in the area. The following all have mutually unintelligible languages: Nauru, the Gilberts, the Marshalls, Kosrae, Pingelap, Mokil, Pohnpei, Ngatik, the Mortlocks, Truk, the Puluwat area, the Woleai area, Yap, the Palaus, and the Marianas. The languages of the islands from the Marshalls and the Gilberts through the Woleai area in the above list appear to be somewhat more closely related to each other than to the remaining languages and have been referred to as Nuclear Micronesian. The languages of the Palaus, Yap, and the Marianas are relatively distinct from each other and from other Micronesian languages, although they are clearly Austronesian in their general affiliation. The Chamorro language, spoken in the Marianas, has undergone much influence from Spanish and probably also from the Philippine Tagalog language after more than four centuries of Western contact.

Traditional Micronesian cultures were characterized by a belief in the stability of society and culture. People suffered occasional natural disasters, such as typhoons or droughts, but their goal after encountering one of these was to reconstitute the previous state of affairs. Wars occurred in most areas from time to time, mainly at the instigation of competing chiefs. At stake was the control of land (a very limited resource in the tiny islands of Micronesia) and followers, but there were usually few casualties.

Living in small communities on small territories, Micronesians learned to adjust to their neighbours and remain on good terms with most of them most of the time and to develop techniques of reconciliation when fights did break out. In general, Micronesians valued good manners. Even today, Micronesian adults and children are likely to impress the outsider as unusually quiet, polite people.

Micronesians traditionally depended on the cultivation of

plant crops and on fishing in shallow reef waters. Because both arable land and reef were in short supply for the relatively dense population, the Micronesian had a strong practical basis for his attachment to his locality and lands. Land rights were usually held through some sort of lineage or extended family group, often backed up by traditions of ancestral origins on the land.

Local loyalties

The strong local loyalties of the Micronesians may also be partly explained by the difficulty of traveling to any place very far from home, especially for the high islanders who lacked oceangoing canoes. Of the high-island peoples, only the Yapese practiced very much navigation in the open seas at the time of their discovery by Europeans, sailing to Palau and to some of the atolls in the central Carolines. The remaining high-island peoples mostly sailed around their home islands without losing sight of land, although they were visited from time to time by low islanders. The low islanders had reason to visit the high islands, with their more fertile soil and greater elevation; they could obtain food and other items not found on their home islands or seek refuge after a typhoon or drought. Low islanders also had reason to visit other low islands in search of mates outside the limits of the incest taboo and for help after typhoons. Some of the low islanders, especially in the storm-swept central Carolines area between Truk and Yap and in the Marshalls, were skilled navigators of oceangoing canoes.

The populations of the Micronesian islands appear to have been in good balance with the natural resources at the time of their discovery by the Europeans. Because the climate is equable and varies little during the year, moderate amounts of labour were sufficient for comfortable survival, and much leisure was available for activities such as dancing, feasting, romantic affairs, and visiting friends and relatives. Some kinds of subsistence activities, especially fishing, were regarded as something like play. The period of youth was often prolonged, since the older people could afford to indulge their children. This positive attitude toward the enjoyment of leisure was especially characteristic of the high islanders, with their more fertile soil and more secure life.

SOCIAL AND POLITICAL LIFE

Settlement patterns and housing. Micronesians preferred to live in dispersed, extended-family homesteads, in contrast to some Melanesians and Polynesians. In atolls the inhabitants generally preferred the lagoon side of the larger islands for ease in launching canoes and for protection from typhoons. On the high islands, people again wanted access to the lagoon, although easily defensible sites were sometimes also preferred, such as the top of a steep, cleared slope.

The "Venice" of Pohnpei

One ancient "city" in Micronesia is the archaeological site of Nan Matol or Metalanim on Pohnpei. This is a sort of tiny Venice, containing about 100 artificial islands, separated by shallow canals, covering one-third of a square mile in all, including water. The islands were used for royal, priestly, and noble residence and for ritual, and the total population may have been several hundred to 1,000. There is a similar site on the islet of Lelu (Lele) in Kosrae.

The typical Micronesian community had one or more meetinghouses. These served both as social gathering places and as places to plan community affairs. The number and elaborateness of the meetinghouses are greatest in Yap and Palau. Palauan meetinghouses were built by the men's clubs, suggesting some Melanesian influence. In Palau, Yap, and the western atolls, meetinghouses were used mostly by men, while farther east, women and children also entered them freely much of the time.

Houses in most areas were built on slightly raised platforms made of coral rock and gravel on the low islands and volcanic rock and dirt on the high islands. They generally had thatched roofs with low eaves and poor ventilation. The smoke from a small hearth may have been used to control mosquitoes, although plaited mosquito sleeping bags were also used at times. This form of construction may have also provided good protection against heavy rainstorms. Some of the houses in the Marianas appear to have been constructed on stone pillars. The *latte* stones of this area—double rows of stone pillars with capstones—are now thought to have been piles for raised houses, perhaps for chiefs and wealthy men, since the *latte* sites are relatively few for the reported population. Other houses may have been built on wooden piles that have since disappeared.

Canoe houses were another important form of building throughout Micronesia. Those big enough to store the larger canoes were on the scale of meetinghouses and often were used as such in some areas.

Small buildings for the isolation of menstruating women were common in the western Carolines and continued in use in Yap until recent times.

Family and kinship. The majority of Micronesians before European contact lived in some form of extended-family group. In most areas the organization of this group probably had considerable flexibility, some newlywed couples living with the husband's family, others with the wife's relatives. Major determinants in the choice of residence were the relative availability of agricultural land and the need for additional labourers on one or the other side of the family. In much of the area the wife's family was considered the ideal place of residence; but frequently exceptions were allowed in practice, and children often had rights to use land on their father's mother's side. In Yap, on the other hand, patrilocal residence and patrilineal inheritance of land were considered ideal.

As in other parts of Oceania, the adoption of children of relatives was frequent. This was, in part, a way of caring for illegitimate and orphaned children; in part, a way of relieving young adults of the chores of child care and providing old people with young children to do minor work for them; in part, a way of ensuring more equitable distribution of land rights; and, in part, a way of providing heirs for specialized knowledge when the proper natural heir was lacking or unsatisfactory.

Descent based on matrilineage

Matrilineages existed in most of Micronesia, with the exception of Yap, and they were traditionally exogamous—*i.e.*, members did not marry within the same lineage. While matrilineage membership was considered basically unalterable in some communities, there was probably some flexibility everywhere. If a lineage grew too large, it tended to split into two parts, one of which would adopt a new name and no longer observe exogamy with the other. If outsiders came into a community, they would often be taken into an existing lineage as honorary or fictive members and eventually might merge with the patron group and observe rules of exogamy with them.

Marriage in Micronesia varied in formality. In the west, in Palau and Yap, marriages were marked by formal payments from the groom's family to the bride's. In the area from the central Carolinian atolls to the Marshalls, marital relationships were usually rather loose and informal, although people of high rank may have had public ceremonies with some exchange of wealth. In this area considerable premarital and extramarital sex was traditionally expected; marriage for ordinary people consisted simply of openly living together and being spoken of by the community as spouses. Apparently, there was more formality to marriage and more control of premarital sex in the Gilberts. Polygyny was generally permitted to some extent in Micronesian societies, although it was not very common. It was more likely to be practiced by high-ranking men and was sometimes restricted to chiefs.

Birth order has traditionally been widely important in Micronesian societies. The seniority of the eldest child is recognized in representing the family or lineage in public and in inheriting political office. The eldest child generally directs the use of lineage or family lands. Younger siblings generally exhibit formal respect to older siblings.

Brother–sister avoidance relationships are well-developed in parts of Micronesia, perhaps most strongly in the central Carolines from Truk through the atolls to the west. In this area women were traditionally expected to avoid speaking to their brothers, to crouch in their presence, and in other ways to show respect. In Pohnpei, farther to the east, a similar relationship existed between a man and his next younger sister but did not extend with the same force to other sisters.

Social stratification

A certain amount of hereditary social stratification was found universally in Micronesia, but its degree varied considerably from some of the smaller Carolinian atolls, which had nominal hereditary chiefs with little special power or wealth, to the high island of Yap, which had several ranked endogamous castes. Other cultures that showed relatively marked social stratification were Palau, Pohnpei, Kosrae, the Marshalls, and the Gilberts. It seems likely that the Marianas may have also had distinct social classes before the Spanish conquest. In all the above areas there appear to have been some chiefs who were supported principally by tribute from their subjects; who were the object of considerable deference; who could punish offenses (especially against themselves and their own relatives) by fines, destruction of property, or death; and whose principal wives were chosen mostly from other high-ranking families. Often they also had subordinate chiefs and officials.

The hierarchical principle was not well-developed in most of these societies. The notion of crime to be punished by the state was weakly developed. Palau and Pohnpei had an abundance of nominal political titles, the principal significance of which might be called recreational and dramatic. In both these societies individuals were able to rise above their initial hereditary status through personal achievement, thus introducing much flexibility into the system of social stratification.

On the high islands of Truk and on many low islands, especially in the central Carolines and the smaller communities in the Marshalls and Gilberts, social relationships were egalitarian. The local chief was expected to work for his living like anyone else. He received ceremonial deference but traditionally had little power to order people to do anything or punish them for offenses. His principal function was to preside at meetings and define and state the community consensus as it developed.

Social hierarchy and political organization. Throughout most of Micronesia the maximum independent autonomous political unit was the high island or the atoll, often subdivided into more than one unit. Satawan Atoll in the Mortlocks had four separate communities aboriginally, each with its own chief; sometimes they fought each other. The Palaus had two confederations of villages or districts aboriginally, each independent of the other, and the villages themselves had considerable autonomy. Pohnpei had five petty states at the time of first European contact, although there were traditions of a unified rule for the whole island at an earlier period. Truk was extremely fragmented politically, with several independent communities on each of the six larger high islands. The Marshalls and the Gilberts had larger political units, integrating groups of separate atolls under a high chief, although the fullest development of these expansionist states was achieved after the introduction of firearms by Europeans.

The Yapese Empire

The low islands between Truk and Yap have been described as belonging to the "Yapese Empire." The so-called empire consisted mainly of a chain of trading and ceremonial relationships with one of the petty states of Yap. The Yapese exerted no military force against the low islanders but did claim the ability to punish them by sending typhoons, disease, and famine if they should fail to fulfill their obligations. According to widely accepted myths, the low islanders originated from the Yapese, although they are culturally and linguistically closer to the Trukese.

In most of the area from the Palaus in the west to the Marshalls in the east, the community was considered to be owned in some sense by a clan, the head of which was also the chief of the community. Other clans had land rights by their relation to former chiefs or to the men of the ranking clan: their women had married men of the chiefly clan, or their men had rendered the chiefly clan service in the past. The chiefly clans in some cases claimed their position by virtue of ancient military conquest and in others by virtue of being the first human beings to occupy the land. The aboriginal pattern of political and community organization in the Marianas has been obscured by the early Spanish conquest, which exterminated most of the population and concentrated the remainder under close Spanish military and religious control. Chieftainship in the Gilberts was by patrilineal succession.

Socialization and education. Micronesians were indulgent with infants. Children were inducted into adult life gradually through observation and participation; there was little in the way of formal schooling or initiation ceremonies in most of the Micronesian societies. Several forms of bodily ornamentation by mutilation were practiced, mostly performed around puberty or in early adulthood. The most widespread of these was tattooing, practiced by both sexes. This and other forms of bodily mutilation were generally done on individual initiative, to demonstrate bravery and increase sexual attractiveness.

Training in cultural specialties, such as medicine, magic, esoteric mythology, canoe and house building, and navigation, was often delayed until young adulthood or middle age. It was usually given by an older relative, although sometimes outsiders could receive it by payment of food and goods. The most formal training in esoteric knowledge appears to have been given by the specialists known as *itang* (in Truk). Each *itang* had himself been trained under an older expert adept in traditional history, oratory, war strategy and tactics, and magic. At the end of his training he was formally awarded the title or degree of *itang* and thenceforth could serve as an orator, ambassador, counselor, and executive officer for a chief.

ECONOMIC SYSTEMS

Production and technology. The islands of Micronesia were probably first settled by small groups of immigrants with few technical specialists among them. The prehistoric Micronesians had basically a subsistence technology with a few part-time specialists producing luxury items, such as personal ornaments and the shell and stone valuables that were used, mostly in the western Carolines, for ceremonial payments.

Fishing and horticulture

Subsistence everywhere was based primarily on fishing and horticulture, with fishing somewhat more important in the low islands and horticulture more important in the high islands. Domestic animals were limited to dogs and chickens, found only in some areas.

Rice was grown in the Marianas but later was largely replaced by corn (maize), introduced from Mexico by the Spanish conquerors. Coconuts and coconut palms were used everywhere, both as food and for other purposes, such as thatch, lumber, and cordage. Some form of taro, either the true taro (*Colocasia*) or the giant swamp taro (*Cyrtosperma*), was probably cultivated everywhere except on some of the drier low islands in the east, where the

© Nicholas DeVore II/Bruce Coleman Inc.



Islanders of Satawal Island, Federated States of Micronesia, sailing a hand-hewn outrigger canoe.

groundwater tended to be too brackish. The breadfruit tree, which produces a large starchy fruit in abundance in the early summer, was also widely distributed. Bananas were an important food crop on the high islands. Tropical yams (*Dioscorea*), which produce large starchy tubers, are found today on a number of the high islands but are of greatest importance as a cultivated crop on Pohnpei, where they have high prestige and provide an important source of food in the winter. The sweet potato and cassava, introduced by Europeans, now serve as alternate subsistence foods on some of the high islands. On some of the drier atolls in the Marshalls and the Gilberts, the pandanus tree is a major subsistence crop. The edible fruit of some cultivated varieties contains starch and sugar that can be made into a flour and stored. Other varieties have large edible nuts. Some varieties of pandanus are cultivated for their leaves, used principally in making plaited mats and sometimes for thatch.

Many kinds of fishing were practiced traditionally. Often there was a division of labour, with men using techniques requiring canoes in deep water, such as seines and line fishing, and women wading in the waters of the fringing reefs. Deep-sea trolling with sailing canoes was practiced mostly by low islanders. The low islanders also at times made expeditions to certain small, uninhabited reefs and low islands to fish and collect turtle and seabird eggs.

Cutting tools were made from stone and shell. Iron was introduced by Europeans, although some iron tools may have been obtained from Oriental sources in western Micronesia.

Simple pottery has been found only in the western high islands of Micronesia: the Marianas, Palau, and Yap. Suitable clay apparently exists in Truk, Pohnpei, and Kosrae, but if pottery was ever introduced to these islands, it was not widely used.

All Micronesians, even the high islanders, relied often on water travel, although the high islanders used canoes principally in the sheltered coastal waters of their home islands. Micronesian canoes had a single hull with one outrigger. Canoes used in protected waters were often simple dugout canoes, but the oceangoing canoes, found especially in the central Carolinian atolls, the Marshalls, and the Gilberts, had sides built up of irregular planks that were caulked and sewn together with cord made from coconut-husk fibre.

Travel on the open sea

Some of the atoll dwellers regularly went on trips requiring several nights in the open sea. Extra provisions were usually taken along as gifts and for emergency needs if the canoe were blown off course. In historic times vessels blown off course have ended up in a variety of islands inside and outside of Micronesia. Conversely, vessels from other areas of the Pacific and from the Orient have been blown to Micronesia. It seems clear that, prehistorically, there was communication in chain fashion from the Ellice Islands (now the state of Tuvalu) and western Polynesia through the Gilberts, the Marshalls, and the Carolines to Palau and the southeast atolls and beyond them into the northern Moluccas in Indonesia. Probably there was also deliberate communication across the larger gap between the central Carolinian atolls and the Marianas; in the latter there is a long-established population of cultural Carolinian immigrants retaining their original language.

Property and exchange. The most important property among the Micronesians was land. Land-tenure customs varied considerably from island to island, even among those with related cultures. Land was owned primarily by extended-family groups or lineages. Individuals acquired use rights to particular plots through their kin connections, acknowledging the rights of the group by periodic offerings of first fruits to the kin-group head. Often a community chief and sometimes a superior chief received offerings from each household or larger kin group holding land under his jurisdiction. In some areas, especially in eastern Micronesia, strong chiefs confiscated land from those who fell out of their favour, awarding it to loyal followers.

Because land was a scarce resource in much of Micronesia, various mechanisms developed to govern its distribution. In most areas matrilineal inheritance of land rights was practiced, and, if necessary, children of the men of the lineage could also inherit use rights to the land worked by their fathers or receive a share outright if the lineage had more than it needed. In general, land was not sold, but it might be given in payment for medicine and nursing services or as compensation for an injury.

In the larger high islands, interior areas not under regular cultivation were considered community property and were used for collecting wild food and for temporary gardens. In the atolls all land was owned by one or another family group, even the smallest islets with only a few coconut trees.

Feasts as exchange rituals

Large feasts were common throughout Micronesia. The people assembled large quantities of food and offered it to the chief, and he in turn redistributed much of it to the people. Sometimes kin groups or communities held alternating semicompetitive feasts in which each tried to outdo the other. Feasts were held to commemorate important transitions in the life cycle, especially marriage and death. A kind of delayed exchange might take place at these feasts, certain relatives being obliged to make presentations of food and goods in return for past or anticipated services.

Substantial payments were sometimes made to practitioners of traditional medicine, especially when the practitioner was not a close relative. Such payments consisted of food and other goods. The precise amount was generally left to the family of the patient, with the understanding that a stingy family might not get the most energetic and effective treatment.

In Palau and Yap, shell and stone valuables were used in transactions at life crises and for certain other specific compensations, such as payment for injuries. These are often called money, but their use was much more limited and specific than that of most money. The best known of them were the large stone disks that the men of Yap manufactured on Palau and carried home in sailing canoes.

A certain amount of trade developed between the low islanders and high islanders, the low islanders providing handicraft products that the high islanders could have made if they had needed to, and the high islanders providing things obtained more easily there. This trade was especially well-developed between the low islanders of the central Carolines and the high islanders of Yap and Truk. The most important high-island export was turmeric, used for medicinal and cosmetic purposes, which formed a bright orange body paint when mixed with coconut oil. The low islanders provided shell beads, plaited pandanus mats, and coarse cloth woven from banana or wild hibiscus fibre and used for women's wraparounds and for men's loincloths. Trade was often with particular partners who regarded themselves as distant relatives. An important function of the trade was to provide the low islanders with aid and a temporary dwelling place when their islands were devastated by the periodic typhoons.

RELIGION AND ART

After 1900 Christianity became well-established in most major centres in Micronesia. Traditional religions were no longer practiced in their full original form, although in Yap and some atolls of the central Carolines, traditional religion continued to be practiced until the middle of the 20th century. Missionaries and travelers have left descriptions of certain aspects of the island religions, but there is no complete and systematic account.

The basic patterns of religion were probably similar throughout most of Micronesia. The Micronesians were polytheists, believing in several high gods, a large number of spirits attached to specific localities or performing specific functions, and a number of ancestors and deceased neighbours who could sometimes make contact with their living descendants and friends. Practices connected with these three major categories of supernatural beings tended to be distinct and to be handled by different specialists, although a figure might gradually shift from one category into the other.

Micronesians generally believed in at least three vertically arranged levels of the universe: the earth proper, the underworld, and heaven or the sky world. Some Micronesians may have believed in multiple heavens, as did

people in Polynesia and Indonesia, but the details of their beliefs are now unclear.

Micronesian origin myths generally tell of the origin of particular islands or descent groups or the initial discovery or later conquest of an island by the ruling descent group. The notion of the supernatural creation of the whole human species or of the whole world is either not found or little emphasized in Micronesian mythology.

Ceremonies for the high gods appear to have been principally seasonal offerings of first fruits, performed often in private by a specialist priest with a few helpers. Special appeals to the high gods were probably also made at times of community crisis, such as wars or typhoons, but human sacrifice apparently was not practiced. The priests were very likely relatives of the ruling chiefs. The lesser spirits were called on by magicians for specific purposes, most notably for the diagnosis and cure of disease but also for such purposes as success in fishing, control of weather, success in love, and prowess in athletic contests, battle, canoe building, and other difficult pursuits.

Communication with spirits

Ancestral spirits were often contacted in dreams and in the trances of spirit mediums, as were the high gods and other nonhuman spirits. They would give people information about the causes of disease, deaths, and other misfortunes and would sometimes prescribe new medicines or new varieties of magic. At times, the spirit mediums would also order their human protégés to perform songs and dances for the entertainment of the spirits, to win their goodwill and ensure the prosperity of the community.

Micronesia has little permanent visual art. The best known consists of the painted bas-relief scenes on the timbers and gables of the men's houses in Palau. Stylized wooden masks and human figures have been found in the Mortlock Islands, southeast of Truk. Decorative lashings of light and dark sennit twine are found in important buildings. Painting and decoration of canoes is common, following traditional patterns for each island. Patterns were woven into clothing. Wooden bowls were carved and painted with aesthetic intent.

Much of the most-developed traditional art of Micronesia was ephemeral: body painting and ornamentation with garlands of flowers, singing, dancing, and the recital of myths, tales, and poetry. Oratory was also an important traditional art and still survives. A good orator was expected to make mythological and historical allusions and to use special figures of speech. (J.L.Fi./R.C.Ki.)

EFFECTS OF EUROPEAN CONTACT

Micronesia has the most complicated colonial history in the Pacific. Guam, the southernmost of the Mariana Islands, became the first inhabited Pacific island to be visited by a European when Ferdinand Magellan landed there in 1521, and the Marianas became the first European colony when Spain took control of the island chain in 1668. In 1670 the indigenous Chamorro people rebelled, and a quarter-century of sporadic warfare followed. That conflict, along with diseases introduced by Europeans, reduced the local population from perhaps as many as 100,000 to 4,000. Most of the Chamorro people were moved to Guam. Much of Chamorro culture was destroyed, but the language survived.

Spain attempted to expand its rule as far east as Pohnpei but did not succeed. The Germans claimed the Marshall Islands in 1885 and the island of Nauru in 1888. In 1892 Britain entered the scene in eastern Micronesia and added the Gilbert Islands to its colonial empire.

The Spanish-American War brought an end to Spain's presence in the Pacific. Upon Spain's defeat in 1898 the United States seized Guam because of the strategic coaling station that Spain had developed there. In the following year Germany purchased Spain's claims to the Caroline Islands (Palau, Yap, Pohnpei, Truk, and Kosrae) and the northern Marianas. The copra industry was greatly expanded under German rule, and its administration was one of indirect rule.

World War I changed the face of Micronesia. Japan replaced Germany in the northern Marianas, the Carolines, and the Marshall Islands. In 1920 the islands became a mandated territory under the authority of the League of Nations. Japan initiated an aggressive program of economic development, and Japanese, Okinawan, and Korean immigrants eventually came to outnumber the islanders by a ratio of two to one. Sugar, fishing, and other industries were developed, and the copra industry was greatly expanded.

Micronesia and the League of Nations

After World War I Nauru, which had been held by Germany, also became a League of Nations mandate, with Australia having administrative responsibility. Guam remained under the firm direct rule of the U.S. Navy, and the Gilberts continued as part of the larger British crown colony of the Gilbert and Ellice Islands with a system of indirect rule.

World War II again altered the political map of Micronesia. After its attack on Pearl Harbor in late 1941, Japan occupied the Gilberts, Guam, and Nauru, and for the only time in its history all of Micronesia was under one colonial master. The war inflicted great suffering, and major military engagements took place between Japanese and American forces in parts of the Gilberts, the Marshalls, Truk, Palau, Guam, and the northern Marianas. At the end of the war the economy was a shambles, a great amount of property had been destroyed, there were food shortages, and many people had been displaced.

After the war Guam was returned to the United States, and British colonial rule was resumed for the Gilberts. Nauru became a trust territory of the United Nations, similar to the old mandate system, and Australia resumed administrative responsibility. The United States replaced Japan in the former mandated islands, and this portion of Micronesia became the United Nations' Trust Territory of the Pacific Islands (TTPI).

The decolonization of the Pacific began when the Polynesian state of Western Samoa became independent in 1962. Nauru was the first Micronesian country to become a sovereign nation when it gained independence in 1968. Because of its rich phosphate deposits, its people have one of the highest per capita incomes in the world. Large sums are invested in trust funds around the world in preparation for the eventual exhaustion of its mineral wealth, which is expected to occur before the turn of the century. Nauruans are almost totally dependent on imports. Little traditional culture remains; housing, foods, and clothing are almost entirely Western.

The Gilbert Islands, along with Kiritimati (formerly Christmas) Island, Banaba (formerly Ocean) Island, and most of the Line Island and Phoenix Island groups, now form the nation of Kiribati, having gained independence in 1979. The Gilberts form the core of the nation and have the vast majority of its population. Outside of Tarawa Atoll, the site of the nation's capital, much of the traditional culture and life-style prevails. Subsistence agriculture and fishing remain the mainstays of the outer island economy, and most buildings and dwellings are of traditional designs and materials.

Negotiations regarding the future political status of the TTPI began in 1969, and it soon became clear that the peoples with such diverse cultures and languages did not feel any real sense of kindred or national unity. Three island groups had strategic value to the United States and used it to their advantage to demand separate negotiations with the United States.

Strategic islands

The northern Mariana Islands were the first to break away. The highly Westernized Chamorro peoples had little in common with other Micronesians, and in 1975 they voted to become the Commonwealth of the Northern Mariana Islands, an integral part of the United States. The islands have a well-developed tourist trade and a sophisticated economy. With commonwealth status, the Chamorros became American citizens, and the population is highly urbanized and concentrated on the island of Saipan, the capital.

Palau and the Marshall Islands became republics, and the remaining islands of the TTPI formed the Federated States of Micronesia (composed of the four states of Yap, Truk, Pohnpei, and Kosrae). Each approved a Compact of Free Association under which the United States handles matters of defense, and strategic access is denied to all other nations. In return, the United States provides

substantial subsidies to the nations, and their citizens have the legal status of "habitual resident," allowing them the right to live in and seek employment in the United States. The compacts have been approved by the United States for the Federated States of Micronesia and the Republic of the Marshall Islands.

Each of the Micronesian entities has its own capital and urban area, and in the case of the Federated States of Micronesia there is a capital in each constituent state. Approximately one-half of all islanders in the former trust territory are urban dwellers today. They are heavily dependent on imports, and, except on the outer islands, much of traditional life has been lost. (R.C.Ki.)

Polynesian cultures

The physical environment of the islands of the Polynesian triangle is not as favourable for human habitation as it might at first seem and certainly presented difficulties when the ancestors of the Polynesians entered the area from the west, well before the beginning of the Christian Era. The islands were devoid of much that was needed for human habitation. Most of the food and useful plants and all of the domestic fowl and animals had to be transported by the settlers. Since that time, the physical environment has continued to exert a marked influence on the nature and extent of the evolution of Polynesian cultures.

The popular image of Polynesian cultures suggests an almost blissfully simple and easy way of life, devoid of harsh extremes of any type, played out on islands of great beauty and natural abundance. But the image, a product of popular literature and the motion-picture and tourist industries, is wide of the mark. Traditional Polynesian cultures were complex, highly specialized, and diversified, and the environment was not always benign.

The principal characteristic of Polynesian cultures is an effective adaptation to and mastery of the ocean environment in which they have flourished. This mastery did not extend merely to the technology involved in ship building, to the techniques of navigation, and to other obvious nautical aspects of the culture but permeated social organization, religion, food production, and most other facets of the culture. The Polynesians could not only sail the seas, but socially they could cope with the human problems of shipwreck, split families, and the sudden loss of large portions of the social group. In short, the cultures were well-equipped to handle the numerous hazards that they had to face in the beautiful but actually quite hostile Pacific environment.

Polynesian conservatism

Another important characteristic of Polynesian cultures was a high degree of conservatism and traditional orientation and the rather detailed structuring that placed every person in a well-defined relationship to society and to the universe. This conservatism is apparent even in a rapid examination of Polynesian materials from islands separated by hundreds or thousands of miles, whose populations had been separated from the common Polynesian stock two or three millennia in the past. The similarity between such items as stone adzes and fishhooks is most remarkable. The same is true for kinship terms, plant names, and much of the rest of the technical vocabulary of the cultures, as well as for art motifs and medical preparations. The ornate and voluminous genealogies, chants, legends, songs, and spells that were passed down and elaborated through the generations show a profound reverence for the past.

The traditions of Polynesian cultures provided a definition of an individual's relation to his society and to all of nature. The creation myths told of the origin of the world, setting forth the order of precedence of Earth, sky, and sea and all the phenomena thereof, including man and woman. The genealogies fixed the individual tightly into the typically aristocratic Polynesian social order. Other legends interpreted natural phenomena, while historical accounts often described, with varying amounts of mythological embroidery, the migrations of a people before they arrived at the island on which they were located, their adventures on the way, and the development of the culture following settlement.

The Polynesians were unusually well-oriented, as would be expected of a nautical culture, to the entirety of nature. Their languages abound with terminology for stars, currents, winds, landforms, and directions and include a large number of grammatical elements indicating, for example, direction of motion implied by verbs, including movement toward or away from the speaker, relative positions of objects with reference to the speaker, and direction of movement along a seashore-inland axis.

Polynesian cultures displayed a thoroughly realistic and practical exploitation of the environment but at the same time exhibited a profound interest in the supernatural. This interest was directly related to the general interest in systematization and structure, the supernatural being for the Polynesian another portion of reality, albeit less manageable, that had to be structured and coped with.

Violence and cruelty were ever-present elements of Polynesian cultures. This is reflected in their oral literature and in all aspects of traditional life. Custom controlled and repressed direct physical expression of aggression within the kin group and the tribe up to a point, but there were definite limits beyond which only violence could restore status or assuage injured pride. Punishments for transgressing rules of behaviour toward chiefs and high priests or for violating various rituals incorporated ritual sacrifice and cannibalism as major features. Intertribal warfare was extemely common, particularly when population pressure had built up and land and resources were limited.

Perhaps the most publicized, and accordingly overrated, aspect of Polynesian culture is its sensuality: as in many other aspects of life, the Polynesians generally took a very direct, realistic, and quite physical approach to gratification of the senses. But there was no abnormal focus or concentration on any aspect of sensual gratification such as is seen in many other cultures where, for example, eating, drinking, or sex have become the object of great cultural elaboration. This lack of elaboration may be due in part to limitations imposed by the environment. The flora of Polynesia did not, for example, provide a very good basis for elaboration of cuisine, as the number of edible plants was rather restricted, and there was a dearth of spices. In general, however, this balanced approach to sensual gratification seems to be just another reflection of the Polynesian straightforward approach to the world. Polynesian culture, for example, is generally thought to have been characterized by extreme sexual promiscuity, but such is not the case. Definite restrictions were placed on sexual behaviour, though the limits of acceptable behaviour were indeed wider than in many other cultures of the world. Children were permitted freedom of sexual expression; from puberty on, sexual activity was strongly encouraged, limited only by incest restrictions (which were not stringent) and considerations of social prestige. Parents would become concerned if daughters were not sought after for sexual favours, and premarital pregnancy only enhanced a girl's attractiveness. Public intercourse was engaged in during certain religious festivals and crisis rite ceremonies as a means to promote general fertility. There was, however, no pathological elaboration of sex for its own sake. Polynesian art, for example, is strikingly devoid of sexual content. There are occasional references made to sexual activity in native Polynesian oral literature, but many of these have to do with creative acts of the Polynesian gods, for the Polynesians saw the process of the creation of the universe as having taken place in a manner similar to human procreation. There is no corpus of Polynesian literature analogous to pornography.

Sexual mores

SOCIAL AND POLITICAL LIFE

Settlement patterns and housing. Two major settlement patterns were present in Polynesia prior to European contact: the hamlet and the village. Their origin and development may depend to a large extent on geographic factors, such as the distribution of food-crop resources, on defense considerations, and on social organization.

The hamlet settlement pattern was found in the larger volcanic islands, where food resources were diversified and scattered over the range of environmental zones that developed on such islands. A typical hamlet settlement pattern was often found in the Marquesas Islands. There,

in prehistoric times as at present, the population spread over the floors and up the sides of the deep and narrow Marquesan valleys in clusters of perhaps four to five houses, often with gardens, taro patches, coconut trees, and breadfruit clustered in the immediate vicinity. The Marquesan houses were built on rectangular platforms, the height and composition of which depended on the prestige of the owner. Individuals of lower status might have a simple paved rectangle no more than a few inches high, while warriors, priests, or chiefs might live in houses perched on platforms seven to eight feet high and containing stones of several tons in weight. The house itself was built on a kind of a step running across the rear of the platform. Composed of a lashed and fitted wooden framework and with a thatched roof, the typical house was open all the way across the front and had square ends. The roof sloped from the high ridgepole directly to the platform floor in the rear. Inside, a polished coconut log often ran the length of the house, serving as a kind of community pillow. The floors were covered with mats, shredded leaves, or bark. Belongings were suspended in bundles from the rafters. Most of the household activity took place on the "veranda," or unroofed front portion of the platform, which was paved with smooth basalt stones from streambeds. Houses of chiefs and other individuals of high status often made use of cut stone slabs for decorating the platform. Many also had rectangular pits in the platforms for storing fermenting breadfruit paste (an important delicacy) and small caches in which were interred the carefully cleaned and packaged bones of important family members.

Among the clusters of houses in each major valley section was a large rectangular ceremonial plaza, or *tohua,* which served as the focal point for all tribal or subtribal ceremonies, as well as a point for daily social gatherings. This complex of structures, often 600 feet (200 metres) or more in length, was generally built upon a huge artificial terrace carved out of a slope. Surrounding the plaza atop the terrace were the houses of the tribal chiefs and priests, temples and other sacred structures, and long sheds for spectators. At one end was usually found a temple with a sacred banyan tree in which were suspended the packaged bones of the illustrious dead of the tribe.

In Samoa, on the other hand, the village settlement pattern was followed after AD 1000 or so. There, fortified villages, consisting of 30 or more houses arranged in a network of paths, grew up along the coast. The houses at the time of European contact were built on oval mounds faced with rough stone slabs, although earlier houses had been built on rectangular platforms much like those of the Marquesas. The typical Samoan house was and is today a large shedlike structure without walls, oval in floor plan, with a beehive-shaped, thatched roof supported by a series of stout wooden pillars. Rolled mats along the eaves were let down to protect the inhabitants from sun, rain, or the night air. These houses, arranged in orderly fashion along the paths and sunken roadways that crosscut the villages, were surrounded by a fortification wall of stone or by wooden palisades.

Maori fortified villages

The Maori of New Zealand constructed large and impressive fortified villages (pas) on hilltops, surpassing those of all other Polynesian cultures. These forts were protected by ditches, palisades, trenches, and terraces. The interiors were partitioned off by additional defensive works so as to facilitate battle even after the outer defenses had been penetrated by an enemy assault. The houses within were of timber, rectangular in plan, and generally dug about one foot into the earth surface.

Other, unfortified Maori villages consisted of loose collections of houses arranged around a level area where ceremonies and meetings might take place. Forts were usually constructed in the vicinity of such villages for the use of inhabitants in time of war.

Family and kinship. The typical Polynesian family was extended; that is, it consisted of three or more generations, and it centred around a descent group. The extended family and the descent group were not the same. In most cases descent was preferably traced through males, with both females and males belonging to their father's descent group, and the extended family consisted of a senior male, his sons and grandsons, and their spouses and unmarried children of both sexes. Residence was usually virilocal; females left their family of birth and joined their husbands' families upon marriage. Polynesian societies were noted for their flexibility, however, and there were many variations in the system of descent. In Hawaii, Tahiti, and elsewhere, and especially if it were to one's advantage, descent could be traced through females and one could become a member of one's mother's descent groups. Thus, while descent through males was preferred, the descent system was actually bilateral (through either parental sex). Adoption was very common and added to the flexibility of the social system. Siblings and cousins frequently adopted one another's children, and grandparents sometimes adopted their own grandchildren. These children behaved toward their adopted parents as they would their own.

The basic Polynesian kinship pattern is a very simple one: it distinguishes among generations, as might be expected in a society so strongly oriented on tradition and genealogy. There are sets of terms for the grandparents' generation, distinguishing by sex only; parents' generation, in which parents are distinguished from various categories of aunt and uncles; members of a person's own generation, in which the terms permit identification by sex, relative age, and sometimes marital potential (marriage of certain cousin categories is preferred in some Polynesian societies); and, finally, children's generation, in which age and sex are again distinguished.

In general, the kinship terminologies in use in the various societies in eastern and marginal Polynesia are relatively simpler than those in the west.

Social hierarchy and political organization. The two major types of settlement in Polynesian cultures—hamlet and village—were bound together by a web of kinship and territorial ties of varying degrees of importance.

There were two main and probably related types of kinship-based descent groups prevalent in Polynesian society. By far the most common and perhaps the most like the ancestral form of Polynesian social organization was the "ramage" type, in which the whole society might be represented in the form of a multibranched tree. There was a senior line of descent, in which descent passed from firstborn son to firstborn son and was traceable back to the mythological past. Branches off this main line were founded by junior sons, and these branches in turn produced further branches. Because the main line was the direct line to the gods and therefore carried the maximum traditional prestige and sanctity, it followed that each of the branches was ranked in terms of its proximity to the main line, and each individual in each branch could place himself in a prestige-ranking scale relative to other members of his household settlement and kin group.

"Ramage"-type descent

The levels of the branches off the main line assumed different levels of importance in various Polynesian societies. In New Zealand, for example, four levels were recognized: (1) The *waka* were a group descended on both maternal and paternal sides (male preferred) from the crew of one of the original canoes in the traditional Maori migration to New Zealand. They occupied a given territory and had a chief who functioned only in times of crisis and members who recognized their common tradition and interest. (2) The *iwi* were component parts of the *waka* but were made up of individuals tracing descent from a single occupant of one of the canoes, through either maternal or paternal line (again, the male line was the prestige line). The *iwi* had a chief and was a more tightly knit kinship group occupying a defined section of land within the *waka* territory. (3) The *hapu,* a component unit of the *iwi,* was the principal Maori descent group, consisting of all descendants of a given ancestor, usually sited in a particular village and controlling the farming land around the village. The *hapu* had a chief responsible for all economic and civil affairs. (4) The *whanau,* or households of the *hapu,* formed the lowest unit within Maori society and usually consisted of three to four generations.

Each of the hierarchical kinship segments just mentioned could be ranked relative to the others, with the exception of the *iwi.* The *hapu* could all be ranked within any given

iwi on the basis of their chiefs' ancestry. Within the *waka,* each *iwi* could also be ranked in terms of the distance of its founding ancestor from the chief of the canoe.

This form of hierarchical branching-descent-group organization with territorial overtones was found in most Polynesian societies, with appropriate variations for local environmental conditions and cultural history.

"Descent-line" kinship groups

The other major form of kinship-based descent group found in Polynesia is that known as the "descent line." Descent-line organization appears to be the result of a breakdown in genealogical ties between the lower levels of a ramage organization. The descent line in Samoa, for example, consists of a group of people tracing descent in the male line from a common mythical ancestor. This group was known as a *sa.* There was no concern for genealogical relationship of one descent line to another, nor was there any concern for ranking based on distance from or proximity to any particular male line of descent. What were passed down through the descent line were titles, each of which had rank and prestige attached to it. Each descent line held a number of these titles, which enabled it to participate in certain ways in the village council (fono). A number of descent lines were represented in each village council, and each descent line was spread through a number of villages in a given area. Within a given village, the senior title, determined by mythological connections, gave its holder the position of chief.

A characteristic of the descent-line system is its flexibility. Depending so heavily on myth and tradition for validation of status and title, it was only natural that ambitious titleholders could advance the prestige of their titles at the expense of others by various kinds of prestige behaviour, such as displays of wealth and power. The traditions governing title seniority could then be judiciously tampered with to produce the realignment that would allow the advance in status to occur. New descent lines might also be founded with similar "rewriting" of oral tradition.

Stratification. The tendency toward social stratification is perhaps inherent in Polynesian society, with its great interest in and emphasis on status and the indicators and prerogatives thereof. There seems to have been a process of social evolution occurring in Polynesia that tended to be somewhat unidirectional, leading toward more sharply defined differences in status and development of social classes that were clearly defined in terms of rights, duties, behaviour, and general life-style.

The chief in ancient Polynesian society was the person of highest status, yet he was regarded by his people and generally conducted himself as merely "first among equals." In most of the more traditional of the societies, the chief could not appropriate the land of his followers, nor did he appear to be too interested in increasing his own group's holdings at the expense of neighbouring groups. In terms of his clothing and behaviour, there was little to distinguish him from other males. Nevertheless, he was the repository of sacred power for the group, a symbol of its tie with the past, and the vehicle whereby this tie would be perpetuated for coming generations. Even in the traditional Polynesian societies, however, it was possible for a man to rise in prestige by various achievements; by giving gifts, holding feasts, or displaying military prowess, for example. These activities seem to have contained the seed of change in the traditional social order.

The breakdown of traditional social order

Warfare. There was a group of Polynesian societies in which the more traditional social order, when breaking down, gave way to one in which chiefs and other individuals of high status no longer held office on the traditional genealogical basis but by virtue of having taken it, often by force, or as a result of some act that raised their prestige higher than that of the actual chief. In these societies the chiefs behaved more autocratically and set themselves off from the rest of the population by various kinds of privileges and behaviour. Changes in leadership were frequent, and kinship groups were split internally and reorganized in the course of internecine wars. Finally, out of this dynamic social situation emerged the most socially and politically advanced Polynesian societies (Hawaii, Tahiti, Tonga), with clear-cut social stratification and concentration of all power at the apex of the social pyramid—the paramount chief—below whom were several levels of nobility usually selected on genealogical relationship to the chiefly line. The paramount chiefs and upper classes had absolute powers over the commoners, inflicting cruel physical punishments, such as human sacrifice, for all manner of affronts to their dignities (touching a chief's shadow, for example). Warfare was frequent, well-organized, and marked by acts of great cruelty. Religious and political powers were closely linked, with religious concepts and the gods themselves mirroring the cruelty and viciousness of the societies that had produced them.

Socialization and education. The Polynesian child was generally born into a large and warm family environment. Even before the child could walk, it was turned over for care to the other children of the household (siblings, cousins), who generally associated in a kind of amorphous play group with children of other families. It was in this context that the Polynesian child received a great deal of its socialization, first carried piggyback by an older sister and then running with a crowd. A particularly warm relationship existed between a child and its grandparents, often taking the form of rather gross bantering but providing the vehicle for inculcation of traditional lore, technical training, and even sexual advice.

There was little by way of formal education in Polynesian society, except for the training in special crafts and skills, such as canoe making or tattooing, and the sacred academies for training the priest-bards who were the repositories of the society's traditions, mythology, and genealogies.

Rites of passage varied in type and importance from society to society, but there were several that were generally held throughout Polynesia. The birth of a child was a matter of great significance, particularly if the child happened to be a firstborn son of a high-status descent group. Various procedures were called for to announce the birth to the community, to the ancestors, and to the gods, and to care for the welfare (both physical and supernatural) of the infant and mother by application of medical and magical techniques.

Circumcision was a major event in the male life cycle and was marked by elaborate rituals, which increased in importance with the status of the male. Polynesians generally have an unusual abhorrence of the uncircumcised, to the extent that in some societies it was believed such people would go to a very unpleasant netherworld after death. Although no such rite involving mutilation is reported for girls, some societies, such as the Marquesas, had orgiastic ceremonies in which adolescent girls made a more or less public debut, in a sexual sense.

Distinctive rites of passage

There were other milestones in life that were marked with ceremonies in Polynesian cultures, such as the formal presentation of a royal heir, the completion of a tattooing operation or ear piercing in a high-status child, and the formal investiture of a priest or chief. The observances of these milestones were marked with a variety of rituals and quite often included human sacrifice, wanton killing of members of the population who might happen into the path of the ceremonial party, or killing of those who made mistakes in performance of the rituals.

Death, the terminal milestone, was universally celebrated with extravagant behaviour, increasing in extravagance in direct proportion to the status of the deceased. In many societies these ceremonies were marked by violence, with mourners mutilating themselves and others, and by human sacrifices, obtained from within the social group or without. Feasts were also common, as was orgiastic behaviour. The extravagance of funeral rites was surpassed, in some societies, by ceremonies to deify a departed chief or priest. These went on for prolonged periods, involving prodigious feasting and drinking, violence, and sexuality.

ECONOMIC SYSTEMS

Production and technology. The Polynesians were above all mariners, but among the many seafaring peoples of the world they were quite unusual, for, far from having a disdain for the land and all that it involves, they were capable of and devoted to horticulture and arboriculture.

The sea provided most of the protein in the Polynesian

diet. Fishing was carried on not only on an individual basis, with spear, line, or net, but on a group basis with large numbers of men spreading and drawing in huge nets in bays or lagoons or driving fish toward shore, where they could be captured in nets held in shallow water. In some Polynesian societies (Marquesas and Samoa, for example) there were specialists who directed the mass fishing efforts and the elaborate religious rituals that went along with them. Sea mammals (porpoises and whales) were also taken, the former often being driven into shore by means of large, well-organized drives: when a school of porpoises was sighted in a bay, a fleet of canoes would seal off the bay entrance, and crew members would leap into the water carrying stones in their hands, which they beat together as they swam slowly toward shore. The sounds terrified the porpoises, and they beached themselves in a frenzy.

The Polynesian fishermen's exploits were not, however, confined to coastal waters, for they were just as much at home on the high seas. The waters around their island homes were explored for miles in all directions. The Easter Islanders, for example, used to travel east to Sala y Gómez reef to fish—a journey of some 300 miles (500 kilometres). Good fishing waters were located by reference to land bodies if visible or just by dead reckoning. Line fishing to depths of 90 feet was not uncommon to take the grouper. School tuna were taken in large quantities on the high seas by means of shiny pearl shell lures with bone points. The flesh of sharks and rays was a delicacy.

Mollusks and crustaceans were important as food, and mollusks also provided the shells from which a wide variety of tools and ornaments were made. Clams, cowries, and various snails and conches were collected, generally by women and children, along the reefs or shorelines and in the shallows. Shrimps were netted with fine nets, while lobsters were collected by men who dove and pried their spiny prey loose from underwater crannies. The octopus was lured from its hiding place by an ingenious lure made of leopard cowrie shells. Various types of seaweeds and algae were also collected and were highly prized for their salty taste. There was, in short, nothing edible in the sea that was not food for the Polynesians.

Horticultural practices

Gardens and groves produced the staples of the Polynesian diet and most of the condiments. The major native Polynesian crops were yams, sweet potatoes, taro, breadfruit, bananas, sugarcane, coconuts, and Tahitian chestnuts (*Inocarpus edulis*). These crops achieved different levels of importance in various Polynesian societies, depending on cultural factors and environmental conditions. The Hawaiians, for example, relied heavily on taro, building extensive irrigation systems to grow the variety that requires muddy soil and planting the "dry" variety in the uplands. Breadfruit was not of great importance in Hawaii, however. In the Marquesas and Tahiti, on the other hand, breadfruit was the major staple, although taro was by no means neglected. In these islands breadfruit was allowed to become overripe and was then beaten into a pulp and stored in large, well-drained pits in the ground, in neat hibiscus-leaf bundles. The breadfruit paste thus stored would ferment but remain edible and nutritious for years, its sour taste being highly prized for imparting flavour to the rather bland fresh breadfruit paste.

In the subtropical climate of Easter Island and in temperate New Zealand, the most important Polynesian food plants could not survive. On Easter Island, for example, the coconut did not survive, although legends maintain that germinating coconuts were brought by the earliest settlers. This deprived the Easter Islanders not only of a variety of condiments and a rich source of vitamins but also of the leaves that were used elsewhere in Polynesia for thatch and baskets and of the husk fibres that were used to make fishing line and netting. In New Zealand it was not only the coconut but also breadfruit, yams, and bananas that did not survive. The radically different flora and fauna of New Zealand, however, offered foods to replace, at least partially, those that it would not support.

Not only the coconut but most food plants in Polynesia also provided materials for other purposes. The breadfruit furnished wood for the hulls of dugout canoes, and the milky sap of the fruit itself was used in caulking the gaps between the planks of larger canoes. The black mud of the taro patches was used to stain wood and stone carvings. The pandanus fruit was highly prized as a kind of chewing gum, as well as a major component of floral necklaces and headpieces, to which it contributed its strong and pleasant colour. Pandanus leaves, when bleached, became the raw material for fine mats. The candlenut was used for torches, its oil was a cathartic, and its wood furnished certain canoe parts.

Polynesian technology can be circumscribed by five items: wood, stone, vegetable fibre, shell, and bone. The canoes, houses, domestic utensils, weapons, religious sculpture, and a host of other incidental tools were fashioned from wood with stone or shell adzes; stone-flake knives; files made of coral, sea urchin, or rough stone; and drills of bone, stone, or shells. Fine carving was done with stone, shell, or animal teeth, particularly those of rats or sharks. Stone, generally fine-grained basalt, was the hardest material available to the Polynesians and was used to produce a variety of adzes.

The absence of metal

The absence of metal for use in attaching or clamping wooden parts posed no obstacle to Polynesian construction. Wooden components were skillfully fitted together and lashed with cordage made of various types of vegetable fibre, such as hibiscus bark, pandanus-leaf fibre, coconut fibre, or banyan bark. Huge double canoes 100 to 150 feet in length, for example, were built of numerous small wooden components all held together only by fitting and lashing, yet they were able to withstand the pounding of wind and wave for thousands of miles.

Vegetable material also furnished a major source of clothing in the form of the beaten bark (tapa) of the paper mulberry tree or the banyan. This material was beaten out into small sheets that could then be assembled to produce loincloths, capes, skirts, and headdresses. Bark cloth was decorated by painting and watermarking and by attaching feathers, shells, teeth, and other ornaments. Fine mats were also used as apparel in western Polynesia.

Home furnishings also consisted mainly of vegetable material in the form of mats for sleeping and sitting, as well as for protection from the weather, and of baskets for holding personal belongings and food. Coconut shells and bottle gourds provided handy, durable containers for various items.

Pottery was in use in western Polynesia at the time of European contacts. It had been in use earlier in the Marquesas Islands, at the extreme eastern edge of Polynesia, but had disappeared from the technology.

Craft organizations are an interesting aspect of Polynesian culture. These organizations provided for the training of specialists in tasks requiring special knowledge and higher levels of manual skills—canoe construction, wood carving, and adz making, for example, as well as complicated operations, such as fishing and tattooing. When specialist groups existed, they were usually designated by name and would ply their craft for pay in the form of goods and services, as well as support for the period of their service.

Property and exchange. Personal property was a well-developed concept in prehistoric Polynesia. Each individual, regardless of rank, had his own tools, small canoes, clothing, ornaments, trophies, weapons, and other items and was generally free to dispose of them as he wished. Other types of property, however, were owned by extended families or descent groups in common and used for the common good. These included items too large to be produced or managed by a single man alone, such as a large double canoe or a fishing net several hundred feet in length, as well as facilities and land intended directly for community use, such as a tribal ceremonial ground, a refuge fortification, or a large breadfruit-paste storage pit.

The matter of ownership of land and the means of production were complicated and varied among Polynesian cultures depending on the type of social organization that they developed. In Polynesian societies of the more traditional type, land was vested in a corporate descent group. Among the Maori this group was the *hapu,* and its members usually inhabited one or a few villages located in close proximity to each other. The land of the *hapu* was viewed as being symbolic of the *hapu* chief's sanctity

Maori landholding practices

and power and as almost a literal extension of his body. Yet, the land was apportioned out to the various family groups within the *hapu* and worked by them. The chief, regardless of his prestige (stemming from a close family tie to the tribe's mythological ancestors), could not evict a family of nonproducers or in any way cut them off from their source of livelihood. He could invoke supernatural powers in the form of a curse or of a ritual prohibition, but expropriation was quite uncommon. In return for this essentially inalienable right to his land, the good Maori was expected to provide certain quantities of produce for the chief on various occasions and also for the chief of the next highest element in Maori social organization, the *iwi*. If his offerings were not of sufficient size and quality, the Maori risked punishment and possibly even death in extreme cases but not expropriation. If the food offered was acceptable, however, the offerer would be assured of an appropriate amount of food of other types in return, demonstrating the chief's desire to live up to his own obligations.

A number of Polynesian societies were essentially similar to the Maori in their landholding practices, including those of the Tikopia, Pukapuka, Futuna, and Uvea (Wallis); but, in others, evolutionary changes in social organization began to exert pressure on the basic landholding descent group (as exemplified by the Maori *hapu*), which seems to have been the basic building block of Polynesian society. As a result, the descent group ultimately lost control of all rights to land, surrendering it to increasingly powerful and autocratic chiefs. Thus, in Hawaii, perhaps the most sociopolitically evolved of all the Polynesian societies, there was a large mass of completely landless commoners dwelling in territorial districts, each of which ran from the seashore to the island's centre, thus including a segment of each of the island's ecological zones.

Each of these territorial districts was ruled by a chief, whose position was somewhat like that of a lord in medieval Europe. The chief was responsible to the paramount chief of the island for supervising, among other things, the production of that district. The district chief or the paramount chief was free to shift at will the population of commoners within the district, regardless of their kin affiliation, although the chief and nobles themselves still relied heavily on their own kinship affiliations and prerogative. The district chief saw to the production of each household in his district and could evict the unlucky farmer who failed to meet his quota. Both he and the paramount chief could also make a variety of arbitrary demands from which the commoners had little or no protection, except to leave the district and submit to the authority of another district chief. The commoners were responsible for the timely production of certain specified quantities of food, which was collected and passed up to the district chiefs and ultimately to the paramount chief. At that level, certain types of food reserved for chiefs only would be set aside, and a redistribution of the remaining food would take place, with portions being passed back down the same channels that had been used to collect the food in the first place. Although kinship groupings, consisting of all individuals related through both male and female lines, still existed, their main functions seem to have been to serve as a channel for low-level exchange of food and other commodities and to assist in the organization of the massive food collections held periodically throughout the year for the nobility.

Economic redistribution in Hawaii

The Polynesian system of exchange of goods and services may be summarized by two terms: redistribution and reciprocity, both already mentioned in connection with the discussion of the landholding systems. The redistributive system was essentially a vertical system with goods moving up from the lower strata of the society to the chiefs and other high-ranking persons and then being apportioned and redistributed, so that all would share in more equal fashion in the productivity of every kinship group or region. This system involved a complex, shifting web of reciprocal obligations, which is very much at the heart of Polynesian culture, even today. Goods and services rendered, even if not requested, necessarily give rise to an obligation for a return in kind. There were no markets in Polynesian cultures, nor was there any standard medium of exchange. All exchange was in the form of barter, often under the general supervision of some senior family or kin-group member. Thus, for example, a portion of the fish catches made on a minor fishing expedition by coastal residents would be passed inland to residents at central villages, who might return dry taro for the fish at some fixed ratio of units of taro to units of fish. The services of any of the numerous specialists in Polynesian cultures (tattooists, fishermen, wood-carvers, for example) was also paid for by barter, usually over and above the cost of the specialist's keep during his period of service. Early European visitors to Polynesia who were able to analyze the importance of reciprocal exchange and put it to their own use generally fared quite well among the Polynesians, although they sometimes found themselves overwhelmed by the increasing size of the obligations they had undertaken. Captain David Porter of the U.S. Navy, for example, in 1813 won the friendship of the chief of Taiohae, Nuku Hiva, with a gift of sperm whales' teeth but within a few months found himself fighting a tribal war essentially on behalf of the same chief, in order to live up to his obligations.

RELIGION AND ART

Religion and magic pervaded all aspects of Polynesian culture. All things, animate and inanimate, were believed to be endowed to a greater or lesser degree with mana, or sacred supernatural power. This power could be nullified by various human actions. Polynesian chiefs had great mana—so great, in fact, that in some islands if a commoner touched the chief's shadow the injury to the chief's mana could only be compensated by the death of the commoner. In many Polynesian cultures even today, it is still considered to be in very poor taste to step over a person's legs, pass one's hand over a person's head, or stand with one's head higher than that of a person of high rank. These actions are believed to sap a person's mana, rendering him profane.

The sacred power of places

It was not only people who had mana, however, but buildings, stones, tools, canoes, and all things. Polynesian life was infused with a wide variety of complicated rules designed to prevent damage to the mana inherent in various things. Groves, trees, temples, or tracts of land were considered sacred and could not be entered by ordinary people because they were pervaded by the mana of a high-status person or god. If anyone inadvertently stepped over a tool left on the ground, it was thus rendered profane and would often be discarded. Women were not permitted in canoes under normal conditions in the Marquesas because their presence defiled the canoe. In many societies, men preparing for war or for any other hazardous or demanding undertaking had to go through a period of purification—avoiding the company of women, eating certain foods only, and often going into seclusion so as to protect their powers from defilement. Some chants and songs were so sacred that every syllable had to be pronounced correctly. The penalty for major violations of these tapu, or prohibitions, was often death, swiftly executed and gruesome. Violations of lesser tapu, such as trespassing in a sacred grove or disturbing the bones of the dead, were believed to result in supernatural punishment, manifested in some form of illness or in bad luck or debilitation.

Mana was not the only force that the Polynesian had to be wary of, however. The universe, in their beliefs, was peopled with spiritual beings of various types, many of whom were malevolent. In addition, there was a host of gods of various degrees of importance, ranging from the great gods of the Polynesian pantheon, such as Tangaroa, Tu, and Lono, to strictly local gods who were deified priests or chiefs of great renown. All of these gods had to be worshiped in their own way, the most important by full-time priests of highest status, those of lesser importance by part-time priests. Worship of the gods involved sacrifices (including humans), chants and recitations, feasting (often with great prodigality), sexual orgies (to promote fertility), and elaborate rituals, often preceded by long fasting and abstinence.

Magic flourished in Polynesian society, with everyone

practicing various rituals for success in love, war, planting, or fishing, or to bring misfortune to rivals. There were also magical specialists who could be resorted to when the problem at hand was too great or complicated to be solvable by ordinary magic. Such magical practices have survived in Polynesian cultures right down to the present time.

The numerous restrictions and demands placed upon the average Polynesian by religious observance and by measures required to protect against supernatural forces were in many cases seen as oppressive by the Polynesians themselves, and, in one case, the whole system was simply abolished. This occurred in Hawaii in 1819, when the young king, Kamehameha II, assisted by the high priest, his adviser, and other leaders of the court, publicly and intentionally broke the main food and eating tapu at a festival, after which the high priest desecrated the images of the gods and burned the temple. While many factors played a part in the prelude to this dramatic act (not the least of which were the presence of Europeans and the desires of the royal women to be free from the burden of tapu), the principal factor motivating Kamehameha II seems to have been a real impatience and frustration at an increasingly constricting system that provided little obvious benefit in return for the great inconvenience. Although no other islands in Polynesia followed the Hawaiian example, there is evidence that the religious system was viewed as extremely frustrating in many other Polynesian cultures.

Polynesian art was highly developed and brilliant, but little of what is known to have existed has survived down to the present day. This is mainly due to the perishability of much of the material (bark cloth, basketry, featherwork) and the dispersion that took place during the era of European contact, when such items were bartered for firearms, liquor, iron tools, or trinkets.

Each Polynesian society developed its own particular area of artistic endeavour—monumental stone sculpture on Easter Island and the Marquesas; highly decorated bark cloth in western Polynesia; fine mats in Tonga. In this development, a vocabulary of art motifs, styles, and artistic principles was elaborated, which also differ somewhat from culture to culture. Certain types of motifs nevertheless are widely distributed in Polynesia. For example, a number of small geometric decorative elements, such as a toothed pattern or units of diagonally sloping lines, are found in most cultures in many media. Stylized floral and animal elements are also widely distributed. In eastern and marginal Polynesia, an anthropomorphic figure and an often grotesque anthropomorphic face with bulging eyes and protruding tongue are prominent in all types of plastic art, as well as in tattooing.

Architecture was developed almost everywhere in Polynesia, but only the ruined stone structures of the eastern Polynesian islands remain to bear witness to the Polynesians' considerable architectural skill. The stone temples of the Society Islands, Easter Island, and the Marquesas, even in their ruined state, display striking proportions and design; clever use of stones of different colours, shapes, and textures; and the evidence of sound combination of practical engineering with artistic objectives.

(R.C.Su./R.C.Ki.)

EFFECTS OF EUROPEAN CONTACT

Polynesian cultures have been radically altered by contact with the Western world. European explorers navigated much of the area in the latter quarter of the 18th century, and the first missionaries arrived in the late 1700s and early 1800s. The colonial partitioning began in 1842 when France annexed Tahiti. Eventually, Chile, Germany, New Zealand, the United Kingdom, and the United States held territory in the area. Germany lost its interests in the islands as a consequence of World War I.

Colonial partitioning

The period between the two world wars was tranquil, although there was some open resistance to New Zealand's administration of Western Samoa under a League of Nations mandate. The missions' influence on the Polynesian peoples increased, and Christianity became an integral part of the islanders' lives. In many areas Christianity in turn was influenced by local traditions and customs. Quite commonly, villages competed to build larger and more elaborate churches, and first-time visitors to Polynesia are often surprised at the intensity of the islanders' commitment to Christianity. Many Polynesians were recruited to missionize other parts of the Pacific, particularly Melanesia.

After World War II, sentiments for decolonization began to spread, and Western Samoa became the first independent Pacific nation when it gained sovereignty from New Zealand in 1962. Tonga freed itself as a protectorate of the United Kingdom in 1970, and the Ellice Islands, in the Gilbert and Ellice Islands Colony of Great Britain, achieved independence as Tuvalu in 1978. Ironically, the conservatism of Polynesian cultures is most manifest in these three independent nations. Three island groups had been monarchies—Tonga, Tahiti, and Hawaii—but monarchy survives only in Tonga, although Tonga also has a British-style parliament that gives special status to traditional nobles.

Monarchy in Tonga

Having been a mandate of New Zealand, a Commonwealth nation, Western Samoa also has a parliamentary system, but only traditional chiefs (*matai*) may vote and run for election. Tuvalu also follows the parliamentary style of government. In Tonga, Samoa, and Tuvalu, the vast majority of people live in subsistence-based villages, where much of traditional culture survives.

The Cook Islands and its former colonial authority created a new political status in 1965 when the Cooks became self-governing in free association with New Zealand. In this innovative arrangement, the Cooks became self-governing with regard to internal affairs, and New Zealand agreed to manage external affairs as requested. New Zealand provides the government of the Cook Islands with a financial subsidy, and Cook Islanders enjoy dual citizenship. As the relationship has subsequently evolved, the Cook Islands has assumed control of most of its external or foreign affairs. Niue Island, another former dependency of New Zealand, followed suit in 1974 when it also became self-governing in free association with the former colonial power. Many of the people of the Cook Islands and Niue have become very Westernized, and many of their citizens as well as those of Western Samoa have immigrated to New Zealand. Indeed, Auckland, N.Z., has the largest concentration of Polynesians of any city or place in the world.

Six other Polynesian entities remain parts of or are territories of four Western powers. Of these, two—French Polynesia and the self-governing territory of Wallis and Futuna—are integral parts of France and vote in French national elections. French Polynesia includes several island groups, but a large majority of the people live in or near the capital city of Papeete on the island of Tahiti and are quite urbanized. Wallis and Futuna is small and remote, and much of its traditional culture survives in a diluted form. Tokelau, New Zealand's last dependency in the Pacific, has small islands and a very small population. Like Wallis and Futuna, it is remote, and much of its former way of life remains.

Association with Western powers

Two island groups are American: Hawaii and American Samoa. Hawaii is the 50th state of the United States. The city of Honolulu, with its large, Western-style commercial district and sprawling suburbs, dominates the island of Oahu, has most of the state's population, and has much in common with cities on the U.S. mainland. American Samoa is a territory of the United States, and it is far less traditional than Western Samoa. The capital town of Pago Pago is a blend of both American and Samoan life-styles. Rural villages, however, have much in common with their counterparts in Western Samoa.

Easter Island is the anomaly of the area. The aboriginal population was decimated by European-introduced diseases and by slavers in the 1860s and almost became extinct. In 1888 Easter Island was annexed by Chile, and its people are the only Pacific islanders controlled by a Latin-American power. Little remains of the traditional culture of Easter Island. The indigenous Polynesian language survives, but most people also speak Spanish. About one-third of Easter Island's small population is from Chile.

(R.C.Ki.)

MAJOR POLITICAL UNITS OF MELANESIA

(As mentioned at the beginning of this article, Australia and New Zealand are treated separately, and Hawaii is treated in the article UNITED STATES OF AMERICA. Other political and geographic units of the Pacific not covered below are treated individually in the *Micropædia.*)

Fiji

Fiji is an independent nation consisting of an archipelago surrounding the Koro Sea in the South Pacific Ocean, about 1,300 miles (2,100 kilometres) north of Auckland, N.Z. It forms part of the Melanesian cultural area.

The archipelago consists of some 300 islands and 540 islets scattered over about 1,000,000 square miles (3,000,000 square kilometres). Of the 300 islands, about 100 are inhabited. The total land area is 7,056 square miles. The capital, Suva, is on the southeast coast of the largest island, Viti Levu ("Great Fiji"). Fiji includes within its borders Rotuma, an island of 18 square miles located 440 miles north of Suva; Rotuma has a Polynesian population and was administered from Fiji in colonial times.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Fiji has a complex geologic history. Based on a submerged platform of ancient formation, the Fiji islands are largely the product of volcanic action, sedimentary deposit, and formations of coral. Viti Levu, the largest island, has an area of about 4,000 square miles and accounts for more than half of Fiji's land. A jagged dividing range running from north to south has several peaks above 3,000 feet (914 metres), the highest being Mount Tomanivi (formerly Mount Victoria) at 4,341 feet. The main river systems—the Rewa, Navua, Sigatoka, and Ba—all have their headwaters in the central mountain area. To the southeast and southwest, and to the south where the range divides, the mountains give way to plateaus, then lowlands. The coastal plains in the west, northwest, and southeast account for only 15 percent of Viti Levu's area but are the main centres of agriculture and settlement.

Vanua Levu, the second largest island, has an area of about 2,140 square miles. It is divided along its length by a mountain range with peaks rising to more than 3,000 feet. On the island's northern coast, away from the mouth of the Dreketi River, the coastal plains are narrow. Most of the other islands, including the Lomaiviti, Lau, and Yasawa groups, are volcanic in origin, but, like the major islands, they are bounded by coral reefs, offshore rocks, and shoals that make the Koro Sea hazardous for navigation.

Climate. At Suva the average summer high temperature is 85° F (29° C) and the average winter low is 68° F (20° C); temperatures typically are lower in elevated inland areas. All districts receive the greatest amount of rainfall in the season from November through March, during which time hurricanes are also experienced perhaps once every two years. While rainfall is reduced in the east of the larger islands from April to October, giving an annual average of 120 inches (3,050 millimetres) per year, it virtually ceases in the west, to give an annual rainfall of 70 inches, thus making for a sharp contrast in both climatic conditions and agriculture between east and west.

Rainy season

Plant and animal life. Almost half of Fiji's total area remains in forest, while dry grasslands are found in western areas of the large islands. Coconut palms are common in coastal areas, and almost all tropical fruits and vegetables can be grown. Much of the shoreline is composed of reefs and rocks, while mangrove swamps are found on eastern coasts. There are few white-sand swimming beaches and, because of the encircling reef, little surf.

Most animals, including pigs, dogs, cattle, and a few horses, are domesticated. Mongooses, introduced to prey on snakes and rats, are often seen.

The people. While the indigenous Fijian people are usually classified as Melanesian, they are larger in stature than Melanesians from Vanuatu, Solomon Islands, or New Guinea; their social and political organization is closer to that of Polynesia; and there has been a high level of intermarriage between Fijians from the Lau Islands of eastern Fiji and the neighbouring Polynesian islands of Tonga. Almost all indigenous Fijians are Christian, mostly Methodist and Roman Catholic.

Since World War II, indigenous Fijians have been outnumbered by Indians, most of whom are descendants of indentured labourers brought to work in the sugar industry. A few, particularly in commerce and the professions, are descended from free migrants. Most of the Indians are Hindus, though a significant number are Muslims.

There are also significant minorities of Europeans, part-Europeans, Chinese, and Pacific islanders from outside Fiji. In the last group are the Polynesian population of Rotuma and the Banabans, who were forced to leave Banaba after destruction in World War II made it uninhabitable. Many Banabans settled on Rabi Island in Fiji.

English is the official language. The widely used Fijian language has many dialects; the one most commonly used is known as Bauan Fijian and comes from Bau, the district that enjoyed political supremacy at the advent of colonial rule. Most people speak at least two languages, including English and the language of their own racial community.

There is very little intermarriage between racial communities. The relative proportions of Fijians and Indians in the population have been changing in recent years because the Fijian birth rate is higher than the Indian and because of the accelerating outward migration of Indians, especially to Canada, Australia, and New Zealand. While Suva has a very mixed population, the sugar regions of Viti Levu and Vanua Levu have predominantly Indian populations. On the smaller islands, and in less-developed rural areas of the larger islands, Fijians live in traditional villages. The two largest urban centres are on Viti Levu: Suva, in the southeast, which has about one-fifth of the total population, and Lautoka in the northwest, which is the centre of the sugar industry and has a major port. Labasa on Vanua Levu is a centre for administration, services, and sugar production.

Indian settlements

The economy. Fiji has an agriculture-based market economy, including a substantial subsistence sector dominated by indigenous Fijians who earn a supplementary cash income from copra, cocoa, kava, dalo (taro), pineapples, cassava, or bananas, or from fishing. The commercial sector is heavily based on sugar, which, for the most part, is produced by independent Indian farmers. Sugar production is concentrated on the western side of Viti Levu and

© S. Chester/Comstock



Indian farmers transporting a load of sugarcane near Natandola, Viti Levu, Fiji.

in the area around Labasa. The Fiji Sugar Corporation has a monopoly on milling and marketing. Sugar is by far Fiji's largest export and accounts for more than one-half of all exports.

Except for a few years early in the 20th century, the alienation of native land has been prohibited since 1874, thus leaving 83 percent of all land under Fijian ownership. Farmers of other races operate on leaseholds of up to 30 years under the Agricultural Landlord and Tenant Act. Fijian land ownership is in the hands of *mataqali,* or clan groups, but may be administered through the Native Lands Trust Board.

Tourism is almost as large an earner of foreign exchange as sugar. Fiji is strategically located for air travelers from Australia, New Zealand, the United States, and Japan and is a major destination for tourist cruises. Tourism is based on the attractions of duty-free shopping and colourful handicraft markets as well as the usual attractions of tropical islands. Many hotels are located on small offshore islands or secluded beaches and offer accommodations in houses of local design and materials rather than in urban-style multistory buildings.

The third major export earner, though well behind sugar and tourism, is gold, which is mined at Vatukoula in northern Viti Levu. Copper deposits are known to exist at Namosi, inland from Suva, but mining is not viable.

Since large-scale systematic planting of pine forests began in the 1960s, a timber industry has developed for domestic use and export. Development plans have emphasized the need to reduce dependence on imported food, especially rice, meat, fish, and poultry products. There is substantial hydroelectricity generation, but fuel remains a major import together with manufactured goods (many for resale to tourists), machinery, and food. Australia, New Zealand, Japan, and the United States are the major sources of imports.

The economy has a strong service and light industrial component serving small neighbouring countries as well as Fiji; activities range from boat building (especially fishing boats and pleasure craft) to brewing and paint manufacture. The government offers incentives (including residence) for investors but insists on potential for job creation and training programs for local employees.

Sugar is sold on long-term contracts to New Zealand, Malaysia, and Singapore and, under the Lomé Convention (which gives former overseas territories access to the European market), to the European Economic Community. Regional trading agreements give Fiji special access to markets in Australia and New Zealand.

Transportation

The larger islands and many smaller ones are served by domestic air services. International airports are located at Nausori, near Suva (mostly for regional flights), and Nadi, on the western side of Viti Levu. A coastal highway circles Viti Levu, and minor roads to the interior give access to most areas of settlement. For many villagers, however, river punts with outboard motors provide the most efficient form of transport, while from more remote areas it may still be simplest to transport produce to market by floating it downriver on bamboo rafts. Regular bus services operate within and between the major towns.

Administration and social conditions. *Government.* Until the coup d'état of May 1987, Fiji was a dominion, a member of the Commonwealth, and a parliamentary democracy, acknowledging the British sovereign, through a governor-general, as head of state. In October 1987 Fiji officially left the Commonwealth and became a republic. The coup leader, then-Lieutenant Colonel Sitiveni Rabuka, appointed a civilian government headed by a president with a largely ceremonial role. The government was composed of a prime minister and cabinet of appointed members, almost all of whom were ethnic Fijians. On July 25, 1990, a new constitution, which concentrated power in the hands of Fijians, was promulgated. Membership in the House of Representatives was raised from 52 to 70, with 37 of the seats reserved for Fijians, 27 for Indians, five for other races, and one for Rotumans. Membership in the Senate was increased from 24 to 34, with 24 of the seats reserved for Fijians, nine for Indians and other races, and one for Rotumans. The constitution also stipulated that the prime minister must be a Fijian.

Local government reflects the plurality of Fiji's social structure. There are elected multiracial councils in the larger towns, a separate Fijian administration incorporating a hierarchy of chiefs and councils for the control of rural Fijians, and direct administration elsewhere.

Before the first of two coups in 1987, Fijian military forces had a largely ceremonial role, though they bore much of the burden of rebuilding and organizing after natural disasters and of civilian development projects. Military forces continued to perform these services after the coup, with the added role of agricultural distribution, together with their major preoccupation with the enforcement of internal security policies.

Education and medical facilities. While the government provides some primary and secondary education, most schools are controlled through local committees run by and for a single racial or religious community. Entry to secondary schools is by competitive examination. Students pay fees but not the full cost of their education, which is subsidized by the government. The University of the South Pacific, near Suva, is a regional institution; Fiji and other Pacific Islands governments fund the budget, and foreign aid meets the costs of buildings and capital development. To extend the reach of the university, lessons are broadcast to distant regional centres by a satellite network. Fiji also provides for its own technical, agricultural, and medical education and teacher training. There are private medical practitioners in all large towns, a national network of clinics and small hospitals, and major hospitals in Suva, Lautoka, and Labasa.

The University of the South Pacific

Effects of urbanization. Rapid urbanization, especially on the fringes of Suva, has seen the emergence of squatter settlements and has caused some social problems. The disparities of income between urban and rural workers, contrasting life-styles within the urban areas, and high urban unemployment can be related to an escalating rate of crime and to the rapid growth of a trade union movement.

Cultural life. Fiji's mixed racial background contributes to a rich cultural heritage. Many features of traditional Fijian life survive; they are most evident in the elaborate investiture, marriage, and other ceremonies for high-ranking chiefs. These ceremonies provide a focus for the practicing of traditional crafts, such as the manufacture of *masi,* or tapa cloth, made from the bark of the paper mulberry; mat weaving; wood carving; and canoe making. Drinking of yanggona (kava, made from the root of *Piper methysticum*) is a part not only of important ceremonies but also of everyday life. Displays of "traditional" Fijian culture, music, and dancing make an important contribution to tourism; model villages and handicraft markets are popular.

Most Indian women continue to wear the sari together with traditional jewelry in gold and silver. Traditional marriage ceremonies are practiced, as are customs such as fire walking and ritual self-torture as part of important religious ceremonies. Cinemas showing imported Indian films are popular. Dīwālī, the Hindu Festival of the Lights, is celebrated every October and is a public holiday.

Fiji has two daily newspapers and a multilingual public radio broadcasting system. Videocassette players are common in the towns (many villages have no electricity).

For statistical data on the land and people of Fiji, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

When Fiji's first settlers arrived from island Melanesia at least 3,500 years ago, they carried with them a wide range of food plants, the pig, and a style of pottery known as Lapita ware. This pottery is generally associated with peoples who had well-developed skills in navigation and canoe building and were horticulturists. From Fiji the Lapita culture was carried to Tonga and Samoa, where the first distinctively Polynesian cultures evolved. Archaeological evidence suggests that two other pottery styles were subsequently introduced into Fiji, though it is not clear whether these represent major migrations or simply cultural innovations brought by small groups of migrants.

Two types of settlement

In most areas of Fiji, the settlers lived in small communities near ridge forts and practiced a slash-and-burn type of agriculture. In the fertile delta regions of southeast Viti Levu, however, there were large concentrations of population. These settlements, which were based on intensive taro cultivation using complex irrigation systems, were protected by massive ring-ditch fortifications.

Traditional Fijian society was hierarchical, with tribal groups led by chiefs chosen according to rank based on descent as well as personal achievement. Tribes were divided into *mataqali*, or clans, and residential subclans. By alliance or conquest, tribes might form confederations led by paramount chiefs; warfare was common.

The first Europeans to sight the Fiji Islands were Abel Janzsoon Tasman, who passed the northeast fringe of the group in 1643, and Captain James Cook, who passed the southeastern islands in 1774. Captain William Bligh passed through the group in his open longboat after the mutiny on HMS *Bounty* in 1789 and returned to explore it in 1792.

Commercial interest in the group began with the discovery of sandalwood at the beginning of the 19th century, leading to a rush to Bua Bay at the southwestern end of Vanua Levu. A few beachcombers, useful as armourers and interpreters, were adopted by influential chiefs from this time. Within little more than a decade the accessible, commercial stands of sandalwood were depleted, but by the 1820s traders were again visiting the group to trade for edible varieties of the marine invertebrate called the sea cucumber, also known as bêche-de-mer or trepang. Whereas most of the sandalwood had been cut by gangs of foreigners, the bêche-de-mer harvest involved large numbers of Fijians in gathering, cleaning, and drying and in the provision of food and firewood.

These opportunities for new wealth and power, symbolized by the acquisition of muskets, intensified political rivalries and hastened the rise of the confederation of Bau, led first by Naulivou, then by his nephew Cakobau. Bau is a tiny island off the southeast coast of Viti Levu, but by the 1850s it dominated western Fiji. Cakobau's main rival was the Tongan chief, Ma'afu, who led an army of Christian Tongans and their allies from eastern Fiji. After a short-lived alliance with Ma'afu, Cakobau became a Christian in 1854, thus bringing most Fijians under the influence of Methodist missionaries. Roman Catholic and Anglican missionaries arrived later but did not enjoy the same success.

By the 1860s Fiji was attracting European settlers intent on establishing plantations to capitalize on a boom in cotton prices caused by the American Civil War. The consequent disputes over land and political power within and between European and Fijian communities, and problems with labour introduced from other Pacific islands, contributed to violent confrontations and exacerbated the implicit instability of Fijian society and ensured that no Fijian chief could impose his rule on the whole group. European attempts at government were doomed by the greed and factionalism of their members and by the interference of European governments and consuls. Imperial intervention thus became inevitable. On Oct. 10, 1874, after negotiations had led to an offer of unconditional cession, Fiji became a British crown colony. The policies of the first governor, Sir Arthur Gordon, were decisive in shaping the history of Fiji. Gordon saw himself as the protector of the Fijian people and thus initiated policies that limited their involvement in commercial and political developments. Sales of Fijian land were banned; the Fijians were taxed in agricultural produce, not cash; and they were governed through a system of indirect rule based on the traditional political structure.

Crown colony status

In order to maintain these policies yet encourage the economic development of the new colony, Gordon promoted the introduction of indentured Indian labourers and investment by the Colonial Sugar Refining Company to establish sugar plantations and processing mills. Indian migrants were encouraged to become permanent settlers at the conclusion of their contracts, even though little land was available for sale and the migrants' political rights were circumscribed. After the termination of the indenture system in 1920, Indian agitation over political and economic grievances caused strikes and continual discontent and challenged the commercial and political domination of the small European community in the islands.

During World War II Fiji was occupied by Allied forces, and a battalion of Fijians saw service as scouts in the campaign for the Solomon Islands. Indians refused to serve because they were not offered the same wages and conditions as Europeans. Indians also refused to cut their cane at the low prices offered. These actions led to the taint of disloyalty being applied to Indians by the other races, and to the army, which was retained after the war, remaining exclusively Fijian except for a handful of European officers. Incipient racial awareness was heightened by the realization that, from 1943, Fijians were outnumbered by Indians. As one response to perceived stagnation in Fijian society, the colonial authorities restructured the Fijian administration, reinforcing chiefly leadership and thus consolidating the conservatism of Fijian society.

Constitutional development toward independence, which began in the 1960s, was more a response to international and British pressures than to any demand from within Fiji. The steps taken were, in practice, a compromise between the principles of parliamentary democracy and the racial divisions within the country. The franchise, previously exercised by Europeans and some Indians, was extended to adults of all races, including Fijians, who until then had been represented by their chiefs. Fijian land rights, guaranteed by the Deed of Cession in 1874, were given constitutional protection, while Fijian chiefs were given an effective veto on all important matters affecting the status of Fijians and on changes to the constitution itself. Facing political reality, Indian leaders relinquished their earlier demands and accepted the communal and cross-voting system, thus giving away the common electoral role that they had sought since the 1930s.

The effect of the constitution

The effect of the constitution was to give power to Fijian politicians so long as they remained in partnership with the General voters and, critically, so long as the Fijian vote remained unified. Despite race riots during by-elections in 1968, independence was achieved in a spirit of cooperation on Oct. 10, 1970, the 96th anniversary of cession.

From that time until April 1987, Fiji was governed by the Alliance Party, which was pledged to policies of multiracialism. Its electoral supremacy was challenged only briefly, in 1977, when Fijian votes were attracted by Fijian nationalist candidates campaigning under a slogan of "Fiji for the Fijians"; only factionalism prevented the formation of an Indian-led government.

In 1987, however, the Indian-dominated National Federation Party joined in coalition with the new Labour Party (led by a Fijian, Timoci Bavadra), which had strong support from Fijian and Indian trade unionists. The coalition was successful in elections held in April. The new government, which had a majority of Indian members in the legislature, was greeted with widespread Fijian protest. After only a few weeks the new government's leaders were arrested and deposed in a coup d'état led by Lieutenant Colonel Sitiveni Rabuka, who demanded greater protection for Fijian rights and an entrenched Fijian dominance of any future government. The governor-general declared a state of emergency and assumed control of the government. With political leaders, he then negotiated a compromise that would have maintained civilian rule pending a constitutional revision and new elections. The compromise prompted Rabuka to lead a second coup in September and reimpose military rule. Toward the end of 1987 he declared Fiji a republic and appointed a new civilian government with Ratu Sir Kamisese Mara, prime minister from 1970 to 1987, as prime minister. A new constitution, designed to concentrate power in the hands of Fijians, was promulgated on July 25, 1990. (B.K.M.)

For later developments in the history of Fiji, see the BRITANNICA BOOK OF THE YEAR.

New Caledonia

New Caledonia (official name Territoire de la Nouvelle-Calédonie et Dépendances, or Territory of New Caledo-

nia and Dependencies) is a French overseas territory in the southwest Pacific Ocean, about 900 miles (1,500 kilometres) east of Australia. It includes the island of New Caledonia (the Grande Terre, or mainland), where the capital, Nouméa, is situated; the Loyalty Islands; the Bélep Islands; and the Isle of Pines. These islands, which form more than 99 percent of the total land area of about 7,300 square miles (19,000 square kilometres), lie between latitudes 18° and 23° S and longitudes 163° and 169° E. The territory also includes a number of far-flung uninhabited islets: Huon and Surprise islands in the D'Entrecasteaux Reefs, the atolls of the Chesterfield Islands and Bellona Reef, Walpole Island, Beautemps-Beaupré Atoll, and Astrolabe Reefs. France also claims Hunter and Matthew islands, but the claim is disputed by Vanuatu. Although New Caledonia has a relatively small population, it has acquired significance as a remaining outpost of European sovereignty in a largely decolonized region.

The main island, with an area of 6,500 square miles, is by far the largest island and contains about 90 percent of the population. It is surrounded by a coral reef that extends from the Huon Islands in the north to the Isle of Pines in the south. Except for the central part of the west coast, which is bordered only by a fringing reef, it is a true barrier reef enclosing a large lagoon. There are numerous passages in the reef, usually at the mouths of rivers.

Metamorphic rock formations

Relief and drainage. Rugged mountain ranges, consisting principally of metamorphic rock formations, divide the island into an east coast, which in many places descends precipitously to the sea, and a west coast, which slopes more gradually and contains flat, undulating land. Ultrabasic, serpentine rock forms a continuous plateau over most of the southern third of the island, rising to a height of 5,308 feet at Mount Humboldt, and continues along the west coast as a series of discrete massifs. Outcrops from this formation form the islands of Art and Pott in the Bélep archipelago in the north and, in the south, the central part of the Isle of Pines, which is bordered by an emerged coral platform. These rocks have weathered to form the striking *terre rouge* (*i.e.,* red soils that overlay the island's extensive deposits of nickel, chrome, and cobalt ore). In the northeast an outcrop of gneiss forms a 40-mile-long mountain range that includes the territory's highest mountain, Mount Panié, 5,341 feet. Elsewhere, the northern half of the island consists mainly of an irregular series of ranges formed from schists. Sedimentary rocks are limited to a narrow zone extending along much of the west coast inland between the serpentine massifs and the northern schist formations. They have weathered to produce broad, undulating plains with some steep-sided hills.

Numerous streams descend from the central mountain chain to the lagoon, often flooding rapidly after rainfall and drying out in dry weather, especially on the west coast. The Diahot River, the longest river in the territory, flows for about 60 miles toward the northern tip of the island along the western escarpment of the Mount Panié range.

The Loyalty Islands consist of three main islands—Ouvéa, Lifou, and Maré—and numerous small islands, the most important being Tiga. The Loyalty Islands form more than 10 percent of New Caledonia's total land area, and about 10 percent of the population lives there. In contrast to the mainland, these islands are raised coral plateaus, nowhere rising more than 330 feet above sea level. Surface water is lacking because of the porous nature of the calcareous rock formation.

Climate. The climate is subtropical with a year-round rainfall. The rainfall is higher on the east coast, where at higher elevations more than 120 inches (3,000 millimetres) may fall annually. On the west coast the rainfall is regularly less than 40 inches. The period from December to March is particularly rainy because of equatorial depressions, including frequent cyclones. Another period of heavy rainfall occurs in July and August; the driest months are September through November. Winds bearing northeast to southeast, including trade winds, predominate throughout the year and relieve temperatures in the hot season, which begins in November. Cyclonic winds are frequent late in the hot season.

The mean annual temperature at sea level ranges between 71° F (22° C) and 75° F (24° C). In the south of the mainland there are very few days when the temperature rises above 86° F (30° C), but their frequency increases to the north. The lowest temperature recorded in Nouméa is 55° F (13° C), but farther north on the west coast temperatures as low as 41° F (5° C) have been known to occur.

The influence of red soils on flora

Plant and animal life. Geographic isolation, contrasting soils, and a wide range in altitudes have produced a rich indigenous flora. *Terre rouge* soils support a highly original maquis in which sclerophyllic (resistant to drying) shrubs with spectacular coloured flowers are dominant. Different forms of rain forest range from those growing on coral platforms, as in the Loyalty Islands, to montane forests above 3,000 feet on the mainland. The savanna woodlands of the west coast are dominated by a ti tree, the ubiquitous niaouli (*Melaleuca quinquenervia*), which is highly fire-resistant and the spread of which is stimulated by bushfires. Although it grows best in wet soils up to an altitude of 2,000 feet, it extends also onto well-drained slopes and crests and forms the dominant species of closed swamp forests growing on the seaboard plains of the Diahot valley. A once-widespread dry sclerophyll forest, dominated by the guaiacum (*Acacia spirorbis*), growing at low altitudes on the west coast, is now confined to relict sites. Mangrove swamps proliferate on the highly indented west coast.

Except for several species of bat, which were present before the arrival of Europeans, mammals are absent from the native fauna. There are no frogs and no venomous land reptiles, although scorpions and centipedes are present. There are no endemic malaria-carrying mosquito species. The kagu, a flightless bird, is the most unusual of some 100 endemic bird species and is now rare. A wide range of marine life is present in the lagoon.

Settlement patterns. About three-fifths of the population live in Nouméa, which since 1965 has expanded to embrace the adjacent municipalities of Dumbéa, Mont-Dore, and Païta. Nouméa has the facilities of a French provincial city, including numerous bars and restaurants, modern shops and supermarkets, a hospital, schools, a local newspaper, and radio and television facilities. About four-fifths of people of migrant origin, including Europeans, Polynesians, and Asians, live there as compared with one-quarter of the Melanesian population. About three-quarters of the Melanesians live outside Nouméa in small, widely dispersed villages with few modern facilities. They engage chiefly in subsistence agriculture based on the cultivation of the yam, taro, sweet potato, and banana. The population is almost entirely Melanesian in the Loyalty Islands, the Isle of Pines, and the Béleps and on the east coast and in the mountain ranges of the mainland.

The people. Melanesians and Europeans are the two principal ethnic groups, forming 43 and 37 percent of the population, respectively. Their differing cultures have given rise to two distinct ways of life, known as *kanak* and *caldoche;* people of mixed descent tend to ascribe to one or the other. The *kanak* identity is based on clan membership, a network of family alliances and specific land rights. The *caldoche* way of life is essentially integrated into a cash economy. The Polynesian minority comprises Wallis and Futuna Islanders (9 percent) and Tahitians (4 percent). Descendants of Indonesian and Vietnamese migrant workers form 4 and 2 percent of the population, respectively, and are essentially an urban group.

French is the official language but is not the mother tongue of most Melanesians and Polynesians. Some 30 Melanesian languages are spoken, most Melanesians being proficient in more than one.

Religions

The Roman Catholic Church claims about two-thirds of the population as adherents, including almost all European, Uvean, and Vietnamese and half of the Melanesian and Tahitian minorities. Two Protestant churches, the Evangelical Church in New Caledonia and the Loyalty Islands (Église Evangélique en Nouvelle-Calédonie et Îles Loyauté) and the Free Evangelical Church (Église Libre), serve one-fifth of the population. Their memberships are almost entirely Melanesian. There are also numerous minor Christian sects and small numbers of Muslims.

The Melanesian population is widely believed to have

declined rapidly in the 19th century, but estimates for the period, including that for the 1887 census, were so poorly based that this remains an impressionistic view. In the 20th century the figure for Melanesians was fairly stable up to 1939 but doubled in the 40 years following World War II. Migration into and out of the territory plays an important role in determining the numbers of the non-Melanesian communities. The birth rate is higher among Melanesians and Uveans than among other groups, but infant mortality is high among Melanesians, and life expectancy is relatively low.

The economy. New Caledonia's economy is highly dependent on France. Exports, which consist largely of partly refined nickel and nickel ore, cover less than half the cost of imports but vary with the world market price of nickel. Half of the territory's trade is with France; other important trade partners are Japan, the United States, and Australia. The trade deficit is made up by the transfer of funds from France, chiefly as grants to the territorial budget and expenditures for health, education, and the maintenance of military and security forces, which accounts for about one-third of the gross domestic product (GDP). The administration, commerce, and transport and services sectors contribute three-quarters of the GDP, compared with 9 to 12 percent for mining and 2 percent for agriculture. Efforts to diversify the economy beyond the commercial and administrative sectors have met with little success. There is, however, considerable potential for tourism, relying on tourists from Australia, New Zealand, and Japan. Import replacement industries, such as the manufacture of soft drinks and beer, soap, cement, fencing wire, and fishing and pleasure boats, have had little impact on the economy because of the small local market.

Although the per capita GDP is one of the highest in the countries of the South Pacific, distribution between ethnic groups is unequal, Melanesian household revenues being on average only about one-quarter of that of European households. The distribution of land resources on the mainland also is very uneven. Although some 35,000 Melanesians live by cultivating the land, two-thirds of the land is in the hands of European families, very few of whom are engaged in agriculture or cattle raising.

Europeans also dominate trades, businesses, and professions and hold most of the high-ranking administrative posts in the government. Among the official unemployed, Melanesians are twice as numerous as Europeans; figures do not include the considerable number of "hidden" unemployed who have returned to their villages.

Taxes in the territory consist primarily of duties on imported goods, sales taxes, and taxes on business revenues. An income tax was introduced in 1982 but has been maintained at a deliberately low level. More than 90 percent of the total tax receipts come from within Nouméa.

Mineral reserves

Resources. The weathering of serpentine rock provides the territory with 20 percent of the world's known reserves of nickel ore, as well as large deposits of chromium, cobalt, iron, and magnesium. The export of nickel ore, which has been mined since 1875, and of partly refined nickel from the Société Le Nickel (SLN) foundry on the outskirts of Nouméa, is subject to cycles of booms and depressions determined by the needs of the world steel industry and competition from other producers. It is a large industry by world standards. A large chromium mine at Tiébaghi in the north of the mainland was reopened in 1982 after a long period of inactivity to export ore to China and the United States. Cobalt and iron-ore deposits, as well as deposits of gypsum on the west coast and of phosphates on outlying islands, are no longer worked. Noncommercial deposits of coal are found on the west coast. The search for oil has not been successful. Hydroelectric power from Yaté and Néaoua provides nearly half of the territory's energy needs; the remainder of the territory's power is produced with imported fuel oil. More than three-quarters of the energy produced is used in the refining of nickel.

Agriculture. Local agricultural products meet barely half of the territory's needs for meat, vegetables, and fruit. Commercial agriculture has not succeeded despite efforts to establish sugarcane, cotton, rice, coffee, and copra industries. The production of coffee and copra that began in the 19th century was kept going after World War II chiefly through the efforts of Melanesian subsistence farmers to diversify into the cash economy, but very little is now exported. A project began in 1982 to restart a coffee industry chiefly on Melanesian land by using varieties of coffee plants that are tolerant to sunlight. In the late 1970s reforestation projects, consisting mainly of plantings of Caribbean pine, were established on Melanesian land on the Isle of Pines and on the mountains inland from Koné on the west coast of the mainland. Similar projects have not been continued. Cattle raising is important to the economy; pigs and horses also are raised but rarely for commercial purposes.

Transportation. The main island of New Caledonia and the inhabited outer islands are ringed by roads. Virtually the entire coast of the island of New Caledonia may be traveled by road, and crossroads penetrate the centre of the island.

The port of Nouméa handles the majority of ocean traffic; a new deep-water harbour was completed in 1974. Regular service is available for cargo and passengers.

The domestic airline, Air Calédonie, provides internal air service from Magenta Airport near Nouméa to the main and outer islands, as well as to France and other countries.

Administration and social conditions. As a French overseas territory, New Caledonia is fully integrated into the French Republic, is subject to its laws, and sends parliamentary representatives to Paris. Special legislation is required, however, to introduce laws for the government of the territory and in areas such as education, health, and labour. Authority over defense, internal security, immigration, justice, radio and telecommunications, and secondary and higher education is retained by the French government and is administered by its appointed representative, the high commissioner. The administration of health, social security, transportation, industry, and infrastructure is shared by the high commissioner and locally elected representatives. Municipal and regional government is administered by the state.

French authority

Health and educational facilities of a high standard are in principle available to all but are of a higher standard in Nouméa than elsewhere. The school system closely follows the syllabus of schools in France, and students of non-European origin tend to progress more slowly than children of French origin. Although all attend school, attendance being compulsory until the 16th birthday, relatively few non-Europeans gain higher qualifications. French is the only language of instruction in state-supported schools.

Cultural life. Nouméa, with its bars, restaurants, cinemas, an excellent library and museum, good bookshops, and radio and television, offers a touch of Europe. For the Europeans, sports tend to be closely related to the sea and include boating, fishing, windsurfing, and swimming or sunbathing at the city beaches of Anse Vata and the Baie de Citron; tennis and cycling are also popular. Although big-league soccer in France is closely followed, soccer playing has virtually been appropriated by Melanesian men and a version of cricket (first introduced by early British missionaries in the Loyalty Islands) by Melanesian women. Melanesian soccer and cricket teams compete annually in territory-wide competitions.

In many areas, Melanesian custom remains strong, particularly in relation to clan and family ties and obligations. Almost without exception Melanesians, regardless of their education or urbanization, return to their villages to take part in elaborate ceremonies and gift exchanges on such occasions as births, marriages, and deaths. "Tradition" has, of course, been modified, an important inclusion now being Christian ceremonies. In rural areas, women still tend to carry out the daily agricultural round of planting, weeding, and harvesting, as well as their domestic tasks of cooking, cleaning, and child rearing. Men perform the heavier tasks of clearing the ground for new gardens. All take part in the annual yam planting and harvesting, which is still something of a ceremonial and social occasion, and these events also sometimes draw urban workers back to their villages for a short period.

Local languages remain surprisingly strong, despite the fact that, through the educational system, French has be-

come the lingua franca. In the 1980s, due to increasing nationalist feeling, there has been a revival of interest among Melanesians in the traditional arts of sculpture, mat and basket weaving, singing, dancing, and wood carving. This has been encouraged by the French government.

State-owned radio and television stations relay programs to all parts of the territory. In Nouméa there are also several privately owned radio stations. There is one daily newspaper, *Les Nouvelles Calédoniennes.*

For statistical data on the land and people of New Caledonia, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Melanesians settled the territory more than 3,000 years ago and, except for rare Polynesian voyagers, probably were cut off from outside contact until the late 18th century. In 1774 Captain James Cook landed at Balade, on the east coast of the mainland, and he named the island New Caledonia for his native Scotland. Cook was followed there by the French navigator Antoine de Bruni, chevalier d'Entrecasteaux, in 1793. Regular contact with Europeans began in 1841 when sandalwood traders from Australia introduced islanders to the use of iron. The arrival of a Protestant mission from the London Missionary Society in the Loyalty Islands in 1841 and a Marist mission, which was set up at Balade with the aid of the French navy in 1843, marked the beginning of the Protestant and Roman Catholic churches in the territory.

The French claim

France took possession of most of present-day New Caledonia in 1853 in ceremonies at Balade and the Isle of Pines with the aim of reserving the territory as a possible site for a penal colony. The French flag was not run up in the Loyalty Islands until 1864.

From the time of colonial settlement until as late as 1917, Melanesian uprisings were common and were constantly feared by the settlers and authorities alike. The insurrection of 1856–59 near Nouméa and the uprising of 1878–79, which extended along the west coast from Bouloupari to Poya, seriously endangered French occupation. Grievances centred around the confiscation of Melanesian lands, the foraging of settlers' cattle on the Melanesians' produce gardens, and, in 1917, the head tax, which had been imposed by the colonial government in 1899. The French suppressed each uprising (with the help of indigenous auxiliaries) by destroying villages and crops and demanding unconditional surrender, and insurgents were punished by deportation or execution and further confiscation of their lands. By 1860 French authority had been established over the southern third of the mainland, and in the next decade policies for the disposal of indigenous land, the regrouping of tribes, and the appointment of a system of tribal chiefs to represent the administration were established. By the end of the 19th century large areas of Melanesian land had been alienated and the inhabitants relegated to reserves. Forced labour, limitations on travel, and curfews were imposed from the early days of colonial rule and became the basis of a system of administrative law codified in 1887 as the *indigénat* ("native regulations"). In 1899 a head tax was levied on male Melanesians to oblige them to obtain employment with settlers and the government and, like the *indigénat,* remained in force until 1946.

Although the prime concern of the early colonial administration was the reception and control of about 22,000 French convicts sent to New Caledonia between 1864 and 1897, the need to attract free settlers and to provide a supply of cheap labour also were continuing preoccupations. Between 1864 and 1939 some 60,000 indentured labourers, as well as the convicts, were imported to construct public works and to work on plantations, ships, wharves, and mines and in commerce and domestic service. These included ni-Vanuatu (the indigenous population of Vanuatu) and Solomon Islanders (1865–1920), Vietnamese (1891–1939), Javanese (1896–1939), and Japanese (1892–1921). Only a small percentage of the survivors of these workers remained in the colony after the expiry of their contracts, and, although few ex-convicts left the territory, not many established families. New Caledonia was also the unlikely home of 4,000 *déportés,* political exiles of the 1871 uprising of the Paris Commune, but few of the survivors stayed after being granted amnesty. Most of the free white settlers were either former members of the French administration and armed forces, settlers from Australia and New Zealand, or former sugar planters from Réunion. Two later waves of free settlers in the 1890s and 1920s were encouraged to migrate to help establish coffee and cotton industries, respectively. By 1936, however, there were only 15,000 residents of European origin among 29,000 native Melanesians. After World War II the European population gradually increased by migration, but the main influx of white and Polynesian settlers occurred in the years leading up to and including the nickel boom of 1969–72. For the first time the Melanesians became a minority in their own country, although they still were the largest single ethnic group.

The main influx of settlers

Since the beginning of French settlement, political debate among the white settlers revolved around the rights of residents to administer their own affairs without endangering the flow of French financial aid. Between 1885 and 1956 a locally elected general council advised the governor on territorial affairs and voted a local budget. By 1953 French citizenship had been granted to all, regardless of ethnic origin. Melanesians then formed a coalition with Europeans to bring the party, the Caledonian Union (Union Calédonienne), to power on a ticket of full self-government in local affairs. Progress toward self-government was made in 1957 when a Territorial Assembly was created with powers to elect an executive to administer, under the presidency of a high commissioner, the affairs of the territory. In the 1970s the Caledonian Union became an increasingly Melanesian party as Europeans deserted it. Many Melanesians also left to form parties that campaigned in support of independence. In 1979 the Caledonian Union, which had converted to the cause of independence, formed an Independence Front in coalition with several minority Melanesian parties. The Socialist French government of 1981 took many steps to stem the growing political polarization and granted complete self-government in territorial affairs under the Lemoine Statute of 1984. The statute was rejected, however, by the Independence Front, which reconstituted itself as the Kanak Socialist National Liberation Front (Front de Libération Nationale Kanake et Socialiste; FLNKS). The FLNKS boycotted the elections held in November 1984 and took control of most of the territory outside Nouméa. Peace was gradually restored as the government proposed to grant the Melanesians independence in association with France in an arrangement through which French citizens would retain residential and other rights. The European community, with the strong support of non-Melanesian ethnic groups, vigorously rejected this proposal.

In 1985 the French government transferred most of the powers of self-government to four regional councils. In elections held in September 1985 the FLNKS won a majority of seats in all except the Nouméa region, which was won by the settler-dominated party, the Rally for Caledonia in the Republic (Rassemblement pour la Calédonie dans la Republique; RPCR). In 1986 France reduced the powers of the regional councils and held a referendum. The referendum was boycotted by the FLNKS and produced a large majority in favour of continued ties with France. It was followed by a new statute for self-government, which altered the regional boundaries in order to limit FLNKS successes to the principal concentrations of Melanesians in the Loyalty Islands and on the east coast of the mainland.

For its part the FLNKS demonstrated its ability to direct the votes of the great mass of Melanesian voters in four electoral consultations since 1984. With the aid of the independent countries of the South Pacific that form the South Pacific Forum, the question of New Caledonia was listed on the agenda of the United Nations Committee of Decolonization. The lack of consensus on the subject in France was paralleled by the increasing polarization within the territory. (D.L.Sh.)

For later developments in the history of New Caledonia, see the BRITANNICA BOOK OF THE YEAR.

Papua New Guinea

The Independent State of Papua New Guinea encompasses the eastern half of the island of New Guinea (the western half, Irian Jaya, belonging to Indonesia) and its offshore islands as well as the islands of the Bismarck Archipelago (New Britain, New Ireland, and the Admiralty Islands) and those of Bougainville and Buka. Together these islands cover 178,704 square miles (462,840 square kilometres) of the southwestern Pacific, stretching from just south of the Equator to the Torres Strait, which separates New Guinea from Cape York Peninsula, the northernmost extension of Australia.

The official languages of the country are all introduced: English, Tok Pisin (Melanesian Pidgin), and Hiri, or Police, Motu, the last being a simplified form of the language of the people who lived around what is now the capital, Port Moresby, when it was first established in 1884. The islands that constitute Papua New Guinea have been settled for tens of thousands of years by the mixture of peoples who are generally referred to as Melanesians, and one of the principal challenges facing those who govern the modern state is the difficulty of welding together hundreds of diverse, once-isolated regional societies into a viable modern nation-state.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Papua New Guinea's magnificent and varied scenery reflects a generally recent geologic history in which movements of the Earth's crust resulted in the collision of the northward-moving Australian Plate with the westward-moving Pacific Plate. The low-lying plains of southern New Guinea are geologically part of the Australian Plate. Indeed, New Guinea was separated physically from Australia only 8,000 years ago by the shallow flooding of the Torres Strait. The southern New Guinea plains, called the Fly-Digul shelf after the Fly and Digul rivers, are geologically stable but very sparsely populated by seminomadic sago gatherers.

Northward lies a belt of limestone country of varying width, most prominent in the Kikori River–Lake Kutubu area. This forms an extraordinarily harsh environment of jumbled karst, dolines, rock towers, and seemingly endless ridges of jagged rock, all covered in virtually impenetrable lowland rain forest. The discovery of mineral deposits about 1970 stimulated mining activity in the previously deserted region.

Farther to the north lie the Highlands, an east–west-trending zone of mountains with elevations in excess of 13,000 feet (4,000 metres) and enclosed upland basins whose floors are usually at 4,500 feet or higher. The basins contain lake deposits, formed in the recent geologic past by impeded drainage; soil wash from the surrounding mountains; and layers of volcanic ash, or tephra, deposited from nearby and recently active volcanoes. Such basins, therefore, are usually very fertile. Temperatures are much cooler than in the lowlands, and frosts occasionally cause serious damage to the sweet potato (*kaukau*), which is the staple diet of the area. Much of the natural vegetation of most of the upland basins has been removed by the inten-


Scale 1:11,236,000
1 inch equals approx. 177 miles
0 50 100 150 200 mi
0 100 200 300 km
Conic Projection
141°
144°
147°
150°
153°
156°
3°
6°
9°
12°
PACIFIC OCEAN
CORAL SEA
SOLOMON SEA
BISMARCK ARCHIPELAGO
BISMARCK SEA
NEW IRELAND
NEW HANOVER
NEW BRITAIN
MANUS
ADMIRALTY ISLANDS
NIMOA GROUP
HERMIT ISLANDS
MANUS ISLAND
LORENGAU
RAMBUTYO ISLAND
MUSSAU ISLAND
ST MATTHIAS GROUP
Yssabel Channel
KAVIENG
TABAR ISLANDS
LIHIR ISLAND
TANGA ISLANDS
NAMATANAI
NUGURIA ISLANDS
Mount Kavogaeng
FENI ISLANDS
GREEN ISLANDS
TULUN ISLANDS
TAKUU ISLANDS
NORTH SOLOMONS
BUKA ISLAND
SOLOMON ISLANDS
BOUGAINVILLE
Arawa
Panguna
Mount Takuan
Buka
CHOISEUL
NEW GEORGIA
SOLOMON ISLANDS
KIRIWINA ISLAND
Losuia
MUYUA ISLAND
RUDIBUDI ISLANDS
MILNE BAY
D'ENTRECASTEAUX IS
GOODENOUGH ISLAND
Misima Island
Bwagaoia
LOUISIADE ARCHIPELAGO
NORMANBY ISLAND
Milne Bay
Alotau
Suregari
TAGULA ISLAND
FYLA ISLAND
Rabaul
Kokopo
GAZELLE PENINSULA
Mount Kowave
2334
Kimbe Bay
WILLAUMEZ PENINSULA
Mount Talawe
1524
Cape Gloucester
Cape Orford
Kandrian
Finschhafen
Lae
Huon Gulf
Markham Valley
Ramu River
Madang
Mount Wilhelm
4509
Mount Hagen
Wabag
Mount Kerigomna
Mount Bosavi
2507
Mount Sisa
Tabubil
3360
Mount Capella
3232
Mount Victoria
4214
Telefomin
Vanimo
Aitape
Wewak
Aitape
Aitape
Angoram
Ambunti
Maprik
Boiken
Boiken
Boiken

MAP INDEX

Political subdivisions

Central ..... 10 00 s 148 00 e
Chimbu ..... 6 15 s 144 50 e
East New Britain . 6 30 s 152 30 e
East Sepik ..... 4 00 s 143 45 e
Eastern Highlands ..... 6 30 s 145 35 e
Enga ..... 5 30 s 143 30 e
Gulf ..... 8 00 s 145 00 e
Madang ..... 5 00 s 145 30 e
Manus ..... 2 00 s 147 00 e
Milne Bay ..... 10 00 s 152 30 e
Morobe ..... 7 00 s 147 00 e
National Capital .. 9 25 s 147 10 e
New Ireland ..... 3 00 s 151 30 e
North Solomons . 6 00 s 155 00 e
Northern ..... 9 00 s 148 30 e
Sandaun (West Sepik) ..... 3 30 s 142 00 e
Southern Highlands ..... 6 15 s 143 25 e
West New Britain ..... 5 00 s 149 30 e
West Sepik, see Sandaun
Western ..... 7 30 s 142 00 e
Western Highlands ..... 5 45 s 144 30 e

Cities and towns

Aitape ..... 3 08 s 142 21 e
Aiyura ..... 6 20 s 145 54 e
Alotau ..... 10 20 s 150 25 e
Ambunti ..... 4 14 s 142 50 e
Angoram ..... 4 04 s 144 04 e
Arawa ..... 6 13 s 155 33 e
Baimuru ..... 7 30 s 144 49 e
Balimo ..... 8 03 s 142 57 e
Banz ..... 5 47 s 144 37 e
Bereina ..... 8 38 s 146 30 e
Bogia ..... 4 16 s 144 58 e
Buin ..... 6 50 s 155 44 e
Bulolo ..... 7 12 s 146 39 e
Bwagaoia ..... 10 42 s 150 50 e
Chuave ..... 6 08 s 145 08 e
Daru ..... 9 05 s 143 12 e
Finschhafen ..... 6 36 s 147 51 e
Goroka ..... 6 05 s 145 23 e
Ihu ..... 7 55 s 145 24 e
Kaiapit ..... 6 16 s 146 16 e
Kainantu ..... 6 17 s 145 52 e
Kandrian ..... 6 13 s 149 33 e
Kavieng ..... 2 34 s 150 48 e
Kerema ..... 7 58 s 145 46 e
Kerowagi ..... 5 51 s 144 51 e
Kieta ..... 6 13 s 155 38 e
Kikori ..... 7 25 s 144 15 e
Kimbe ..... 5 33 s 150 09 e
Kiunga ..... 6 07 s 141 18 e
Kokopo ..... 4 21 s 152 16 e
Kundiawa ..... 6 01 s 144 58 e
Kupiano ..... 10 05 s 148 11 e
Kwikila ..... 9 48 s 147 41 e
Lae ..... 6 44 s 147 00 e
Laiagam ..... 5 17 s 143 46 e
Lorengau ..... 2 01 s 147 16 e
Losuia ..... 8 32 s 151 04 e
Madang ..... 5 13 s 145 48 e
Malalaua ..... 8 04 s 146 10 e
Maprik ..... 3 38 s 143 03 e
Mendi ..... 6 09 s 143 39 e
Minj ..... 5 54 s 144 41 e
Mount Hagen ... 5 52 s 144 13 e
Mumeng ..... 6 59 s 146 35 e
Namatanai ..... 3 40 s 152 27 e
Panguna ..... 6 15 s 155 30 e
Popondetta ..... 8 46 s 148 14 e
Porgera ..... 5 26 s 143 12 e
Port Moresby ... 9 29 s 147 11 e
Rabaul ..... 4 12 s 152 11 e
Saidor ..... 5 38 s 146 28 e
Samarai ..... 10 37 s 150 40 e
Sogeri ..... 9 25 s 147 25 e
Tabubil ..... 5 15 s 141 15 e
Tapini ..... 8 22 s 146 59 e
Tari ..... 5 42 s 142 57 e
Telefomin ..... 5 08 s 141 35 e
Vanimo ..... 2 41 s 141 18 e
Wabag ..... 5 29 s 143 42 e
Wau ..... 7 20 s 146 43 e
Wewak ..... 3 33 s 143 38 e

Physical features and points of interest

Adelbert Range .. 4 43 s 145 15 e
Admiralty Islands ..... 2 10 s 147 00 e
Alice, see Tedi
Arop (Long) Island ..... 5 20 s 147 06 e
Balbi, Mount, *volcano* ..... 5 55 s 154 59 e
Bangeta, Mount . 6 16 s 147 04 e
Bismarck Archipelago .... 5 00 s 150 00 e
Bismarck Sea ... 4 00 s 148 00 e
Bosavi, Mount ... 6 35 s 142 50 e
Bougainville, *island* ..... 6 00 s 155 00 e
Bougainville Strait ..... 6 40 s 156 10 e
Budibudi (Laughlan or Nada) Islands .. 9 15 s 153 40 e
Buka Island ..... 5 15 s 154 38 e
Capela, Mount .. 4 58 s 141 06 e
Carteret, see Tulun Islands
Central Range ... 5 00 s 143 00 e
Coral Sea ..... 12 00 s 150 00 e
Crown Island .... 5 07 s 146 58 e
Crown Island Wildlife Management Area ..... 5 07 s 146 58 e
D'Entrecasteaux Islands ..... 9 39 s 150 42 e
Father, The, see Ulawun, Mount
Feni Islands .... 4 05 s 153 42 e
Fergusson Island . 9 30 s 150 40 e
Finisterre Range . 5 50 s 146 07 e
Fly, *river* ..... 8 30 s 143 41 e
Gazelle Peninsula ..... 4 40 s 152 00 e
Goodenough (Morata) Island . 9 22 s 150 16 e
Great Admiralty, see Manus Island
Green (Nissan) Islands ..... 4 30 s 154 10 e
Hermit Islands ... 1 32 s 145 05 e
Huon Gulf ..... 7 10 s 147 25 e
Huon Peninsula .. 6 25 s 147 30 e
Kaiserin Augusta, see Sepik
Karkar Island .... 4 38 s 145 58 e
Kikori, *river* ..... 7 39 s 144 17 e
Kimbe Bay ..... 5 10 s 150 30 e
Kiriwina (Trobriand) Island ..... 8 40 s 151 06 e
Konogaiang, Mount ..... 4 05 s 152 42 e
Kutubu, Lake ... 6 24 s 143 20 e
Laughlan, see Budibudi Islands
Lihir Island ..... 3 05 s 152 35 e
Long, see Arop Island
Long Island Wildlife Management Area ..... 5 20 s 147 05 e
Louisiade Archipelago .... 11 12 s 153 00 e
McAdam National Park ... 7 13 s 146 38 e
Manam Island ... 4 05 s 145 02 e
Manus (Great Admiralty) Island ..... 2 05 s 147 00 e
Markham, *river* .. 6 44 s 146 58 e
Milne Bay ..... 10 22 s 150 30 e
Misima (St. Aignan) Island .. 10 40 s 152 45 e
Morata, see Goodenough Island
Mortlock, see Takuu Islands
Murray, Lake .... 7 00 s 141 30 e
Mussau Island ... 1 25 s 149 38 e
Muyua (Woodlark) Island ..... 9 06 s 152 50 e
Nada, see Budibudi Islands
New Britain, *island* ..... 5 40 s 151 00 e
New Hanover, *island* ..... 2 30 s 150 15 e
New Ireland, *island* ..... 3 20 s 152 00 e
Ninigo Group, *atolls* ..... 1 22 s 144 17 e
Nissan, see Green Islands
Normanby Island . 10 00 s 151 00 e
Nuguria Islands .. 3 20 s 154 45 e
Ok Tedi Mine ... 5 10 s 141 15 e
Ottilien, see Ramu
Owen Stanley Range ..... 9 20 s 148 00 e
Pacific Ocean ... 3 00 s 155 00 e
Papua, Gulf of ... 8 30 s 145 00 e
Purari, *river* ..... 7 46 s 145 10 e
Rambutyo Island . 2 18 s 147 49 e
Ramu (Ottilien), *river* ..... 4 02 s 144 41 e
Rooke, see Umboi Island
Rossel, see Yela Island
St. Aignan, see Misima Island
St. George's Channel ..... 4 30 s 152 30 e
St. Matthias Group ..... 1 30 s 149 48 e
Schouten Islands . 3 30 s 144 30 e
Sepik (Kaiserin Augusta), *river* . 3 51 s 144 32 e
Solomon Islands . 6 00 s 157 00 e
Solomon Sea .... 8 00 s 152 00 e
Star Mountains .. 5 00 s 141 00 e
Strickland, *river* .. 7 35 s 141 23 e
Suckling, Mount . 9 45 s 148 58 e
Sud-Est, see Tagula Island
Tabar Islands ... 2 50 s 152 00 e
Tagula (Sud-Est) Island ..... 11 30 s 153 30 e
Takuan, Mount .. 6 27 s 155 36 e
Takuu (Mortlock) Islands ..... 4 45 s 157 00 e
Talawe, Mount .. 5 32 s 148 24 e
Tanga Islands .... 3 25 s 153 15 e
Tedi (Alice), *river* . 6 10 s 141 08 e
Torres Strait .... 10 25 s 142 10 e
Torricelli Mountains ..... 3 23 s 142 15 e
Trobriand, see Kiriwina Island
Tulun (Carteret) Islands ..... 4 45 s 155 20 e
Ulawun (The Father), Mount, *volcano* ..... 5 03 s 151 20 e
Umboi (Rooke) Island ..... 5 36 s 148 00 e
Varirata National Park ..... 9 27 s 147 20 e
Vitiaz Strait ..... 5 35 s 147 00 e
Vitu, see Witu Islands
Wamtakin, Mount ..... 5 11 s 141 48 e
Wawoi, *river* .... 7 48 s 143 15 e
Wilhelm, Mount .. 5 48 s 145 02 e
Willaumez Peninsula ..... 5 15 s 150 02 e
Witu (Vitu) Islands ..... 4 40 s 149 18 e
Woodlark, see Muyua Island
Yela (Rossel) Island ..... 11 21 s 154 09 e
Ysabel Channel .. 2 00 s 150 00 e

sive agricultural technology of the Highlanders. Throughout the Highlands, carefully tended gridiron gardens with their drainage ditches or perfectly circular earth-covered mounds of compost dominate the landscape.

The mountains drop away sharply to the north, and the intensive cultivation of fertile soils gives way to swidden (slash-and-burn, or shifting) cultivation of taro and yams in the forests of the foothills. These thinly populated areas in turn give way to the sago swamps along the courses of the great Ramu and Sepik rivers, the latter area famous for its magnificent folk art but equally noted for its immense numbers of mosquitoes. In the slightly more elevated areas away from the main rivers, the high water table combines with the human activities of swidden cultivation and hunting (the burning of vegetation to drive animals toward hunters) to create extensive areas of poor grassland.

The most northerly zone consists of a complex, unstable volcanic arc stretching from the Schouten Islands off Wewak to the Huon Peninsula and through New Britain island, at which point it bifurcates, one arm sweeping northwestward through New Ireland and the Admiralty Islands, the other proceeding southeastward through Bougainville and the Solomon Islands. The north coast, unlike the swampy south coast, drops sharply to the sea, and stands of mangrove and, in brackish waters, nipa palm are rather limited. This northern volcanic fringe contains some of the most fertile soils of the islands. Thus, despite the hazards of volcanic activity and frequent earth tremors, the area is generally well-populated. The island of Karkar and the Gazelle Peninsula of New Britain island are centres of particularly dense population where yams, taro, bananas, and fish are the basic foods. Elsewhere, previously little-used volcanic soils are the focus of large-scale resettlement and oil-palm-cropping schemes, especially in western New Britain.

Drainage. Steeply sloping mountain areas, exceptionally heavy rainfall, geologic instability in all except the most southerly areas, and rapid growth of both population and commercial enterprise combine to create some of the highest soil-erosion rates in the world, rivaling those of



Sweet-potato farming, Southern Highlands province, Papua New Guinea.
Bob and Ira Spring

the Himalaya region. Consequently, while rivers are usually quite short in length, they carry extraordinarily high sediment loads, which have built up vast swampy plains and deltas, especially along the Sepik, the Fly, and the Purari river systems. Once they leave the Highlands, often through spectacular gorges, such rivers meander slowly across the sediment plains. For example, 512 miles (824 kilometres) from its mouth the Fly River is a mere 60 feet above sea level. Such high deposition rates create major problems for any proposed human use of these rivers, for transportation or hydroelectricity generation, for example.

Climate. Although all the climatic regions of Papua New Guinea are basically tropical, they are nevertheless varied. In the lowlands, mean annual maximum temperatures range between 86° and 90° F (30° and 32° C), and the minimums are between 73° and 75° F (23° and 24° C). Seasonal variation in temperature is slight, and the daily variation approximates the annual variation. Cooler conditions prevail in the Highlands, where night frosts are common above 6,500 feet; daytime temperatures there generally exceed 72° F (22° C) regardless of season.

Rainfall patterns

Rainfall, rather than temperature, is the determinant of season. Precipitation is dependent on two wind systems, the southeast trades and the northwesterly turbulence zone (the monsoon), and on the three site characteristics of latitude, altitude, and exposure. The southeasterlies blow for approximately seven months (May to November) on the extreme southeast of the country (Milne Bay) and for gradually shorter periods northward, predominating for only three months in the Admiralty Islands. Conversely, northwesterlies are more common on the north coast and in the Bismarck Archipelago, but they affect Port Moresby for only three to four months of the year (December through March). The Highlands seem to have their own airflow systems, receiving rain throughout the year totaling between 100 and 160 inches (2,500 and 4,000 millimetres). With the northwesterlies, rain is frequently from convectional storms, and rain shadow effects are reduced. With the southeasterlies, however, exposure is particularly important. The Port Moresby coastal area is parched throughout the period of the southeasterlies, which flow parallel to the coast, yet where mountainous land lies athwart the airflow, as in New Britain or the southward-facing slopes of the Highlands, rainfall is extremely heavy, frequently exceeding 300 inches. Port Moresby receives less than 50 inches of rain annually, which occasions problems of water and hydroelectric power supply.

Plant and animal life. In most areas the coastline is lined by mangrove swamp, succeeded inland by nipa palm (*Nipa fruticans*) in brackish waters. Farther inland, particularly along the valleys of the larger rivers in the north and along the deltas of the south coast, large stands of sago palm are scattered. Primary lowland rain forest covers much of the island up to approximately 3,300 feet above sea level. The forest is characterized by a large number of species, by the absence of pure stands of any one species, by fairly distinct layering of the forest into two or three levels, by the limited development of undergrowth, and by the small amount of human impact upon it. Dense undergrowth is usually a sign of human interference, except on soils that are particularly poor or where the low height of the forest allows sunlight to penetrate to the soil surface. In those lowland areas where drier conditions prevail, agriculture and hunting have often changed the delicately balanced natural ecosystem into grassland (for example, in the Sepik and Markham river valleys).

Lowland rain forest

In the uplands above 3,300 feet, stands of single species become more common, and trees such as oak, beech, red cedar, and pine become increasingly dominant. Above 6,500 feet, cloud or moss forest characterized by conifers, tree ferns, and a wealth of fungi and epiphytes (such as mosses and orchids) appears.

Papua New Guinea possesses a rich variety of reptiles, marsupials (animals that carry their young in pouches), native freshwater fish, and birds but is almost devoid of large mammals. Since the Tertiary Period (which occurred from 66,400,000 to 1,600,000 years ago), the island has been isolated from Southeast Asia by the sea, and its animals thereafter evolved in isolation. This appears not to have been true for plant life, for New Guinea has been a centre of dispersal for many plants to all neighbouring regions. The extraordinary profusion in Papua New Guinea of such plants as orchids, figs (*Ficus*), and false beech (*Nothofagus*) and of such animals as cassowaries (large, flightless birds), birds of paradise, parrots, butterflies, and marsupials—including the tree kangaroo and phalanger—gives the island an unparalleled biogeography.

Papua New Guinea's unique biological species have long been sought by collectors all over the world. The government has introduced several conservation and protection measures; the export of birds of paradise is now banned, and hunters thereof are restricted to the use of traditional weapons. Similarly, the export of many other birds and butterflies and of crocodile skins is strictly regulated. Other policies encourage the controlled expansion of selected exports of "farmed" orchids and crocodiles and of "cultivated" butterflies and other insects.

Wildlife conservation

Settlement patterns. More than 80 percent of the country's population lives in rural areas. Rural settlement patterns are extremely varied. In isolated areas of the southern interior there still remains a handful of the previously common giant communal structures that house the whole male population, with a circling cluster of women's huts. In many coastal areas, villages stretch between the beach and an inland swamp in long lines, broken into clan or family segments. In the Highlands numerous village forms exist: in the Eastern Highlands and Chimbu provinces, villages previously clustered along ridgetops for defensive purposes are dispersing increasingly downslope; in the Western Highlands and Enga provinces, the traditional form is of scattered households each surrounded by its own land, with separate houses for men and women; in the Telefomin area, clustered villages are supplemented by scattered garden houses at a distance from the central settlement. Housing styles originally reflected environmental circumstances, ranging from houses built over the sea in sheltered coastal areas, which avoided mosquitoes and simplified sanitary disposal problems, to houses dug into the ground to retain heat in the colder parts of the Highlands. Such adjustments are disappearing as Western-style housing becomes more common.

The 14 percent of the population that is urban lives in towns whose original location was determined either by access to a good harbour for early colonial planters or, in the interior, by the sufficient availability of level land for an airstrip. Despite the greatly diminished importance of plantations and the removal out of town of most of the

airstrips, these origins help determine the existing urban layout. Within the towns, of which Port Moresby and Lae are the largest, there are great contrasts in housing. Port Moresby has, for example, grand modern apartment blocks overlooking the sea, but nearly half its population lives in squatter settlements. Rapid urban growth, at about 6 percent annually, and considerable income inequalities mean that public or low-cost housing cannot be built in sufficient quantities to accommodate much more than a relatively small portion of the urban population.

More than 700 ethnic groups

The people. *Ethnic composition.* Papua New Guinea's ethnic composition is extremely complex. There are more than 700 ethnic groups; these are often separated into two major divisions, Papuan (constituting more than four-fifths of the total population) and Melanesian (constituting all but about 1 percent of the rest of the population). Very small minorities of chiefly Micronesian and chiefly Polynesian ethnic groups can be found on some of the outlying islands. Within the larger divisions, characteristics vary widely; the Melanesians, for example, who generally inhabit the coastal regions and offshore islands, range from the relatively tall, light-skinned Trobriand Islanders to the black-skinned people of Buka. Ethnic Papuans, who live mainly in the interior, are often physically characterized by other citizens as stocky and muscular.

While at independence in 1975 the expatriate community of 40,000 was predominantly of Australian and Chinese origin, a decade later the slightly smaller foreign community was more mixed, with the largest non-Western group being from the Philippines.

Linguistic composition. There are two radically different indigenous language types—Austronesian, or Melanesian, and non-Austronesian, or Papuan—and the language areas generally reflect ethnic divisions. Some 200 related Austronesian languages occur, mainly in the islands and along the New Guinea coast. The approximately 550 non-Austronesian languages have small speech communities, the largest being the Enga, in the Wabag area. Because of the multiplicity of tongues, Tok Pisin has developed as an effective lingua franca.

Religious composition. About three-fifths of the populace consider themselves Protestant, and the largest portion of these are Lutheran. Nearly a third are Roman Catholic. The remainder include Anglicans and Bahā'is. Despite the numbers enrolled in introduced religions, traditional religious beliefs persist, and rituals of magic, spells, and sorcery are still widely practiced.

Demographic trends. Population growth is high, about 2.7 percent annually, and two-fifths of the population is under 15 years of age. Since employment in the commercial sector has grown much more slowly (it declined in the 1980s), the government has attempted to concentrate its policies on rural, village-based development in an attempt to reduce urban migration and demands for formal employment. In the early 1980s the birth rate, though falling, was 50 percent higher than the world average, while the death rate was only 25 percent higher and falling much faster. Consequently, full employment is likely to remain a problem.

The economy. *Resources.* From 1970 onward, major mineral discoveries have transformed the Papua New Guinean economy from one dependent on tropical crops (coffee, copra, cacao) to one that depends for much of its exports on minerals. Large gold or gold and copper deposits have led to major development at Panguna on Bougainville, Ok Tedi in the Star Mountains, Misima Island in Milne Bay, Porgera in the Western Highlands, and Lihir Island near New Ireland. After 70 years of exploration, major natural gas and oil finds were made in the late 1980s; at the end of the decade plans for their exploitation were being prepared.

Dispute over mineral revenues

Such resource exploitation has caused local groups to contest, principally with the national government, the distribution of mineral revenues. Open warfare broke out on Bougainville, and the island's gold and copper mine was closed in 1989. Despite this, revenues from mining provided one-fifth of government revenue during the 1980s.

Forestry exploitation, extensive particularly around Madang, in the northeast, and on New Britain and New Ireland islands, accounted for some 10 percent of exports. While fisheries have great potential, this industry has been especially erratic.

Agriculture. Most agricultural production is for subsistence, and production remains difficult to estimate. Almost all commercial crop production is exported. High-quality Arabica coffee is grown throughout the Highlands, while cacao is grown in the islands, oil palms in western New Britain and the southeastern mainland, and robusta coffee on the north coast; copra is produced in all lowland areas, together with some rubber. Production of all crops has been stagnant since 1985. However, after 1975, smallholder growers increasingly took over the bulk of production, replacing plantations.

Industry. Industrial output is of little significance, accounting for only one-tenth of gross domestic product despite the government's attempts to promote its expansion. Processing and manufacturing are centred chiefly in Lae and Port Moresby; products include food, beverages, tobacco goods, wood products, textiles, and metal goods.

Finance. The kina, Papua New Guinea's currency, remained stable and very strong following independence, although problems on Bougainville in 1989 resulted in its devaluation. Responsible financial management remains a strong feature of the Papua New Guinean economy. Foreign investment and taxes thereon have dominated the economy and government receipts since independence.

Trade. Papua New Guinea's trade is roughly in balance. Its exports, principally gold and copper, go primarily to Japan and Europe. Australia has consistently supplied just under half of the country's imports since independence. After machinery and transport equipment, foodstuffs are a major import; the basic diet of most urban dwellers consists of rice and canned fish or meat, very little of which is produced within the country.

Transportation. Despite the construction of the Okuk Highway, which links Lae to all major Highland towns, air transport remains the most important and, from Port Moresby, essentially the only form of interurban transportation. Papua New Guinea has more regularly operative airstrips per 1,000 population than almost any other country in the world. International air access is via Jackson's Airport, Port Moresby.

Administration and social conditions. Papua New Guinea is a member of the Commonwealth, with the British monarch as its head of state. The government follows a unicameral parliamentary system. The head of state appoints the prime minister, who presides over the National Executive Council; the governor-general, who is nominated by the council and represents the crown; and the chief justice, who presides over the six- to eight-member Supreme Court. Political parties are extraordinarily divisive and numerous and have weak local organization. A feature of parliamentary politics in Papua New Guinea during the 1980s was the frequency of motions of no confidence in the government (allowed every six months under the 1975 constitution). Despite the frequent changes of government, policies have remained notably stable. Papua New Guinea's major foreign policy concerns are its bridging role between Southeast Asia and the rest of the Pacific and its relations with Indonesia, which have been occasionally strained as a result of events in neighbouring Irian Jaya. A feature of Papua New Guinea's political life has been its active and free press staffed almost exclusively by national journalists.

Education. Despite Papua New Guinea's impressive progress in education, three-fifths of all adults are illiterate. Only about two-thirds of all children attend primary school, only half of those attending finish primary school, and only one in six primary-school students receives any secondary schooling. The rapid growth and extreme youthfulness of the population (42 percent are under 15 years of age) mean that educational demand outstrips supply even though education is a major item of government expenditure. Families opt to educate sons rather than daughters: for every 10 boys in primary schools, there are 8 girls; at secondary school the figure is 6 girls, and at university 2 girls.

Health and welfare. The village and the family in Papua

New Guinea provide largely those welfare services which, in more developed countries, are assumed by the state. Health expenditure in Papua New Guinea as a proportion of total government spending is relatively high for a developing country. Pneumonia and malaria are among the leading causes of mortality. Emphasis on primary health care at the village level has reduced maternal and infant mortality. It has also led to a rapid increase in life expectancy.

Cultural life. Despite incorporation into the modern world, Papua New Guinea retains a rich variety of village cultures, expressed in its human-molded landscapes and its sculpture, painting, storytelling, dance, and body decoration. The Papua New Guinea National Museum and Art Gallery at Waigani has a significant collection of ethnographic artifacts, and the government encourages the continuation and, in some cases, the revival of activities associated with traditional cultures.

For statistical data on the land and people of Papua New Guinea, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The peopling of New Guinea. Relatively little archaeological work has been carried out in New Guinea. On the basis of current evidence, it has been postulated that parts of New Guinea were occupied as early as 50,000 years ago. Remains of swamp-drainage channels and other water-management works indicate the existence of intensive agriculture on the island around 7000 BC. The intensity and length of time of human occupation of the Highlands are evidenced by the extent of man-made landscapes in the region. These discoveries are made even more interesting by the fact that the sweet potato, the present staple crop of the region, does not seem to have arrived in the area from the Americas until 300 or 400 years ago. It is presumed that taro was the earlier staple, as it still is in some isolated Highlands basins such as that at Telefomin.

Early agriculture

The colonial period. Malay and possibly Chinese traders took spoils and slaves from New Guinea for hundreds of years. The first European visitor may have been Jorge de Meneses, who possibly landed on the island in 1526–27 while en route to the Moluccas. The first European attempt at colonization was made in 1793 by Lieutenant John Hayes, a British naval officer, near Manokwari, now in Irian Jaya. The Dutch, however, claimed the western half of the island as part of the Dutch East Indies in 1828; their control remained nominal until 1898, when their first permanent administrative posts were set up at Fakfak and Manokwari. Captain John Moresby of Great Britain surveyed the southeastern coast in the 1870s, and European planters had moved onto New Britain and New Ireland by the 1880s. By 1884 the southeastern quadrant of New Guinea had been established as a British protectorate, and in the same year the German New Guinea Company began its administration of the northeastern quadrant. Despite early gold finds in British New Guinea (after 1906 administered by Australia as Papua), it was in German New Guinea, administered by the German imperial government after 1899, that most early economic activity took place. Plantations were widely established in the islands and around Madang; labourers were brought from the Sepik River, the Markham Valley, and Buka Island.

German New Guinea was taken over by Australia as a mandated territory of the League of Nations in 1921, after World War I. It remained administratively separate from Papua, where the protectionist policies of Sir Hubert Murray (lieutenant governor of Papua, 1908–40) did little to encourage colonial investment. The discovery in the 1920s of massive gold deposits at the Bulolo River and Edie Creek in the mandated territory increased the disparity of colonial impact in the different regions. In the early 1930s an even greater discovery was made—nearly 1,000,000 people previously unknown to Europeans were contacted in the Highlands basins of the Australian mandate. At first, the Highlanders were utilized as a massive new source of labour for the coastal plantations, a role they continue to play. At the end of World War II, however, the growing of Arabica coffee by small landholders spread rapidly throughout much of the Highlands. Cacao also was rapidly adopted as a plantation and smallholder crop in the islands and around Madang.

Discovery of gold

In 1945 Australia combined its administration of Papua and that of the mandate into the Territory of Papua and New Guinea, with the common capital at Port Moresby, on the south coast of Papua. From 1946 Australia administered the mandate of New Guinea as a United Nations trust territory. Despite the general lack of economic development in Papua, its one large town of Port Moresby grew rapidly and attracted large numbers of migrants, particularly from the poorer areas.

Independence. Self-government was achieved on Dec. 1, 1973, and full independence from Australia on Sept. 16, 1975. At independence there were attempts in both Bougainville and Papua to secede from the new state. The crisis on Bougainville was subdued as a result of the introduction of provincial government and a devolution of some national powers to such governments. The Papuan movement was also pacified by this action, but it, too, lacked cohesive support. In late 1988 secessionism was rekindled on Bougainville as a result of disputes over the distribution of mineral revenues, and in 1990 Bougainville declared independence, though its independence was not recognized by any country. (R.T.J.)

For later developments in the history of Papua New Guinea, see the BRITANNICA BOOK OF THE YEAR.

Solomon Islands

Solomon Islands, an independent country in the southwestern Pacific Ocean, consists of a double chain of volcanic islands and coral atolls in Melanesia, occupying a total land area of 10,954 square miles (28,370 square kilometres). Buka and Bougainville islands, at the northwestern end of the Solomons chain, form the North Solomons province of Papua New Guinea. Honiara, on Guadalcanal Island, is Solomon Islands' capital and largest city.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The main islands of the group are large and rugged, rising to 8,028 feet (2,447 metres) at Mount Makarakomburu (Makarakombou) on Guadalcanal. They lie in two parallel chains: the western includes Vella Lavella, Kolombangara, New Georgia, and Guadalcanal; the eastern, Choiseul Island, Santa Isabel, and Malaita. The chains converge on Makira (San Cristóbal) Island. Geologically, they are part of the volcanic arc extending from New Ireland, in Papua New Guinea, to Vanuatu.

Principal islands

The climate is tropical oceanic; that is, hot and humid, but relieved by cool winds and abundant, year-round rainfall. Temperatures seldom exceed 90° F (32° C) and rainfall generally averages 120–140 inches (3,000–3,500 millimetres) a year. Heavily wooded, mountainous terrain is characteristic and, although there are extensive plains, only those on the northern side of Guadalcanal have been developed for large-scale agriculture. As in most island groups, animal life is limited. There are hot springs on Savo, where a volcano last erupted in the 1840s. Solomon Islands has a number of other volcanoes. Tinakula in the Santa Cruz group and Kavachi, a submarine volcano near New Georgia, for example, have erupted regularly every few years, and Simbo Island has a solfatara (a volcanic area or vent that yields only hot vapours and sulfurous gases). Earthquakes and destructive cyclones also occur regularly.

The people. Most of the population is Melanesian. Polynesians, forming a small minority, live mainly on outlying atolls, principally Ontong Java, Rennell and Bellona, the Reef Islands, Sikaiana, Tikopia, and Anuta. There are also Gilbertese from Micronesia, who were resettled there between 1955 and 1971, and small numbers of Chinese and Europeans. Anglican, Roman Catholic, South Sea Evangelical, United (a union of Methodists and Congregationalists), and Seventh-day Adventist denominations exist in Solomon Islands, and almost all of the islanders are Christian. More than 60 languages and dialects are spoken. English is the official language, but Pijin, an English-based pidgin, is the one most widely used and understood.

Most of the people live in small rural villages. They engage mainly in subsistence gardening, pig raising, and fishing but are also involved in the cash economy.

The economy. Solomon Islands' main resources, fish and timber, have been exploited excessively, with little regard for their depletion. Its other export products are derived from plantation crops: palm oil from oil palms, copra from coconuts, and cocoa from the cacao plant. Japan, Thailand, and Australia are the major customers. The chief imports are machinery, fuels, manufactured goods, and food, and Australia is the main supplier. A government plan to reduce food imports by growing rice locally was scuttled in 1986 when a cyclone destroyed the rice cultivations on Guadalcanal.

The islands have significant reserves of bauxite (on Rennell Island) and phosphates (on Bellona Island), and some gold has been extracted on Guadalcanal. Manufacturing includes fibreglass products, boat building, batteries, and clothing, in addition to fish freezing and canning. Traditional handicrafts, including woodwork, shell inlay, mats, baskets, and shell jewelry, are made both for the tourist market and for export. Indigenous currencies such as shell money (from Malaita) and red-feather money (from Santa Cruz) are also made for use in customary transactions. Henderson Field, near Honiara, is the principal international airport, although airfields at Munda, Nusatupe (at Gizo), and Graciosa Bay (at Santa Cruz) may also serve as international points of entry. Solomon Islands Airlines, which is owned by the government, provides domestic air service. Ports handling overseas cargoes include Honiara, Tulagi (the prewar capital), and Gizo. Aola Bay, Viru, and Graciosa Bay are used mainly for log exports. Interisland shipping is operated both privately and by the government.

Traditional handicrafts

Administration and social conditions. Solomon Islands is an independent parliamentary state and a member of the Commonwealth. As established by the constitution of 1978, the British monarch is the head of state and is represented by a governor-general who is appointed by the monarch upon recommendation by Parliament. The governor-general must be a citizen of Solomon Islands. Parliament, which is elected every four years by universal adult suffrage, is composed of 38 members. Executive power is exercised by a prime minister chosen by and from Parliament and a cabinet of 12 ministers (also members of Parliament). Although political parties exist in name, their organization and discipline tend to be loose. Political loyalties are based on personal and local factors rather than on broad class interests or ideology. As a result, the prime minister rarely commands a clear majority in Parliament, and so he owes his position to the sometimes shaky support of a coalition of parties or factions. In 1981 considerable decentralization occurred with the establishment of seven provincial bodies that assumed control of matters regarding transportation, economic development, health, and education.

Before World War II, schooling was solely the concern of the missions. Since then the government has become involved. Primary education was provided mostly by subsidized mission schools until 1974, when local authorities assumed responsibility. There are several secondary schools, a multi-faculty college of higher education, and a branch (1971) of the University of the South Pacific. Some students study overseas, especially in Fiji and Papua New Guinea. There is a hospital at Honiara.

Much of traditional culture endures. Crafts are promoted by the Solomon Islands National Museum, established in 1969 in Honiara, and dances and music are regularly performed. Panpipes and percussive "bamboo bands" (assemblages of drums made of bamboo) are popular. Since the 1970s an indigenous literary movement has developed. The writing is in English and mostly published in Solomon Islands.

For statistical data on the land and people of Solomon Islands, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The Solomon Islands were initially settled by at least 2000 BC—well before the archaeological record begins—probably by people of the Austronesian language group. Lapita pottery was in use in Santa Cruz and the Reef Islands about 1500 BC. Material radiocarbon dated to about 1000 BC has also been excavated at Vatuluma Cave (Guadalcanal), Anuta, Santa Ana, and Tikopia.

The first European to reach the islands was the Spanish explorer Álvaro de Mendaña de Neira, in 1568. Subsequently, unjustified rumours led to the belief that he had not only found gold there but had discovered where the biblical King Solomon obtained the gold for his temple in Jerusalem. The islands thus acquired the name Islas de Solomón. Later Spanish expeditions to the southwest Pacific, in 1595 and 1606, were unable to confirm the discoveries reported by Mendaña. Geographers came to doubt the existence of the group, and it was not until the late 18th century, after further sightings by French and English navigators, that the Solomons were accurately charted. After the settlement by the English of Sydney in 1788, naval and commercial shipping began increasingly to pass through the Solomons' waters.

The naming of the Solomons

Roman Catholic missionaries in the 1840s failed to establish a settlement but did so in 1898. Anglican missionaries, who had been taking islanders to New Zealand for training since the 1850s, began to settle in the Solomons in the 1870s. Other missions arrived later.

By the late 19th century the islands were being exploited for labour to work the plantations of Fiji and other islands and of Queensland, Australia. About 30,000 labourers were recruited between 1870 and 1910. To protect their own interests Germany and Britain divided the Solomons between them in 1886; but in 1899 Germany transferred the northern islands, except for Buka and Bougainville, to Britain (which already claimed the southern islands) in return for recognition of German claims in Western Samoa and parts of Africa. The British Solomon Islands Protectorate was declared in 1893, partly in response to abuses associated with labour recruitment and partly to regulate contacts between islanders and European settlers, but mainly to forestall a threat of annexation by France. Colonial rule began in 1896. Although generally humane, administrators were more concerned with promoting the interests of European traders and planters than with those of the islanders, and islanders were punished harshly for offenses against colonial law and order. The murder of government tax collectors on Malaita in 1927 was answered with a savage punitive expedition, which was backed by an Australian warship. Together with some of his associates, Basiana, the leader of the assassins, was later hanged, and his young sons were forced to witness the execution.

With the outbreak of World War II in the Pacific, the Japanese began occupying the protectorate early in 1942, but their advance farther southward was stopped by U.S. forces, who invaded on August 7. Fighting in the Solomons over the next 15 months was some of the most bitter in the Pacific; the long struggle on Guadalcanal was the crucial battle of the Pacific war. Throughout the campaign the U.S. forces and their allies were strongly supported by the islanders. At the end of the war, because of the proximity of an airfield and the availability of flat land and of the military's buildings, Honiara on Guadalcanal became the new capital, replacing Tulagi.

Another result of the war was to stimulate political consciousness among the islanders and so inspire a nationalist movement known as Maasina Rule, which lasted from 1944 to 1952. Subsequently, in response to the worldwide movement for decolonization, the Solomons set out on the path of constitutional development. The nation was formally renamed Solomon Islands in 1975, and independence was attained on July 7, 1978. (H.M.La.)

For later developments in the history of Solomon Islands, see the BRITANNICA BOOK OF THE YEAR.

Vanuatu

Vanuatu is a republic consisting of a chain of 13 principal and many smaller islands in the southwestern Pacific Ocean, 500 miles (800 kilometres) west of Fiji and 1,100 miles east of Australia. The islands extend from north to

south for 400 miles in an irregular Y shape and have a total area of 4,707 square miles (12,190 square kilometres); they include Vanua Lava and Gaua (Santa Maria) in the Banks and Torres groups; and Espíritu Santo, Aoba, Maéwo, Pentecôte (Pentecost), Malakula, Ambrym, Épi, Éfaté, Erromango, Tanna, and Anatom. Formerly the jointly administered Anglo-French condominium of the New Hebrides, Vanuatu achieved independence in 1980. The capital, largest city, and commercial centre of the republic is Vila, on Éfaté Island.

PHYSICAL AND HUMAN GEOGRAPHY

The land. A diverse relief—ranging from rugged mountains and high plateaus to rolling hills and low plateaus, with coastal terraces and offshore coral reefs—characterizes the islands. Sedimentary and coral limestones and volcanic rock predominate; frequent earthquakes indicate structural instability. Active volcanoes are found on several islands, including Mount Séré'ama on Vanua Lava, Mount Garet on Gaua, Mount Benbow and Mount Marum on Ambrym, and Mount Yasur on Tanna. There are also several submarine volcanoes in the group, and some islands have solfataras or fumaroles. The highest point is Mount Tabwémasana, 6,165 feet (1,879 metres), on Espíritu Santo, the largest island. The climate divides into two seasons—hot and wet from November to April and cooler and drier from May to October. The southeast trades are the prevailing winds, although northerlies during the hot season provide most of the heavy rainfall. Rainfall varies from 80 inches (2,000 millimetres) in the south to 160 inches in the northern islands. Much of the group is covered by dense rain forest, but drier regions have patches of savanna grassland. Abundant bird and insect life contrasts with the sparse fauna. Of the 11 types of bats found in Vanuatu, three are found only there.

Volcanoes

The people. The indigenous population, called ni-Vanuatu, is overwhelmingly Melanesian, though some of the outlying islands have Polynesian populations. More than 100 local Melanesian languages and dialects are spoken; Bislama, an English-based pidgin, is the national language and, along with English and French, is one of three official languages. Religions include Christianity, traditional ancestor worship, and cargo cults.

The economy. Subsistence agriculture has traditionally been the economic base of Vanuatu, together with an elaborate exchange network within and between islands. Economic changes occurred with the development of European plantations in the island group after 1867: cotton was the initial crop, followed by corn (maize), coffee, cocoa beans, and coconuts (for copra). Cattle ranching was instituted later. By the 1880s French planters had reversed the initial British domination of the plantation sector, though they, too, found it increasingly difficult to compete with ni-Vanuatu producers, who could fall back on subsistence agriculture in times of economic downturn. French hopes of economic hegemony, based on high world prices for copra and the importation of Vietnamese labour in the 1920s, were dashed by the worldwide depression of the 1930s. By 1948 most of the copra in the island group was being produced by the ni-Vanuatu themselves, though it was not until the development of cooperatives in the 1970s that they were finally able to assume control of the trade.

Copra, cocoa, fish, and beef remain the most important exports. Imports, of food and manufactured goods, come principally from Australia. Though Vanuatu looks toward large-scale agricultural development for its long-term economic growth, since independence tourism has emerged as the largest earner of foreign income.

On most of Vanuatu's islands, roads link coastal settlements; there are few interior roads. Interisland transportation is by boat or airplane. Bauerfield airport, near Vila, and Pékoa airport, near Luganville on Espíritu Santo, are international airports, and many smaller airfields are scattered throughout the islands.

Administration and social conditions. The constitution of 1980 provides for a president, elected to a five-year term by an electoral college comprising the parliament and the presidents of the Regional Councils. Members of the parliament are elected to four-year terms on the basis of universal franchise. The parliament elects the prime minister from among its members, and the prime minister then appoints the Council of Ministers from among members of the parliament. The constitution also provides for a National Council of Chiefs, composed of "custom chiefs," which is concerned with matters relating to custom and tradition. Island government councils have taken the place of the Regional Councils proposed in the constitution. The Supreme Court hears and determines civil and criminal proceedings.

Although attempts have been made since independence to institute a single, and English-speaking, education system, much French economic aid has been channeled into the maintenance of the Francophone school system, and thus French aid has ensured that the 25 percent of ni-Vanuatu children who are French-speaking attend French-language schools. Only a small proportion of ni-Vanuatu children undertake post-primary education, but with more than 300 primary schools, the nation's elementary school attendance is one of the highest in the Pacific.

The education system

Health care in Vanuatu consists of two hospitals in Vila supplemented by smaller hospitals, clinics, and dispensaries on the other islands. Malaria, tuberculosis, hookworm, and gastroenteritis are the most common diseases.

Cultural life. The overwhelming majority of ni-Vanuatu are subsistence agriculturalists, living in small rural villages where activities revolve around the land. The constitution has guaranteed that land cannot be alienated from its "indigenous custom owners," or traditional owners, and their descendants. More than an economic resource, land is the physical embodiment of the metaphysical link with the past, and identification with a particular tract of land (expressed by the Bislama phrase *man ples*) remains one of the fundamental concepts governing ni-Vanuatu culture. On many islands the men of a clan still gather nightly at their local *nakamal* to drink kava and communicate with the spirits of their ancestors, whose bones typically are buried nearby. Through magic stones they will contact and try to control the all-pervasive spiritual realm. Among the vast majority of rural dwellers *kastom* (custom), along with Christianity, continues to guide daily life.

For statistical data on the land and people of Vanuatu, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Archaeological evidence indicates that by 1300 BC islands in northern Vanuatu were settled by the makers of the distinctive Lapita pottery from Melanesian islands to the west. Since then, there have been successive waves of migrants, including people of Polynesian origin on the southern islands of Futuna and Aniwa. Around AD 1200, a highly stratified society developed in central Vanuatu with the arrival (from the south, according to tradition) of the great chief Roy Mata (or Roymata). His death was marked by an elaborate ritual that included the burying alive of one man and one woman from each of the clans under his influence. European contact began with the Portuguese explorer Pedro Fernández de Quirós (1606), followed by the French navigator Louis-Antoine de Bougainville (1768) and the British captain James Cook (1774). Cook mapped the group and named it the New Hebrides. European missionaries and sandalwood traders settled on the fringes of islands from the 1840s, but their impact on the indigenous people was minimal. Significant cultural change occurred only after the 1860s, as thousands of ni-Vanuatu men and women who had been indentured to work on plantations in Fiji and New Caledonia, and in Queensland, Australia, began to return to their homes. Many established new forms of political influence within the network of Protestant (mainly Presbyterian) missions or successfully competed against European traders and planters in the group. To protect the interests of the mainly British missionaries and mainly French planters, the British and French governments established rudimentary political control with a Joint Naval Commission in 1887. This was succeeded in 1906 by an Anglo-French condominium, under which resident commissioners in the capital, Vila, retained responsibility over

The return of indentured workers

their own nationals and jointly ruled the indigenous people. This clumsy administrative arrangement had only a slight impact, however, on most ni-Vanuatu, whose chief European contact continued to be with either missionary or planter. The islands became a major Allied base during World War II, when the spectacle of free-spending black American troops inspired the transformation of the John Frum cargo cult on Tanna into an important anti-European political movement. After the war, local political initiatives originated in concern over land ownership. At this time more than one-third of the New Hebrides continued to be owned by foreigners.

Independence was agreed upon at a 1977 conference in Paris attended by British, French, and New Hebrides representatives. Elections were held and a constitution was drawn up in 1979. Despite an unsuccessful attempt in mid-1980 by Jimmy Stevens, the Na-Griamel Party leader, to establish Espíritu Santo Island as independent from the rest of the group, the New Hebrides became independent within the Commonwealth under the name of the Republic of Vanuatu ("Our Land Forever") on July 30, 1980. (R.Ad.)

For later developments in the history of Vanuatu, see the BRITANNICA BOOK OF THE YEAR.

MAJOR POLITICAL UNITS OF MICRONESIA

Guam

Guam, the largest, most populous, and southernmost island of the Mariana Islands and an unincorporated territory of the United States, lies in the Pacific Ocean about 5,800 miles (9,300 kilometres) west of San Francisco and 1,600 miles east of Manila. The total land area is 209 square miles (541 square kilometres). Agana is the capital. The most populous settlements are Tamuning, Apra Harbor, and Mangilao.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The island is sharply divided into a northern limestone plateau with a general elevation of about 500 feet (150 metres) and an area of higher, volcanic hills to the south. The plateau is covered with a thick growth of jungle; the volcanic hills support mainly sword grass. The hills rise more than 1,000 feet above sea level, and their lower slopes to the east (and also in part to the west) are covered with younger limestones, generally similar to those of the northern limestone plateau. The higher hills are Mount Lamlam, which rises to an elevation of 1,332 feet (406 metres); Mount Bolanos (1,220 feet); and Mount Sasalaguan (1,109 feet).

Guam has a pleasant tropical climate. Temperatures range from 70° to 90° F (21° to 32° C) and are fairly even throughout the year. Average annual rainfall is about 95 inches (2,400 millimetres), three-fourths of which falls during the wet season, generally starting in May or June and lasting through November. The even climate is punctuated by destructive typhoons that occur at irregular intervals.

The people. The native Guamanians, ethnically called Chamorros, are basically of Malayo-Indonesian descent with a considerable admixture of Spanish, Filipino, Mexican, and other European and Asian influences. Their vernacular, called the Chamorro language, is a distinct language with its own vocabulary and grammar mixed with many Spanish words. The word Chamorro is derived from Chamorri, or Chanioli, the ancient name for chief. English and Chamorro are the official languages; although Chamorro is still used in many homes, English is the language of education and commerce. The predominant religion is Roman Catholicism.

Before World War II the villages were the main social and economic units, preserving customs and traditions similar to those of 19th-century Spain. The fiesta held in memory of a patron saint was the great social and religious event of the year for each village, and it brought together people from other villages. Fiesta customs are still observed in modern Guam. Changes in the social life and institutions of Guamanians, however, have come about with economic improvement and increasingly international contacts.

Fiesta tradition

The economy. The development of Guam into an important U.S. military base brought about profound changes in the island's agricultural patterns after World War II. Foodstuffs were imported in increasing amounts, discouraging the cultivation of many formerly important truck crops. Guam now imports most of its food.

All branches of the U.S. armed forces are represented at military facilities on Guam. Andersen Air Force Base, at the northern end of Guam, is a part of the U.S. Strategic Air Command; the armed forces also operate a naval station with a ship repair yard, a naval air station, a communications station, and a hospital. The availability of work at the military facilities has drawn many islanders from their former lives of subsistence agriculture and fishing.

Tourism is the most prominent component of the economy after the military, with more than half a million tourists visiting from Japan alone in 1987. There are several luxury hotels along Tumon Bay, which has been highly developed as a tourist area. A modern international airport links Guam with Japan, Hawaii, the U.S. West Coast, the Philippines, Australia, and other parts of Asia. Guam is a duty-free port, and this status has attracted a number of small manufacturing companies from countries in Asia and has also prompted some immigration. In the 1970s poultry farming, garment-finishing plants, and an oil refinery were established. Major imports are petroleum products, machinery, automobiles and transport vehicles, manufactures, and food.

Administration and social conditions. Guam is an unincorporated territory of the United States governed under the Organic Act of Guam, passed by the U.S. Congress and approved by the president on Aug. 1, 1950. The Organic Act made all Chamorros U.S. citizens, but they were not given the right to vote in national elections.

The act was amended in 1968 to provide for the popular election of a governor and lieutenant governor for four-year terms. All persons 18 years of age or older are permitted to vote. The legislature is a unicameral body with 21 senators directly elected at large for a term of two years. Guam has no constitution. The people of Guam, in seeking to improve their political relationship with the United States, voted in 1982 in favour of pursuing a commonwealth relationship similar to that established in the Northern Marianas. A draft Commonwealth Act was approved by the people in 1987, and negotiations with the U.S. Congress were initiated.

The judiciary consists of the Federal District Court of Guam, whose judge is appointed by the U.S. president for a term of eight years. There are two levels of local trial courts: the Superior Court of Guam, for criminal and civil cases, and the traffic, juvenile, and small-claims courts. Judges are appointed by the governor with consent of the legislature and are reconfirmed by majority public vote every four years. Appeals may be made to the U.S. Court of Appeals for the Ninth Circuit and to the U.S. Supreme Court.

Village commissioners

Each of the island's villages is headed by a popularly elected commissioner. A chief commissioner, elected by the commissioners, acts as liaison between the governor and the districts.

Education is compulsory between the ages of six and 16. The University of Guam, opened in 1952, is a four-year institution with colleges of agriculture, business and public administration, arts and sciences, and education. Graduate programs at the master's level are also available. Health facilities include public, private, and military hospitals and local clinics.

Cultural life. Guam is culturally diverse, with Chinese, Japanese, Filipino, Korean, and other Asian communities of significant size in addition to its indigenous and mainland U.S. populations. As a centre of transportation and communication for the island region it also has sizable

numbers of islanders from various parts of Micronesia, such as Palau, Yap, Truk, Pohnpei, Kosrae, and the Marshall Islands. The extended family is the main social unit for most groups on Guam, although many of the younger members travel and live in the United States. Folk arts and handicrafts of Guam have enjoyed a revival in recent years with the advent of various public and private groups that have promoted music, dance, and other traditional cultural arts for the benefit of both the local community and tourists.

U.S. national holidays are celebrated on the island, as are several significant local dates such as Discovery Day, March 6, which commemorates the arrival of Ferdinand Magellan in 1521. The museum and library on Guam have active programs for the community, and the university is also a promoter of regional arts and culture. There are daily and semiweekly newspapers and quarterly and monthly magazines published on Guam, and several radio and television stations broadcast local and international news and features daily.

For statistical data on the land and people of Guam, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Guam, like the other Mariana Islands, was settled before the second quarter of the 2nd millennium BC by an Indonesian–Filipino people. Archaeological research shows that by AD 800 they had developed a complex society that erected elaborate stone pillars (*halege*), which served as supports for communal houses (*latte*).

Ferdinand Magellan probably landed on Guam in 1521. The island was officially claimed by Spain in 1565, and there was no attempt to conquer it until the latter part of the 17th century. After an uprising in 1670 and 25 years of intermittent warfare, the Spanish subdued the population with considerable bloodshed. Disease, particularly smallpox and influenza, also played an important role in the decimation of the population. Typhoons in 1671 and 1693 caused further destruction and loss of life. Guam remained a Spanish possession until 1898, when, in the course of the Spanish-American War, the U.S. warship *Charleston* steamed into the harbour of Apra and shelled the old fort. Guam was ceded to the United States, and the other islands of the Marianas were sold by Spain to Germany in 1899. From that time until 1950 (except for the period of its occupation by the Japanese during World War II) the governor of the island was a naval officer appointed by the president of the United States.

During World War I Japan occupied Germany's island possessions north of the Equator, and after the war Japan received the Marianas (except Guam) and the Caroline and Marshall groups as a mandate under the League of Nations. Japan retained possession of the islands even after it withdrew from the League in 1933.

In World War II the Japanese landed on Guam just after the attack on Pearl Harbor and occupied the island by Dec. 12, 1941. U.S. forces retook Guam by Aug. 10, 1944. It was a major air and naval base for the squadrons of bombers that attacked Japan near the end of the war. Under the jurisdiction of the U.S. Navy, it was made a territory (1950) that was administered by the U.S. Department of the Interior. From 1973 Guam was administered by the Office of Territorial Affairs within that department. Guam remains the site of major U.S. naval and air bases; about one-third of the land in Guam is owned by the U.S. armed forces.

In the 1970s Guam gradually began to move toward representative self-government. The first popularly elected governor ran for office in 1970, and in 1972 Guam was given the right to send one nonvoting delegate (entitled to vote in committees, however) to the U.S. House of Representatives. In 1978 the U.S. Senate accorded Guam the right to adopt a territorial constitution. In 1982, in a referendum offering six options, the option of commonwealth status won a plurality (48 percent) of votes.

(D.A.B.)

For later developments in the history of Guam, see the BRITANNICA BOOK OF THE YEAR.

Kiribati

Kiribati is a republic of low-lying coral islands in the central Pacific Ocean. Of its 33 islands, only 20 are inhabited. The population is concentrated in the 16 Gilbert Islands, which straddle the Equator. From the Gilbert Islands, Kiribati extends 250 miles (400 kilometres) to the west to Banaba (formerly Ocean Island) and 1,800 miles to the east, to the Line Islands, three of which are inhabited. In between lie the islands of the Phoenix group, which have no permanent population. The capital is at Bairiki on Tarawa atoll in the northern Gilberts. Kiribati and Tuvalu were formerly joined as the Gilbert and Ellice Islands Colony.

PHYSICAL AND HUMAN GEOGRAPHY

The land. A few of the islands are compact with fringing reefs, but most are atolls. The largest atoll (and one of the largest in the world) is Kiritimati (Christmas Island) in the Line group, which has a land area of 150 square miles (388 square kilometres) and accounts for about one-half of the country's total. Kiritimati was used for U.S. and British nuclear testing in the 1960s; it now has a large coconut plantation and fish farms. Banaba rises to 265 feet; its rich layer of phosphate was exhausted by mining from 1900 to 1979. It is sparsely inhabited. Average rainfall in the Gilbert group ranges from 120 inches (3,000 millimetres) in the north to 20 inches in the south. Most rain falls in the season of westerly winds, from November through March; from April to October, northeast trade winds prevail. Temperatures are usually in the range of 80° to 90° F (27° to 32° C).

The importance of coconuts

Coconut palms dominate the landscape on each island. With the products of the reef and the ocean, coconuts are the major contributors to village diet—not only the nuts themselves but also sap tapped from the fruit-bearing spathes. The resulting "toddy" is used in cooking and as a sweet beverage; fermented, it becomes a powerful intoxicant. Breadfruit and pandanus also are grown. *Cyrtosperma chamissonis,* a coarse tarolike plant, can be cultivated in pits, but plants such as taro, bananas, and sweet potatoes are scarce. Pigs and chickens are raised.

The people. The people are Micronesian and speak a single language, which has 13 sounds; "ti" is pronounced as "s" or "see"; thus Kiribati, pronounced "Kiri-bas," is the local rendition of "Gilberts." English is widely spoken, especially on Tarawa. Half of the population is Roman Catholic, and most of the rest is Congregational Protestant.

For many years the population of most islands has remained static because of migration to the rapidly growing urban centres of South Tarawa, where one-third of the population now lives. Betio, the port and commercial centre, has a population density of 13,000 per square mile (5,000 per square kilometre), most living in single-story accommodations. The rural population lives in villages dominated by Western-style churches and large open-sided thatched meetinghouses. Houses of Western-style construction are seen on outer islands and are common on Tarawa.

The economy. Until 1979, when Banaba's phosphate deposit was exhausted, Kiribati's economy depended heavily on the export of phosphate. Before the cessation of mining a large reserve fund was accumulated; the interest now contributes to government revenue. Other revenue earners are copra, mostly produced in the village economy, and license fees from foreign fishing fleets. An Exclusive Economic Zone of 1,350,000 square miles is claimed. Kiribati depends on foreign aid for most capital and development expenditure. Food accounts for one-third of all imports. The main shipping links are with Australia and Fiji. While South Tarawa has an extensive wage economy, most of the people living on outer islands are subsistence farmers with small incomes from copra, fishing, or handicrafts. These are supplemented by remittances from relatives working elsewhere. Interisland shipping is provided by the government, and most islands are linked by a domestic air service. Tarawa and Kiritimati have major airports.

Administration and social conditions. Kiribati is a member of the Commonwealth. It has a president chosen by national election from among four candidates chosen by

and from members of the legislature (who are also popularly elected, for a four-year term). The president chooses his ministers from the legislature, and they remain responsible to it. One seat is reserved for the people of Banaba, who have been resettled in Fiji. Local government is through elected island councils.

Primary education is compulsory, and selected pupils attend government- or church-run secondary schools. Kiribati conducts its own technical and teacher training, but other higher education is by overseas scholarship. Tarawa has two hospitals, and there are clinics on all of the inhabited islands.

Cultural life. Kiribati society remains conservative and resistant to change; ties to family and traditional land remain strong, and conspicuous displays of individual achievement or wealth are discouraged. The building and racing of sailing canoes is a popular pastime. Musical composition and dancing in customary styles are regarded as art forms and are the basis of widespread competition. Volleyball and soccer are now widely played.

For statistical data on the land and people of Kiribati, see the *Britannica Book of the Year*.

HISTORY

The first settlers in the Gilbert Islands and Banaba came from Southeast Asia, by way of Micronesia, some 4,000–5,000 years ago. About the 14th century AD the southern islands received an influx of Samoans, and soon thereafter the islanders adopted a gerontocratic style of government (based on rule by elders). The Line and Phoenix islands had no prehistoric population.

Rule by elders

Spanish explorers sighted some of the islands in the 16th century, but most did not appear on charts until the early 19th century, when first whalers and then coconut-oil traders reached the islands. From the mid-19th century, labour was recruited in the Gilberts for plantations elsewhere in the region.

The Gilbert Islands became a British protectorate in 1892, and Banaba was annexed in 1900 after the discovery of its rich phosphate deposits. Both were linked with the Ellice Islands as the Gilbert and Ellice Islands Colony from 1916; the colony subsequently was extended to include most of the Phoenix and Line groups and, for a time, Tokelau. Administration was through island governments, which sometimes became enmeshed in sectarian rivalries between Roman Catholics and Protestants. As a response to drought and perceived overpopulation in the 1930s, a resettlement scheme was initiated for the Phoenix Islands; a later scheme saw other Gilbertese resettled in the Solomons.

During World War II the Gilberts were occupied by the Japanese, who were later ejected by American forces. The colony had few services until aid-funded development programs were introduced after the war. An elected House of Representatives was established in 1967. The subsequent emergence of racial tensions led to the division of the colony in 1975–76. In 1976 the British High Court rejected Banaban claims for massive compensation over mining issues. The Gilbert Islands achieved independence as Kiribati in 1979.

After independence a high priority was given to economic development, especially the exploitation of marine resources. Kiribati belongs to the South Pacific Forum and the Commonwealth. (B.K.M.)

For later developments in the history of Kiribati, see the *Britannica Book of the Year*.

Marshall Islands

The Republic of the Marshall Islands, a self-governing republic, consists of some of the easternmost islands of Micronesia. The Marshalls are composed of two parallel chains of coral atolls—the Ratak, or Sunrise, to the east, and the Ralik, or Sunset, to the west. The chains lie about 125 miles (200 kilometres) apart and extend some 800 miles northwest to southeast. The islands and islets number more than 1,200 and have a total land area of 70 square miles (181 square kilometres). Dalap-Uliga-Darrit, on Majuro Atoll, is the capital of the republic.

The Marshalls were administered by the United States as part of the U.S.-administered Trust Territory of the Pacific Islands from 1947 to 1986, when the Trust Territory was dissolved by the U.S. government.

PHYSICAL AND HUMAN GEOGRAPHY

The land. None of the 29 low-lying coral atolls and the five coral islands in the Marshall group rises to more than 20 feet (six metres) above high tide. The islands are coral caps set on the rims of submerged volcanoes rising from the ocean floor. The island units of the Marshalls are scattered over about 180,000 square miles of the Pacific. The largest atoll in the group and in the world is Kwajalein, which has a land area of only six square miles but surrounds a 655-square-mile lagoon. The Marshall Islands' nearest neighbours are Wake Island (north), Kiribati and Nauru (south), and the Federated States of Micronesia (west).

Scattered island group

The climate is tropical, with a mean annual temperature for the entire group of 82° F (28° C). Annual rainfall varies from 20 to 30 inches (500 to 800 millimetres) in the north to 160 inches in the southern atolls. The wettest months are October and November. Several of the northern atolls are uninhabited owing to insufficient rainfall. Most of the Marshall Islands are true atolls, consisting of an irregular, oval-shaped coral reef surrounding a lagoon; the islets lie along the coral reef. The islands and islets of the Ratak chain tend to be more heavily wooded than those of the Ralik. Coconut and pandanus palms and breadfruit trees are the principal vegetation. Soils are generally sandy and low in fertility.

The people. The native people of the Marshalls are Micronesians. The most populous atolls are Majuro and Kwajalein, which offers employment at the U.S. missile testing range; together they have more than two-thirds of the country's total population. The remaining one-third of the population lives in traditional villages on the outer islands away from the two urban centres. The American missionary effort began in the Marshalls in the 1850s and was exceedingly successful. The Marshallese today are predominantly Christian. The Marshallese and English languages are spoken, but only a minority are fluent in the latter.

The economy. Substantial U.S. subsidies to the Republic of the Marshall Islands under the Compact of Free Association (see below) and the leasing of land for the U.S. missile testing range on Kwajalein are the main sources of revenue for the nation. Employment and modern amenities at both Majuro and Kwajalein serve as magnets that draw people to the two urban centres.

On the outer islands, subsistence farming, fishing, and the raising of pigs and poultry are the principal economic activities. Coconut, pandanus, breadfruit, and taro are the major food crops. The production of copra is the chief source of income for the outer islands. The principal import is processed foods. Other major imports include machinery and transport equipment, manufactured goods, and fuels, primarily from the United States, Japan, and Australia.

Transportation among the atolls and islands is by boat or air. Government-owned ships make scheduled trips among the islands. Several commercial cargo lines also serve the islands. Majuro has a commercial dock complex, and many of the atolls have good anchorage within their lagoons. Majuro and Kwajalein have international airports, and domestic and regional flights link some of the other atolls and islands.

Administration and social conditions. Under the constitution adopted in 1979, the government consists of a president elected by a unicameral, 33-member parliament known as the Nitijela. The Council of Iroji (Chiefs) has mainly a consultative function, concerned with traditional laws and customs.

A hospital on Majuro, a small sub-hospital on Ebeye island (part of Kwajalein Atoll), and dispensaries on other islands provide health care. There are primary schools, both public and church-run, on the inhabited islands and islets. Majuro and Jaluit atolls each have a public secondary school. Majuro is also the site of a centre for

Education

continuing education operated by the Federated States of Micronesia's College of Micronesia.

HISTORY

The Marshall Islands were settled initially around the beginning of the Christian Era by Micronesians who may have been influenced by early Polynesian (Lapita) culture. Radiocarbon dates from earth-oven charcoal samples that were excavated in Laura village on Majuro yielded dates of about 30 BC and AD 50. The early Marshall Islanders were skilled navigators and made long canoe voyages among the atolls.

Sighted in 1529 by the Spanish navigator Álvaro Saavedra, the Marshalls lacked the wealth to encourage exploitation or mapping. The British captain Samuel Wallis chanced upon Rongerik and Rongelap atolls while sailing from Tahiti to Tinian. The British naval captains John Marshall and Thomas Gilbert partially explored the Marshalls in 1788, but much of the mapping was done by Russian expeditions under Adam Johann Krusenstern (1803) and Otto von Kotzebue (1815 and 1823). U.S. whalers frequented the islands from the 1820s, and U.S. and Hawaiian Protestant missionaries began efforts to convert the islanders in the 1850s. Germany established a coaling station on Jaluit Atoll by treaty with island chiefs and in 1886, by agreement with Great Britain, established a protectorate over the Marshalls. Japan seized the islands in 1914 and later (after 1919) administered them as a League of Nations mandate. Occupied by the United States in World War II, following heavy fighting at Kwajalein and Enewetak, the Marshall Islands were made part of the United Nations Trust Territory of the Pacific Islands under jurisdiction of the United States in 1947. After their populations were removed to other atolls, Bikini and Enewetak served as an official testing ground for U.S. nuclear bombs (1946–58). The tests stopped in 1958 and cleanup efforts began in the late 1960s. During the trial resettlement of the Bikinians, however, their atoll was found to be too contaminated for permanent habitation, and by the late 1970s the people had to be evacuated once again. The Enewetak people were returned to their homeland, and a program to monitor Bikini was put in place.

Compact of Free Association

After voting to separate from the other entities of the Trust Territory of the Pacific Islands, in 1978, the Marshall Islands drew up a constitution that voters approved in 1979. It formed the republic and brought internal self-government. In 1982 the government signed the Compact of Free Association with the United States. This agreement, approved by the voters in 1983, requires that the United States remain responsible for defense and external security and that it provide financial assistance for the republic. The compact entitles the United States to use the missile testing range on Kwajalein Atoll, and provides for the republic to become fully independent and to alter its status with the United States at any time, subject to approval of the residents through plebiscites. In 1983 the islanders reopened negotiations on the compact's provision regarding compensation for the people of the four atolls affected by the U.S. nuclear tests. The United States agreed to set up a separate fund for the people of the four atolls and established an open-ended fund to cover personal injury claims among the islanders; it also agreed to set up a joint U.S.-Marshallese claims tribunal. A further agreement set up a fund to improve living conditions on Ebeye island, where all Micronesians working on the Kwajalein missile range and base resided. The Trust Territory was dissolved in 1990 with the approval of the UN Security Council, and on Sept. 17, 1991, the Marshall Islands became a member of the UN. (R.C.Ki.)

Micronesia

The Federated States of Micronesia is an internally self-governing country in the western Pacific Ocean. It is composed of more than 600 islands and islets in the Caroline Islands archipelago and is divided, roughly along cultural and linguistic lines, into the states of Yap, Chuuk, Pohnpei, and Kosrae. The total land area is 271 square miles (701 square kilometres). The capital is Palikir, on the island of Pohnpei. To the west of the Federated States of Micronesia lies the Republic of Palau, also in the Caroline archipelago, and to the east is the Republic of the Marshall Islands. These two nations, together with the Commonwealth of the Northern Marianas and the Federated States of Micronesia, were administered by the United States as the Trust Territory of the Pacific Islands from 1947 to 1986.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The islands are of two general types: high volcanic islands that support a large variety of plant forms and low-lying coral atolls with poorer soil. The main landmass in Yap is distinctive in that it was formed by folds in the Earth's crust and is continental in geologic composition. Each of the four states has as its population centre an elevated landmass of fairly large extent. These range in size from Pohnpei, with an area of 129 square miles, to Yap and Chuuk, each with 39 square miles. The volcanic islands of Chuuk are an anomaly in the Pacific since they are encircled by a reef but have not yet subsided beneath sea level to become a classic atoll. Yap Island and six islands in the Chuuk group rise to more than 500 feet (150 metres), and Kosrae and Pohnpei islands have peaks of 2,061 feet and 2,593 feet, respectively. The coral atolls consist of groups of small islands, formed atop a coral reef enclosing a lagoon.

The climate is tropical, with high humidity and a mean annual temperature of 81° F (27° C). There is little seasonal variation in temperature. Rainfall averages 120 inches (3,000 millimetres) per year throughout the area, although Pohnpei receives more than 200 inches yearly. There are distinct rainy and dry seasons, the latter occurring during the height of the northeasterly trade-wind season between December and April. Yap, which alone is situated in the monsoon area, has westerly winds for part of the year. Numerous typhoons originate in the east each year, usually spinning off to the northwest toward Yap and the Mariana Islands and seldom striking any of the other islands.

Coastal settlements

On the high islands, mangrove swamps grow along the shore, and grassland or scrub ascends to tropical rain forests in the interior mountain areas. Settlements are almost without exception located near the coast. Volcanic islands, with their richer soil, support many different species of plant life. On coral atolls the predominant forms of vegetation are the coconut palm and pandanus and breadfruit trees. Atoll dwellers typically locate their houses on the lagoon side of the island.

The people. The native people of the Federated States, while generally classified as Micronesian, are very diverse in physical types, cultures, and languages. The people of Yap Island have Melanesian-like features and speak a language only distantly related to the other languages of the area. Inhabitants of the coral atolls in Yap state are similar in language and culture to the people of Chuuk, although their languages are not mutually intelligible. Both Chuuk and Pohnpei contain several dialects, and the inhabitants of Kapingamarangi and Nukuoro, two atolls in the southwestern portion of Pohnpei state, are Polynesians and speak languages unrelated to Pohnpeian. Only Kosrae has complete ethnic and linguistic unity. Altogether eight local languages are recognized as distinct, and dialectal differences in the outlying atolls add further variety.

Almost half of the total population lives in Chuuk. Pohnpei has about 30 percent, and Yap and Kosrae have about 10 percent and 7 percent, respectively. In recent years there has been some migration to the towns for employment and education, and the four main towns—the capitals of each state—now contain about one-quarter of the total population. The birth rate remains high, and mortality rates have been sharply reduced. The population is growing by about 3 percent yearly. Almost the entire population is Christian. English is spoken widely in commerce and government throughout the nation.

The economy. Subsistence farming and fishing are the principal economic activities. The main staples are breadfruit, taro, coconuts, and bananas, but yams, cassava, and sweet potatoes also are grown on some of the high islands.

Poultry, pigs, and dogs are sometimes also raised for food. The main source of revenue is U.S. government grant aid, which continues under the Compact of Free Association signed in 1982 and which accounts for an estimated 70 percent of the gross national product. Government employment is by far the major source of income, and this in turn supports the service industries that make up the private sector. Exports, which account for less than 5 percent of the gross national product, include copra, a few marine products, black pepper, and handicrafts. Additional income is derived from the sale of fishing rights and from a slowly expanding tourist industry. Local commercial fishing is developed only on a small scale, and large-scale agriculture has repeatedly proved a failure. Much of the Federated States' food, including rice, fruits, vegetables, meat, and fish, and almost all manufactured goods and fuel oil must be imported.

Interisland and international transportation are by air or ocean freighter. The major commercial ports are at Takatik near Pohnpei, Kosrae, Colonia on Yap, and Moen. International airports are located only on Yap, Pohnpei, and Moen. Smaller ships operated by the government provide service to the outlying atolls in each state. Some of the atolls are also linked by weekly flights on smaller airplanes. On the major islands all villages are connected by a road system, and the primary roads in the commercial centres are paved.

Administration and social conditions. Under the 1979 constitution, the government of the Federated States of Micronesia consists of a president and vice president elected by the unicameral National Congress, made up of one at-large senator from each state, with four-year terms, and single-district senators (elected on the basis of population), with two-year terms, including one each from Yap and Kosrae, three from Pohnpei, and five from Chuuk. Each of the four local state governments has a popularly elected governor and a unicameral legislature. There are no political parties. The judicial system consists of the Supreme Court of the Federated States of Micronesia, state courts in each of the four states, and municipal courts in each political subdivision of each state. Recognized customary law is in force wherever it does not conflict with constitutional law.

Education

There are elementary schools on every island, and each state has at least one public high school. The Community College of Micronesia, with its main campus on Pohnpei and extension programs in other states, is the only institution for higher education in the Federated States of Micronesia, but several U.S. universities offer extension programs for teacher training. A hospital is maintained in each of the state capitals, and dispensaries have been built on all major islands. Local healers continue to practice massage and other traditional medical skills.

Cultural life. Although only the people of Yap and a few outlying atolls in Chuuk still wear traditional dress (loincloths for men and grass skirts or lavalavas for women), most villagers live much as their ancestors did. They support themselves by subsistence farming and fishing, use traditional food-preparation techniques, and gather in community meetinghouses for ceremonies and leisure activities. Islanders from some of the remote atolls have maintained their skills in the construction and navigation of large single outrigger canoes. The seamanship of the islanders is probably their most remarkable achievement and has astonished Westerners since initial contact. Wood carving, originally practiced for religious purposes, has become a commercial enterprise. Women in some of the islands weave lavalavas on traditional belt looms, and nearly everywhere large sleeping mats are plaited of pandanus leaves. Tattooing, once used widely to indicate social status, is seldom practiced today, but the traditional dances remain a favourite form of entertainment.

For statistical data on the land and people of Micronesia, see the *Britannica Book of the Year.*

HISTORY

The eastern Caroline Islands, like the Marshall and Gilbert islands, were probably first settled from the New Hebrides–Fiji area in the millennium before the Christian Era. Archaeological and linguistic evidence suggests that the earliest migrants worked their way up the chain of islands to the east and gradually spread westward from the Marshalls. Pottery dating to the beginning of the Christian Era has been found in Chuuk, and artifacts of similar antiquity have been unearthed in other islands in the east. The languages of the area, with the exception of the two Polynesian outliers and Yap, are closely related to one another and show striking similarities to Vanuatuan tongues. The high island of Yap appears to have been settled from the west, probably from the Philippines or Indonesia. Pottery and shell adzes found there date from the 2nd century AD and show a resemblance to types discovered in the Marianas. Yap's caste system and other features of its social organization are unparalleled elsewhere in the area. The renowned Yapese stone money was quarried in nearby Palau and the Marianas. In later centuries a reciprocal tribute and trade system was evolved with the surrounding coral atolls, sometimes termed the Yap Empire or Yapese Empire. In the eastern islands of Pohnpei and Kosrae, some seven centuries ago, a major social upheaval occurred, possibly under the impact of invasion from the south. Fortified settlements were built of huge basaltic logs, their ruins still visible today, and political authority over each island was centralized even as the society became more stratified.

Early contact with Europeans

The islands were visited occasionally during the 16th century by Spanish navigators and were later named for the Spanish king Charles II. During the 19th century European naval ships followed whalers and traders to the islands, and the inhabitants experienced their first intensive contact with foreigners and their wares. In 1886 Spain finally colonized the Carolines, but its short rule was ended in 1899 when the islands were sold to Germany at the conclusion of the Spanish-American War. Germany, in turn, yielded the islands at the outbreak of World War I to Japan, which gained formal title to them in 1920 as a League of Nations mandate. At first Japan attempted to develop a solid economy; later it used the islands as an outlet for surplus population; and finally it fortified them just before World War II. In July 1947 the Carolines, together with the Marshalls and northern Marianas, became a United Nations trust territory under U.S. administration. They were designated as a strategic area to allow the United States to set up military bases as deemed necessary; consequently, ultimate responsibility for the Trust Territory of the Pacific Islands rested in the UN Security Council rather than in the UN General Assembly, as was the case with other trust territories.

Micronesian leaders began negotiations with the United States for full self-government in 1969. The constitution that was drafted in 1975 became the basis for the federated government that went into effect in 1979. By that time, however, the Marianas, the Marshalls, and Palau had voted to separate from the proposed federation; this left only the "inner states" of Yap, Truk (later Chuuk), Kosrae, and Ponape (later Pohnpei) to compose the new political entity. In a 1983 plebiscite, voters approved a Compact of Free Association with the United States. The compact provides for internal self-government for the Federated States of Micronesia, while requiring that the United States remain responsible for defense and external security and that it provide financial assistance. The articles of the compact provide for the Federated States of Micronesia to become fully independent and alter its status with the United States at any time, providing this change of status is approved in plebiscites. The compact has a 15-year duration but is renewable by mutual consent. The Federated States of Micronesia became a member of the UN on Sept. 17, 1991. (F.X.H.)

For later developments in the history of Micronesia, see the *Britannica Book of the Year.*

Nauru

Nauru is an independent nation consisting of a solitary raised coral island located in southeastern Micronesia, 25 miles (41 kilometres) south of the Equator. It is about 2,480 miles northeast of Sydney, Australia, and a slightly

greater distance southwest of Honolulu. Nauru has a total land area of about 8 square miles (21 square kilometres) and a circumference of 12 miles. The capital is Yaren administrative centre.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Most of Nauru rises somewhat abruptly from the ocean, and there are no harbours or protected anchorages. A fairly fertile but relatively narrow belt encircles the island and surrounds the shallow inland Buada Lagoon. Farther inland, coral cliffs rise to a plateau 100 feet (30 metres) above sea level, with the highest point at about 213 feet. The plateau is largely composed of rock phosphate, leached from guano, or bird droppings. The mineral deposit covers 70 percent of the island, and its extraction has left irregular, pinnacle-shaped outcrops of limestone that give the landscape a forbidding, otherworldly appearance.

Otherworldly appearance of the landscape

© L. Zann/Australasian Nature Transparencies



Coral pinnacles harbouring phosphate, Nauru.

Nauru's climate is tropical, with daytime temperatures around 82° F (28° C), tempered by ocean breezes. Rainfall, averaging 81 inches (2,060 millimetres) annually, is extremely variable, and prolonged droughts occur. The only locally available water is collected from roof catchment systems, and water is imported as ballast on ships returning to Nauru for loads of phosphate. There are no rivers or streams.

Soils are generally poor and highly porous, and the irregular rainfall limits cultivation to the coastal belt and the lagoon's fringe. Subsistence crops, consisting mainly of coconut palms, pandanus, bananas, pineapple, and some vegetables, are not adequate to support the population. Varieties of plants and trees, however, number about 200. Nauru is a favourite stopover point for migratory birds, and chickens have been introduced. There was an absence of mammals until rats, mice, cats, dogs, and pigs were also imported.

The settlement pattern is dispersed. People are scattered along the coastal zone, and there is one small village, Buada, inland near the lagoon.

The people. Of the fewer than 10,000 residents of Nauru, about three-fifths are reported to be indigenous Nauruans. Other residents are employees of the government of Nauru or the Nauru Phosphate Corporation. A number of Europeans serve as administrators and technicians for the phosphate industry and some government agencies. There are a number of Chinese shopkeepers.

Nauruan is the national language. No adequate written grammar of the language has been compiled, and its relationships to other Micronesian languages are not well-understood. English is widely spoken.

Missionization came later to Nauru than to many other Pacific islands. The first Protestant evangelist arrived in 1899 and was followed three years later by the first Roman Catholic missionary. Today, the Nauru Congregational church claims about 60 percent of the population, with the remainder belonging to the Roman Catholic church.

The economy. The quality of Nauru's phosphate is the highest in the world. The mineral rock is Nauru's main resource and sole export, totally dominating the island's economy. The phosphate industry and government services together provide almost all of the island's salaried employment. Agriculture, with the exception of that along the coastal and lagoon perimeters, fishing, manufacturing, and tourism are of minor value to the overall economy.

Importance of phosphate mining

Phosphate has been mined on Nauru since 1907. For most of the 20th century, the phosphate industry was owned and operated by a British corporation. The government of independent Nauru gained control in 1970, and the Nauru Phosphate Company, an agency of the government, now manages the industry. The phosphate rock is crushed on the island and is transported by truck and narrow-gauge railway to a cantilever system that loads the product on ships moored to buoys in the open ocean.

On a per capita basis, Nauru is one of the wealthiest countries in the world. Landowners receive royalties from the phosphate earnings, and, by choice, many Nauruans are unemployed. Australia and New Zealand are the main customers for Nauru's phosphate. Australia supplies more than 90 percent of Nauru's imports. Most food, water, and manufactured goods are imported. With the exception of those levied on alcohol and tobacco, there are no import duties, and there is no income tax. By any standard, consumption levels are high, and much of the island's manual labour is done by outsiders.

Because it is projected that Nauru's phosphate deposits will be exhausted by about the year 2000, a major portion of the earnings are invested abroad. The future economic well-being of Nauruans depends on the success of the investment program, which includes a shipping line, an international airline, and extensive real-estate holdings in more than a half-dozen countries.

Nauru has its own banking system; the Bank of Nauru is wholly owned and operated by the government. As one consequence of its colonial history, Nauru is within the Australian monetary system, and Australian currency is the nation's legal tender.

Transportation on the island is good. A paved road system links all villages, and, reflecting the nation's high level of income, there is an abundance of motor vehicles of most varieties. Surface transportation to other destinations is difficult. Because there are no wharves or natural harbours, passengers and cargo are shuttled by barge between oceangoing vessels and a small artificial anchorage. Most regional and international travel is by air. The country launched its national airline, Air Nauru, in 1970, and in the mid-1980s the international airport with modern terminal facilities was expanded. Air Nauru has a fleet of Boeing jets.

Administration and social conditions. Nauru's constitution, implemented with independence in 1968, calls for broadly phrased fundamental rights and freedoms for individuals and a government that combines parliamentary and presidential systems. The parliament has 18 members. Unless dissolved by a vote of no confidence, it has a tenure of three years and elects the president, who is both head of state and head of government. The president appoints a cabinet of four or five members from the parliament. Nauru is a special member of the British Commonwealth, having all benefits of membership except the right to attend meetings of the Commonwealth heads of government.

The tripartite judicial system comprises a Supreme Court,

Tripartite judicial system

a District Court, and a Family Court. The Supreme Court, presided over by a chief justice, has both original and appellate jurisdiction. At Nauru's request, final appeals may be taken to the High Court of Australia.

Basic services in education and health are provided free to all citizens. Education is compulsory between the ages of 6 and 16. The government provides several kindergartens and elementary and secondary schools. The Roman Catholic mission has its own school system at the same three levels. Students go abroad, mainly to Australia, for higher education.

Nauru's wealth is a mixed blessing. The predominance of sedentary life-styles and high levels of consumption of food, alcohol, and tobacco have resulted in widespread personal health problems. The rates of obesity, high blood pressure, and diabetes are among the highest in the world, and, while vital statistics are not readily available, life expectancy for Nauruans is known to be lower than that for most other Pacific islanders.

HISTORY

When and from where the first inhabitants of Nauru reached the island remain unknown. A long period of relative isolation is believed to account for the distinctiveness of the indigenous language. By the time of the arrival of Europeans in the early 18th century, Nauruan society consisted of 12 matrilineal kinship groups, each having a chief.

An English sailing vessel sighted the island in 1798, but extensive contact with Europeans did not begin until the 1830s, when the whaling industry penetrated eastern Micronesia, and Nauru became a port of call for vessels in search of food and water supplies. Shortly thereafter, a small number of European beachcombers settled on the island, bringing with them alcohol, firearms, and foreign diseases. Intraisland warfare among competing districts escalated, becoming particularly intense in the 1880s. Encouraged by a few German traders concerned about their own interests on the island, Germany incorporated Nauru into its Marshall Islands protectorate in late 1888. The German administration and the arrival of the missionaries shortly thereafter brought an end to armed hostilities. In 1906 the Pacific Phosphate Company, a British concern, negotiated an agreement with the German administration to begin the mining of Nauru's phosphate deposits, and the operation began the following year.

With the onset of World War I, a small Australian force occupied Nauru and removed most German nationals. In 1920 Nauru became a mandated territory within the framework of the League of Nations. Australia, Britain, and New Zealand were named as the responsible authorities, but in actual practice the administration remained in Australian hands. The phosphate industry was taken over by the newly formed British Phosphate Commission, a joint Australian, British, and New Zealand enterprise.

World War II brought another occupier when Japanese forces arrived in August 1942. In the following year, 1,200 Nauruans were taken to Truk (now Chuuk) to serve as forced labourers on Japanese military installations there. A Japanese airstrip on Nauru became the target of American bombers, and the island suffered air attacks for the next two years. In September 1945, Australian troops again took possession of Nauru. On Jan. 31, 1946, with their numbers depleted by almost 500, 737 Nauruans were returned home.

United Nations Trust

In November 1947, Nauru became a United Nations trust, an arrangement paralleling the former League of Nations mandate. The same three metropolitan powers were the responsible authorities, but Australia continued to provide the actual administration.

A series of developments in the 1950s and particularly in the early 1960s led to self-government and eventually political independence and ownership of the phosphate industry. In October 1967 an agreement granting Nauruan independence was concluded. Jan. 31, 1968, the 22nd anniversary of the return of Nauruans from Truk, was chosen as Independence Day for the Republic of Nauru. Political parties are of lesser importance than personalities in Nauruan politics. One man, Hammer DeRoburt, has dominated the political scene, serving as president for most of the period since independence. (R.C.Ki.)

For later developments in the history of Nauru, see the BRITANNICA BOOK OF THE YEAR.

Northern Mariana Islands

The Northern Mariana Islands, a self-governing commonwealth in political union with the United States, is composed of 22 islands and islets in the western Pacific. It is a part of the Mariana Islands, a chain of volcanic mountain peaks and uplifted coral reefs extending from latitude 12° to 21° N and longitude 144° to 146° E. (The Marianas chain also includes the politically separate island of Guam.) The Northern Marianas has a land area of 184 square miles (477 square kilometres). Saipan (47 square miles), Tinian (39 square miles), and Rota (33 square miles) are the principal islands and, together with Alamagan and Agrihan, are inhabited. Another island, Pagan, was evacuated in 1981 after a severe volcanic eruption.

PHYSICAL AND HUMAN GEOGRAPHY

The land. Rota, the southernmost island, consists of a volcanic base capped with coral limestone, giving it a terraced appearance. Four southern islands (Farallon de Medinilla, Saipan, Tinian, and Aguijan) are composed of limestone and have gently rolling elevations and few mountains. The islands farther north are volcanic peaks. Mount Pagan erupted in 1922 and again in 1981, 1983, and 1988; Farallon de Pajaros (also called Urucas), the northernmost of the Marianas, and Asuncion are also active volcanoes. Agrihan volcano, the highest of the Northern Mariana group, rises to 3,165 feet (965 metres). Besides Guam, the nearest neighbours are the Bonin Islands (north) and the Federated States of Micronesia (southeast).

Tropical climate

The climate is tropical, with average yearly temperatures on Saipan ranging between 79° F (26° C) and 82° F (28° C) and annual rainfall averaging about 70 inches (1,800 millimetres). Typhoons strike the islands periodically. Rainfall is significantly less on the northernmost islands. The four limestone islands have tropical or scrub forests at higher elevations and coconut palms and casuarina trees along the coast, with the exception of Farallon de Medinilla, which is barren. Where level or gently sloping areas occur, cattle are grazed. The steep slopes of the volcanic islands from Guguan northward are mostly barren. The soils in these areas are generally shallow and low in fertility.

The people. The native people of the Northern Mariana Islands are Micronesians. Three-fifths of the population are Chamorros, descendants of the original inhabitants, who intermingled with Spaniards, Mexicans, Filipinos, and various other Europeans and Asians. About 14 percent of the population are Filipino. About 12 percent of the population are Carolinians, descendants of people who migrated from the central Carolines during the 19th century. There are small numbers of Europeans and Asians (other than Filipinos) as well.

Having lost most of their original Pacific Islands culture, the people of the Northern Mariana Islands have a mode of life that is Spanish Roman Catholic but is influenced by American culture. Saipan has more than 85 percent of the commonwealth's total population. Chamorro, related to Indonesian, is the principal language; English is the official language, and Japanese is widely used.

The economy. Tourism is the principal economic activity. Saipan and Rota are the main tourist centres and offer luxury hotels. The tourists are mainly Japanese and Americans. Subsistence farming, including the cultivation of taro, cassava, yams, breadfruit, vegetables, and bananas, is practiced extensively by many islanders to supplement their cash income. Copra production, fishing, pig and cattle raising, truck farming, and handicrafts have dwindled in recent years. Small manufacturing industries were introduced from Korea, China, and the Philippines, and such economic activity, with the continuing rise in tourism, has encouraged much immigration from Asia.

Saipan, Tinian, and Rota have paved roads. Transportation between the islands is largely by air, with some boat

traffic primarily for cargo. Saipan is the largest port, followed by Tinian and Rota. Saipan has an international airport.

Administration and social conditions. Part of the Trust Territory of the Pacific Islands granted to the United States by the United Nations in 1947, the Northern Marianas voted in plebiscite for status as a U.S. commonwealth in 1975. Some aspects of the commonwealth status were implemented in 1976, and the full commonwealth became effective upon the dissolution of the Trust Territory for the Marianas by the U.S. government in 1986. Eligible residents of the commonwealth became U.S. citizens at that time.

U.S. citizenship

According to the constitution of 1978, residents elect a governor, a lieutenant governor, and a bicameral legislature made up of nine senators and 14 representatives. The administrative headquarters of the commonwealth are at Chalan Kanoa on Saipan.

All inhabited islands have primary schools and hospitals. There are public and church-run secondary schools on Saipan. The College of the Northern Marianas, a public junior college, was established in 1976.

Cultural life. In recent years there has been an increase in the cultural diversity of the Northern Marianas. There are growing Filipino, Chinese, and Korean communities, although none of these has political representation. The Carolinian community, which has lived at Saipan for more than 100 years, is represented in the legislature. The local Chamorro population, with its tradition of extended families, is dominant. A local arts council has promoted folk arts and cultural events in the community, and a small museum and library have been established. Public beach parks and preserves as well as two large golf courses provide recreation. American national holidays and several local holidays are celebrated. Two local weekly newspapers are supplemented by the Guam daily newspaper, which provides some local coverage; all are published in English. There are several radio and television stations.

HISTORY

Archaeological evidence at Chalan Piao on Saipan indicates that the Northern Marianas were settled by an insular people originating in Southeast Asia. They made a distinctive red-slipped ware, sometimes incised with lime-filled decoration, closely related to Philippine wares. By AD 800 a plain, unslipped ware was in use, and stone architecture had developed. Parallel rows of upright pillars topped with hemispheric capstones (*halege*) were erected. According to early Spanish accounts, the pillars were supports for structures called *latte* (after which term the culture is named), which may have served as men's houses or as canoe sheds. Each village had from one to several *latte* structures. Stone and shell tools were used and the betel nut was chewed, as shown by extended burials most often located between the rows of *latte*.

Ferdinand Magellan was the first to open the Marianas to the West when he stopped there briefly during his first circumnavigation in 1521. There is some historical question as to which island he actually visited, but Magellan named the islands the Ladrones (Spanish: "Thieves") because while he was there some of the islanders took a small skiff that he had trailing behind one of his ships.

In 1565 Miguel López de Legazpi landed at Umatac, Guam, and proclaimed Spanish sovereignty over the Ladrones; some priests went ashore to say the mass. No colonies were started at that time, however, because the Spanish were more interested in conquests in the Americas, the Philippines, and the Moluccas. The British adventurer Thomas Cavendish was the next to visit the Marianas, in 1588 aboard the *Desire*. He traded briefly, and as he left he ordered his men to open fire from the rear of the ship to discourage the islanders from following.

Effects of early explorers

The effects of the early explorers and those who followed them to the islands were not unmixed. On the one hand the Europeans provided ironware and cloth, which was traded for fresh produce. On the other hand they introduced infectious diseases, including influenza, smallpox, leprosy, venereal diseases, and tuberculosis, which severely depleted the indigenous population.

The permanent colonization of the islands began with the arrival of Father Diego Luis de Sanvitores in 1668. With him were priests, laymen, women, and some Filipino soldiers. Mariana of Austria, the regent of Spain, financed his mission, and he renamed the islands the Marianas in her honour. Sanvitores and his colonists established churches and religious schools. A series of revolts attended these efforts, since the islanders resisted conversion to a religion that did not fit traditional beliefs. The Spanish during this time, in order to control the islanders more effectively, moved the population of the Marianas into enclaves and segregated the people into villages. Many islanders were killed in the process of relocation. Others died during their adjustment to new social environments, and the population was further decreased.

In 1680 the Spanish sent reinforcements led by Don José Quiroga, who was interim governor of the Marianas from 1680 to 1696. He subdued the islanders after a series of revolts, sieges, murders of missionaries, and burnings of churches that was known as the "Chamorro wars" and that resulted in many islanders fleeing to the hills. In reprisal, the entire native population was relocated from Saipan and Rota in the northern Marianas to the island of Guam. Finally, the Chamorro people took the oath of allegiance to the king of Spain, accepted Spanish customs, and began to wear Western-style clothes, cultivate corn, and learn to eat meat. Artisans were sent to the villages to teach sewing, spinning, weaving, tanning, iron forging, stone masonry, and other crafts. By 1698 the subjugation of the Marianas was complete.

The Spanish branched out from the Marianas into the rest of Micronesia, meeting only mild resistance. Guam became a regular stop for the Spanish galleons traveling between the Philippines and Mexico. By the 19th century the Marianas had become involved in European colonial rivalries. The Germans as well as the British began to encroach on the Spanish claims in Micronesia, and difficulties were averted in 1886 by the mediation of Pope Leo XIII, whose efforts in this regard prevented war between Germany and Spain. But Spain's empire was weakening, and by 1898 war with the United States was at hand. After American naval forces under the command of Commodore George Dewey defeated the Spanish fleet in the Philippines and took Guam, Spain made the decision in 1899 to withdraw entirely from the Pacific and sold its possessions—including all of the Marianas except Guam, which the Americans still held—to Germany.

German administration

The Germans promptly established an administrative centre at Saipan, which had served as the seat of the Spanish administration after the Americans captured Guam in 1898. This period marked the permanent division between Guam and the northern Marianas. Schools, a hospital, and other public buildings were erected in the northern Marianas, and colonists were encouraged to emigrate from Germany. Large coconut plantations were started, as were other agricultural projects. Copra production was the main German agricultural interest.

German control in the northern Marianas ended abruptly with the outbreak of World War I. In October 1914 the Japanese navy took possession of the northern Marianas and the rest of Micronesia. Japan's authority for this seizure was based upon several secret agreements with the British designed to keep the peace in Asia in the event of war. After World War I Japan received the northern Marianas by the terms of the Treaty of Versailles on June 28, 1919, and then later as a mandate under the League of Nations on Dec. 17, 1920. The United States, which continued to hold Guam, recognized this mandate on Feb. 11, 1922.

Japanese rule in the northern Marianas was direct and allowed the islanders little part in local government. The basic laws of Japan were extended to the islands with only such modifications as were needed to meet local conditions. Formal educational facilities were restricted, emphasis being placed upon learning the Japanese language. Public health conditions, however, were improved, and hospitals were established. Economic development was Japan's main interest, and large sugar plantations and refineries were started at Saipan and Rota. Large amounts

of labour and capital were made available. There was full employment, and the islanders were made more productive than they had been before.

When World War II began, the Japanese immediately took Guam from the United States and made their domination of the Marianas complete. The larger islands of Saipan, Tinian, Rota, Pagan, and Agrihan, along with Guam, became bases for Japanese expansion to the south and east. In 1943 U.S. forces began to reduce Japanese strength in Micronesia, and in June–July 1944 the Marianas were neutralized with the recapture of Saipan, Tinian, and Guam. The three islands then became Allied bases for the bombing of the Japanese home islands; the atomic bombs dropped on Hiroshima and Nagasaki were flown from a base at Tinian. Most of the individual islands of the Marianas were bypassed, and with most of these communication was not reestablished until the war was over.

U.S. invasion

The U.S. invasion of the Marianas during World War II completely destroyed the Japanese-created economy in the islands. The United States began the task of rebuilding. President Harry S. Truman signed a Trusteeship Agreement with the United Nations for the administration of the northern Marianas as a district within the Trust Territory of the Pacific Islands in Micronesia on July 18, 1947, and delegated responsibility for the civil administration to the U.S. Navy. Then, on July 1, 1951, following the signing of the peace treaty with Japan and the renouncing of all Japanese claims to the northern Marianas, U.S. administrative responsibility under the UN Trusteeship was transferred to the secretary of the interior. A special designation of "strategic" trusteeship was made, which placed the ultimate jurisdiction of the Marianas under the United Nations Security Council. The other Micronesian island groups of the Marshalls and the Carolines were also covered under this Trusteeship Agreement.

The United States pledged to "promote development . . . towards self-government or independence as may be appropriate." In 1969 political status talks began between the Micronesians from the entire Trust Territory and U.S. representatives. In 1973 the northern Marianas began separate negotiations with the United States that led to a plebiscite in 1975 in which the people approved a commonwealth status in association with the United States. The new government began operations under its own constitution and an elected governor, who was installed in January 1978. Eligible residents of the Northern Marianas became U.S. citizens. The United States declared the trusteeship officially ended on Nov. 4, 1986.

(D.A.B.)

Palau

The Republic of Palau (Palauan: Beluu era Belau) is an internally self-governing republic. It consists of about 340 coral and volcanic islands perched on the Kyushu-Palau Ridge in the western Pacific. The Palau archipelago lies in the southwest corner of Micronesia with the Philippines 550 miles (885 kilometres) to the west, New Guinea 400 miles to the south, and Guam 830 miles to the northeast. It reaches from latitude 2°55′ to 8°12′ N and from longitude 131°05′ to 134°44′ E. A huge barrier reef system, continuous on the west and broken on the east, encircles most of the archipelago's 188 square miles (488 square kilometres) of land: its major populated islands are Babelthuap (Babeldaob), Koror, Malakal, Arakabesan, and Peleliu. The sparsely populated Kayangel Islands to the north of Babelthuap and the raised coral islands of Angaur, Sonsorol, Pulo Anna, and Tobi south of Peleliu lie outside the barrier reef system. Koror Island, rising to 2,061 feet (628 metres) just south of Babelthuap, is the provisional capital and largest population centre.

Palau was a member of the UN Trust Territory of the Pacific Islands established in 1947 and administered by the United States. The U.S. government dissolved the trusteeship in 1986, but repeated measures to win the required support for a compact of free association between Palau and the United States were unsuccessful until 1993. The Republic of Palau officially became a sovereign state on Oct. 1, 1994.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief and drainage.* All but six of Palau's islands lie within an expansive lagoon, enclosed by the barrier reef, that stretches northeast to southwest for almost 70 miles. Babelthuap, the largest island (153 square miles), is volcanic, mainly andesite, and is bounded by thick mangrove forests broken occasionally by sandy beaches on the east coast. Its highest point, Rois Ngerekelehuus, in the northwest, is 794 feet high. Babelthuap is essentially a rolling upland, part grassland and part jungle, that has been incised by stream action to form a well-developed drainage system of three rivers. With 150 inches (3,800 millimetres) of rain annually, considerable erosion has taken place on Babelthuap in spite of the stability provided by laterite soils, clays, and vegetation. The Palauan practice of burning the grassy upland areas during the dry season has contributed to erosion.

Mangrove forests

Since 1977 Babelthuap has been connected by a concrete bridge to the 3.6-square-mile Koror, which, in turn, is linked by causeway to Malakal island, the site of Palau's deepwater port, and Arakabesan island. The combined area of the three smaller linked islands is seven square miles. All are of volcanic origin. Beginning adjacent to southern Babelthuap and eastern Koror, however, and filling the huge lagoon for 28 miles south to Peleliu are 320 green velvet "rock" islands. These are uplifted reef structures of coralline limestone, each deeply undercut at sea level. Some of the islands are large, towering 600 feet, and have interior brackish lakes, containing unique organisms, that are connected to the lagoon by subterranean channels. Plant growth is thick on the rock islands and, together with the chemical action of heavy rains, has sculpted and broken the surfaces, producing razor-sharp edges and points and broken rubble. The limestone islands have rich deposits of phosphate, and the more accessible ones have been mined.

The inhabited coral islands outside Palau's reef–lagoon–island system sit on volcanic substructures and consist of the Kayangel Islands, 25 miles north of Babelthuap, and Angaur, six miles south of Peleliu, which was heavily mined for its phosphate first by the Germans and later by the Japanese. Sonsorol, Pulo Anna, and Tobi, all with areas less than one square mile, are 180 miles southwest of the Palau archipelago. All are flat platform structures with fringing reefs.

Climate. Palau's climate is tropical. Rainfall varies from 120 to 160 inches per year. Humidity is fairly constant, ranging from 77 to 84 percent, and temperatures vary not more than 10° F (5.5° C) diurnally, monthly, or annually from a mean of 81° F (27° C). Northeast trade winds prevail from December to March, and southwest monsoons from June to October. Prevailing oceanic currents offshore are the North Equatorial Current and the Equatorial Countercurrent.

Geologically, Palau sits on the Philippine Sea Plate only 30 miles west of the 26,200-foot-deep Palau Trench, the western boundary of the upthrusting Pacific Plate. Despite its close proximity to this subduction zone, Palau rarely experiences earthquake activity.

Palau's marine environment exhibits a rich fauna balanced by an abundant terrestrial flora. This richness derives from Palau's close proximity to Indonesia, New Guinea, and the Philippines. Palau has more species of marine life than any other area of similar size in the world; corals, fish, snails, clams, sea cucumbers, starfish, sea urchins, sea anemones, jellyfish, squid, and feather worms exist in profusion and variety. Common flora are the beach morning glory, Polynesian ironwood tree, pandanus, and various species of palm and fern. The birds of Palau are abundant and colourful, and many migrate to or through Palau twice annually. A few species of reptile and amphibian live in Palau, including a unique frog that gives birth to live young. Insects are also abundant. The accidentally introduced coconut rhinoceros beetle did enormous damage to coconut palms until it was brought under control by the parasitic scolia wasp.

Unique frog species

The people. Palau's people are generally of medium height with diverse skin tones and facial types. The islands were inhabited from 2,000 to 3,000 years ago by



Emperor angelfish (*Pomacanthus imperator*) swimming among the corals in the barrier reef channel of Aulong Island, Palau.
© Douglas Faulkner/Sally Faulkner Collection

successive waves of Malays from Indonesia, Melanesians from New Guinea, Philippine natives, and some Polynesians from outlying Polynesian islands in Micronesia, who formed the basic genetic stock.

During the past two centuries Europeans, Japanese, and Americans have made genetic contributions. The southwest islanders, who are culturally and linguistically distinct from the Palauans, are the only minority group; they trace their origin to a group of ancestral survivors of one or more canoes that drifted to Sonsorol from Ulithi Atoll, northeast of Yap. Palauan is a Western Austronesian language and is very complex in that it has many irregularities that make formulation of grammatical and lexical rules difficult.

The indigenous Palauan religion of powerful ancestral and nature spirits was supplanted by Christianity. The only remnant of past belief systems is the Modekngei religion, which developed as a syncretic movement in response to colonial pressure. In addition to the well-established Roman Catholic, Lutheran, and Seventh-day Adventist churches, missionaries of the Jehovah's Witness, Mormon, Bahāʾī, and Assemblies of God faiths have carried their message to Palau.

Palauans are atypical Micronesians in their desire to migrate overseas. There are a number of substantial Palauan communities on Guam, in Hawaii, and on the U.S. West Coast. This continued migration and the apparent decline in birth rates have resulted in a small negative population growth rate.

The economy. Since the end of World War II, the major employer in Palau has been government—first the U.S. Navy, then the Trust Territory of the Pacific Islands, and finally the government of Palau. Nevertheless, in the rural areas outside Koror the subsistence economy is active. Women typically gather and cultivate wet taro, sweet potato, and cassava, and men fish and tend pigs, which are used at customary feasts. Tourism has grown considerably during the late 20th century and has also made some contribution to the republic's economic growth.

With the majority of Palau's population in Koror, that island has a good system of paved roads. There are stretches of paved road on Babelthuap, and the roads built in 1944–46 by U.S. military forces on Peleliu and Angaur are still usable. Transportation between islands is by boat or airplane. There is regular commuter service from Koror to Peleliu and Angaur, and trips by speedboat to coastal villages on Babelthuap usually can be completed in a few hours. There is an international airport located on Babelthuap.

Nearshore reef fishing is carried out on a subsistence and small-venture basis, but it does not generate significant government revenue. Offshore tuna fishing by foreign vessels provides a small amount of government revenue through the sale of licenses.

Administration and social conditions. The constitution of the Republic of Palau established a presidential form of government, which was installed in 1981. The executive consists of separately elected offices of president and vice president, a Council of Chiefs to advise the president on traditional laws and customs, and a Cabinet. The National Congress consists of a Senate and a House of Delegates. Both executive and legislative branches are elected for four-year terms. The Palau judiciary consists of a Supreme Court and a National Court of Common Pleas with both trial and appellate divisions. At the local level, each of the 16 traditional settlement areas comprises a state with an elected governor and legislature. Palau has no armed forces.

Each of the rural states has an elementary school. Three church-run high schools and a public high school take most of those who complete the eighth grade. The Micronesian Occupational College on Koror was established in 1969 and was reaccredited in 1987.

Health care is provided by the hospital on Koror and dispensaries in each rural state. The incidences of mental illness, suicide, and alcoholism in Palau are higher than in most countries.

Cultural life. Pragmatic adaptation, persistence of wealth-exchange customs, and competition characterize Palauan society. Palauans adapted to a century of colonial intrusion—Spanish, German, Japanese, and American—by viewing reality as something imposed from the outside to which one must adjust quickly if it is to be manipulated. Reciprocity and redistribution customs carried out between clans exchange food and services for money and gifts at births, house completions, and funerals. Women are the strength of society and control land, money, and titles. Men, previously occupied as fishermen and warriors, continue their traditional tasks in the rural areas and, as an adaptation to modern society, compete for elected office, in business, and in baseball. Traditional art forms persist in chants and storyboards, which are now made for sale to tourists rather than for decoration of men's clubhouses. The Palau National Museum in Koror has a small but instructive collection of artifacts.

Adaptation

HISTORY

Large hillside terraces, numerous stone ruins, and megaliths on Babelthuap give evidence of a vital culture before contact with European explorers. The first extensive contact of Palauans with Westerners took place after the shipwreck of the East India Company's packet *Antelope* in 1783. George Keate's *An Account of the Pelew Islands* (1788), which recounted the friendship and high adventure found in Palau, served to fuel the European myths of the noble savage and island paradise. The first 70 years of the 19th century were punctuated by the occasional visits of whalers and traders, who left beachcombers and firearms behind. Diseases communicated by contact with Europeans led to the deaths of many islanders, and firearms were prized for intervillage warfare, which was ended in 1883 through the peaceful intervention of Captain Cyprian Bridge of HMS *Espiègle*. Spanish and German colonial influence was expressed through Roman Catholic missionaries. The Japanese navy expelled the Germans at the beginning of World War I, and, although the Japanese period is locally remembered as one of economic development and order, the Palauans were a marginal minority by 1936. Japan lost Palau in World War II in a struggle that was socially destabilizing and confusing to the Palauans.

After a short period of administration by the U.S. Navy, Palau became part of the United Nations Trust Territory of the Pacific Islands under U.S. administration in 1947. A constitution was adopted in 1981 (following two prior referendums), and elections were held in the same year. The country became internally self-governing in 1981. Palau signed a Compact of Free Association with the United States in 1982, but the required number of voters failed to pass the referendum until 1993. The compact required that the United States remain responsible for

external security and defense and that it provide financial assistance for Palau, but conflict arose over Palau's constitutional prohibition on the operation of U.S. nuclear-powered or nuclear-armed vessels and aircraft within the jurisdiction of Palau. According to the terms of the compact, the United States reserved this right as well as the right to neither confirm nor deny the presence or absence of such weapons in Palau. Several attempts were made to revise the constitution, revise the compact agreement, and secure Palauan approval, and the United States dissolved the trusteeship in 1986. In 1992, voters approved an amendment that reduced from three-fourths to a simple majority the popular vote required to override the antinuclear provision of the constitution. This cleared the way for approval of the compact in 1993.

In 1985 volatile internal politics resulted in the assassination of the first president, Haruo I. Remeliik. In August 1985 Lazarus E. Salii was elected to serve out the four-year term begun by Remeliik in January 1985, but Salii himself met a violent death in August 1988. (D.R.Sh.)

Tuvalu

Tuvalu, an independent nation composed of nine small coral islands, lies south of the Equator in the central Pacific Ocean. The islands are scattered in a chain lying approximately northwest to southeast over a distance of some 420 miles (676 kilometres). The islands are small; Tuvalu has a land area less than 10 square miles (26 square kilometres). The capital is Fongafale, on Funafuti Atoll. Tuvalu was formerly called Ellice Islands and with colonial Kiribati formed the unit known as Gilbert and Ellice Islands Colony.

PHYSICAL AND HUMAN GEOGRAPHY

The land. The group includes both atolls and reef islands. The atolls—Nanumea, Nui, Nukufetau, Funafuti, and Nukulaelae—have islets encircling a shallow lagoon; the reef islands—Nanumanga, Niutao, Vaitupu, and Niulakita—are compact with a fringing reef. The islands are low-lying, most being 13 to 16 feet (4 to 5 metres) above sea level. There are no rivers; rain catchment and wells provide the only fresh water. Rainfall averages 100 inches (2,500 millimetres) in the north and 125 inches in the south. The prevailing winds are southeast trades; westerly storms occur from November to February. Daytime temperatures range from 80° to 85° F (27° to 29° C).

Porous soils

Because the soils are porous, agriculture is limited. Coconut palms thrive, and breadfruit trees, pandanus, taro, and bananas are grown. Pigs and chickens are raised, and seabirds, fish, and shellfish are caught for food. The islands increasingly depend on imported food.

The people. The Tuvaluans are Polynesian, and their language, Tuvaluan, is closely related to Samoan. Nui, however, was heavily settled in prehistoric times by Micronesians from the Gilbert Islands (now Kiribati). English is taught in the schools and widely used. The vast majority of the population belongs to the Church of Tuvalu (the former Ellice Islands Protestant Church).

Although most people live on the outer islands in extended family households clustered into villages, about one-third of the population lives on Funafuti, the centre of government and commerce. Population growth has been slowed by family planning; life expectancy at birth is about 60 years. About 10 percent of the population lives overseas, either pursuing education, working in the Nauru phosphate industry, or working on merchant ships.

The economy. Most Tuvaluans are subsistence farmers and are aided by remittances from relatives working overseas. A small quantity of copra is produced for export, the sale of stamps accounts for modest earnings, and fees are collected from foreign fishing fleets, but the country depends heavily on foreign aid. It imports most of its food, fuel, and manufactured goods. Retailing is handled by community-based cooperative societies. Tuvalu uses Australian currency but also issues its own coinage. There is a single bank, a joint government-commercial venture.

Tuvalu has air links with Kiribati and Fiji; for international shipping, it depends on irregular regional services. Seaplanes have been used for interisland travel, but generally the outer islands depend on a single government vessel. Motorcycles are common on Funafuti, but there are few automobiles.

Administration and social conditions. Tuvalu is a constitutional monarchy within the Commonwealth, with the British sovereign (through a governor-general) as head of state. The government is a parliamentary democracy with a unicameral legislature elected by universal adult suffrage. There are no political parties: the prime minister is chosen by and from the legislature. Tuvalu is a member of the South Pacific Forum.

The government provides universal primary education and, under a joint arrangement with the Church of Tuvalu, secondary education to school certificate level for selected pupils. A few are sent overseas for further education and training. Medical facilities are centralized on Funafuti, but all other islands have clinics with trained medical staff.

Mixture of traditional and Western ways

Cultural life. The Tuvaluan lifestyle has been Westernized to an extent, but Western-style amenities are few. Only Funafuti has a regular electricity supply; the government publishes a brief news sheet, but there is no newspaper; a few motion pictures are shown, but there is no television; and there is only a single radio station. Most Tuvaluans live in villages of a few hundred people, tend their gardens, and fish from handcrafted canoes. Traditional music and dancing still enjoy a strong following, along with Western forms. Volleyball, soccer, and cricket are popular. Tuvaluan life, despite modernization, still rests on a firm traditional base that emphasizes the importance of community consensus and identity.

For statistical data on the land and people of Tuvalu, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

The first settlers were from Samoa and probably arrived in the 14th century AD. Smaller numbers subsequently arrived from Tonga, the northern Cook Islands, Rotuma, and the Gilbert Islands. Niulakita, the smallest and southernmost island, was uninhabited before European contact; the other islands were settled by the 18th century, giving rise to the name Tuvalu, or "Cluster of Eight."

Europeans first discovered the islands in the 16th century through the voyages of Álvaro de Mendaña de Neira, but it was only from the 1820s, with visits by whalers and traders, that they were reliably placed on European charts. In 1863 labour recruiters from Peru kidnapped some 400 people, mostly from Nukulaelae and Funafuti, reducing the population of the group to less than 2,500. A few were later recruited for plantations in Queensland, Australia, as well as in Fiji, Samoa, and Hawaii. Concern over labour recruiting and a desire for protection helps to explain the enthusiastic response to Samoan pastors of the London Missionary Society who arrived in the 1860s. By 1900, Protestant Christianity was firmly established.

With imperial expansion the group, then known as the Ellice Islands, became a British protectorate in 1892 and part of the Gilbert and Ellice Islands Colony in 1916. There was a gradual expansion of government services, but most administration was through island governments supervised by a single district officer based in Funafuti. Ellice Islanders sought education and employment at the colonial capital in the Gilbert group or in the phosphate industry at Banaba or Nauru. During World War II, U.S. forces were based on Nanumea, Nukufetau, and Funafuti, but hostilities did not reach the group.

From the 1960s, racial tension and rivalries over employment emerged between Gilbertese and Ellice Islanders. Ellice Islanders' demands for secession resulted in a referendum in 1974, transition to separate colonial status between October 1975 and January 1976, and independence as Tuvalu in 1978. After independence the main priorities were to establish the infrastructure for a separate, if small, nation, and to seek foreign assistance to match political independence with economic viability. (B.K.M.)

For later developments in the history of Tuvalu, see the BRITANNICA BOOK OF THE YEAR.

MAJOR POLITICAL UNITS OF POLYNESIA

American Samoa

American Samoa, a territory of the United States, consists of the eastern part of the Samoan archipelago. It is situated in the central Pacific Ocean about 1,600 miles (2,600 kilometres) northeast of New Zealand and 2,200 miles southwest of the Hawaiian Islands. The territory includes the six Samoan islands east of the 171° W meridian and is part of Polynesia. Samoa (formerly Western Samoa), its closest neighbour and a self-governing nation, consists of the nine Samoan islands west of that meridian.

American Samoa, with a total land area of 77 square miles (199 square kilometres), includes the inhabited islands of Tutuila, Tau, Olosega, Ofu, and Aunuu and an uninhabited coral atoll named Rose Island. Swains Island (inhabited), a coral atoll 280 miles northwest of Tutuila and geographically not a part of the archipelago, was made a part of American Samoa in 1925. The capital of American Samoa is Pago Pago, on Tutuila.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Except for the coral atolls, the islands of American Samoa were formed within the past 7,000,000 years by volcanic activity; their interiors are high and rugged. The main island of Tutuila, with an area of 53 square miles, rises steeply above deep inlets. The most notable of these inlets is Pago Pago Harbor, which almost divides the island in two. Tutuila's highest peak is Mount Matafao (2,141 feet [653 metres]). The Manua island group (Tau, Olosega, and Ofu islands), situated about 60 miles east of Tutuila, constitutes the second largest island area. Coral reefs are common to the extremities of the islands, particularly Tutuila; some of the reefs form barriers that enclose lagoons.

Coral reefs

Climate. The climate is tropical. Pago Pago receives about 200 inches (5,000 millimetres) of precipitation annually. Most streams do not reach the ocean but filter into the porous basalt rocks. Hence, coastal wells provide much of the water supply. Temperatures are unusually constant; average temperatures range from 68° to 90° F (21° to 32° C). Average humidity is 80 percent. The moderate southeast trade winds prevail, but severe storms can occur during the wet season, from November to March.

Plant and animal life. Rain forests with tall ferns and trees cover the mountainous interiors of the islands. Plantations of taro, coconut, and other food crops are located on the coasts. Although the islands are not rich in animal life, some of their bird species—such as the rare toothbilled pigeon—are unique. Wildlife includes the flying fox, lizards, rats, harmless snakes, and pigs. The islands also have a rich insect life.

Settlement patterns. Most people live in coastal villages. Pago Pago, the only town, is the main port and administrative and commercial centre. Since the mid-20th century many American Samoans have migrated to the United States, and there are now more American Samoans abroad than on the islands.

The people. The Samoans are a Polynesian people closely related to the native peoples of New Zealand, Tahiti, Hawaii, and Tonga. The Samoan way of life, or *fa'a Samoa,* is traditionally communal. The basic unit of social organization is the extended family (*aiga*). The extended families are arranged in villages, which are grouped into districts. The extended families are headed by chiefs (*matai*), who are selected by their extended families on the basis of consensus. Most chiefs' titles are very old. The village chiefs together make up a village council (*fono*), which controls and runs village affairs. Despite the increasingly pervasive influence of American culture, most Samoans still live according to *fa'a Samoa,* and nearly all of them are fluent in the Samoan language. Most of the American Samoans, nonetheless, also speak English. The Congregational Christian Church has the largest following among religious institutions; most of the remaining population is either Roman Catholic, Mormon, or Methodist.

Fa'a Samoa

The economy. *Industry and agriculture.* The U.S. administration is the main employer. Tuna canning (by American-owned canneries) and tourism are major industries. Agriculture is organized on a semicommercial basis for the production of taro, bananas, tropical fruits, and vegetables. Traditional family gardens produce coconuts, breadfruit, and yams. Production nearly meets domestic needs, and the U.S. government has implemented programs to increase production to self-sufficiency levels.

Social legislation ties American Samoa more closely to U.S. costs of living than its South Pacific neighbours.

Transportation. A major public works program on American Samoa has increased the number of miles of paved roads, mainly on the island of Tutuila. Pago Pago is the only port of note. An international airport is located on Tutuila, and smaller airstrips operate from the islands of Tau and Ofu.

Administration and social conditions. *Government.* American Samoa is an unincorporated, unorganized territory, the people of which are U.S. nationals but not citizens. The territory's chief executive, according to the constitution of 1967, is the governor. Until 1977 the governor was appointed by the U.S. Department of the Interior; in response to a referendum of 1976, however, the offices of governor and lieutenant governor are filled by popular elections every four years. The minimum voting age is 18. The Fono (bicameral legislature) is autonomous in its disposition of local revenues and is the sole lawmaking body, subject to the governor's approval. The 21 members of the House of Representatives are elected by universal suffrage; one is a nonvoting delegate elected from Swains Island. The 18 senators are chosen by councils of chiefs, in accordance with Samoan custom. In 1981 the first official nonvoting delegate from American Samoa to the Congress of the United States was elected. Apart from Swains Island, the islands are divided into three administrative districts (each with an appointed district governor), which are subdivided into 14 counties. Chiefs representing each family form village and district councils. Each village has a village court with authority to adjudicate on minor misdemeanours.

Village courts

Education. Education is compulsory between the ages of six and 18 in American Samoa. Televised instruction, given mainly in English and mostly by U.S. teachers and technicians, is available to local schools. The American Samoan Community College offers vocational training and nursing programs, with university education available in Hawaii or on the U.S. mainland. (P.R.C./A.We.)

HISTORY

The Samoan islands were settled by Polynesians (probably from Tonga) about 1000 BC. By about AD 200 Samoa had become the centre of much of the settlement of eastern Polynesia.

The Dutch navigator Jacob Roggeveen sighted Samoa in 1722, and other European explorers, beachcombers, and traders followed. The London Missionary Society sent its first missionaries to the islands in the 1830s. More missionaries traveled to the islands as missionary influence spread to Tutuila and later the Manua Islands. In 1878 the United States signed a treaty for the establishment of a naval station in Pago Pago Harbor. Great Britain and Germany signed similar agreements the next year. The three powers increasingly competed for influence in the archipelago until 1889, when a typhoon sank six of their warships. They subsequently agreed to declare the islands neutral; however, in 1899 eastern Samoa was annexed by the United States, and Germany followed suit by annexing the western islands—Western Samoa (now Samoa).

By 1904 the eastern islands had all been ceded to the United States, but the U.S. Congress did not formally accept the deeds of cession until Feb. 20, 1929. Under the administration of the U.S. Navy (1900–51) American Samoa became a strategic naval base, but the Samoan leaders had little administrative power. In 1951 control of the

territory was transferred to the Department of the Interior. The U.S. government appointed a governor who had full powers to administer the territory. He appointed political advisers and senior civil servants from the United States to help him. The Samoans agitated for control of their country's affairs, and in 1977 Peter Coleman, a Samoan, became the territory's first elected governor. Since then, all members of the territory's House of Representatives have been elected by the citizens. Except for defense, the governor and the Fono run American Samoa's affairs. American Samoans are U.S. nationals and can move freely between their country and the United States. (A.We.)

Cook Islands

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief and drainage.* The Cook Islands is an internally self-governing island state in free association with New Zealand in the south Pacific Islands. The 15 tiny islands have a total land area of 91 square miles (236 square kilometres) but are spread over 770,000 square miles of sea. The capital is Avarua on Rarotonga. Each island is the top of one or more volcanoes, but only on the largest islands do the plugs and craters of now-extinct volcanoes still dominate the skyline; these rise to 2,139 feet (652 metres) on Rarotonga, an island only four miles (six kilometres) wide. Most other islands of the southern group (such as Atiu, Mangaia, Mitiaro, and Aitutaki) show various combinations of atoll and high-island formation. The northern group islands (including Penrhyn, Manihiki, Rakahanga, and Nassau) are atolls, narrow and low-lying sandbanks resting on circular reefs around lagoons rich in marine life.

The land areas being so small, there are no rivers, and only the largest islands have even small streams. There are small freshwater lakes on the high islands of Mangaia, Atiu, and Mitiaro; saltwater lagoons inside all the atolls; and fringing lagoons between most islands and their outer reefs. The rain that falls on the atolls permeates the island coral and is naturally stored in a lens-shaped layer above the heavier salt water. The islanders must rely on wells and rainwater storage tanks to conserve their limited sources of water.

Soils. Soils on the low-lying atolls are very limited in depth and quality. Most of the high island of Rarotonga is ruggedly mountainous, with narrow valleys having small but fertile pockets of soil. The coast consists of makatea, or upraised coral reef, of limited fertility. Between the mountains and the coast, however, is a ring of fertile volcanic soil. On the other high islands much of the area is likewise taken up by eroded central slopes encircled by makatea, but again there are areas of fertile soil between. The problem of erosion has been greatly accentuated by the planting of pineapples and other export crops on soils that, in the long term, are too fragile for plantation farming.

Ring of fertile volcanic soil

Climate. All the islands lie within the tropics, but the southernmost just barely. Because the Cook Islands are small, mid-ocean islands swept by the southeast trade winds, temperatures are generally moderate. The mean annual temperature on the southern island of Rarotonga is 75° F (24° C), but on the northernmost island of Penrhyn it is 82° F (28° C).

Seasons are not clearly marked. The English terms summer, winter, spring, and autumn are used, but Cook Islanders also recognize the traditional local patterns of prevailing winds, rainfall, and temperature. Rainfall, though erratic over the years, tends to be uniform across the various islands. It averages 80 inches (2,000 millimetres) on Rarotonga, though with considerable difference between the windward and lee sides of the central mountains, and is slightly lower on Aitutaki and slightly higher on Penrhyn.

The spectacular climatic hazard is hurricanes, which strike with destructive force between December and March about once or twice every 10 years. Less spectacular, but at times equally destructive of agriculture, are droughts, to which the northern group is more vulnerable than the southern.

Plant and animal life. Only a limited range of plant life thrives in the north, the coconut and the pandanus being predominant. On the fertile areas of the southern islands, a wide range of tropical fruits and vegetables flourishes. Indigenous species include taro, yams, bananas, breadfruit, and sweet potatoes. Introduced species—in many cases grown for export—include citrus fruits, tomatoes, pineapples, papayas, beans, and zucchini.

The original Polynesian settlers brought with them pigs, dogs, chickens, and a type of small rat. These are still the main fauna, though a few goats, horses, and other animals have been introduced. Some native birds became extinct in the 19th century after Europeans introduced cats. The *kakirori,* or Rarotongan flycatcher, an attractive tiny bird unique to Rarotonga, had been reduced by the early 1990s to about 30 breeding pairs, which were protected by the government conservation service.

Settlement patterns. Most Cook Islanders live in villages, the largest of which is Avarua on Rarotonga, though some people (in Rarotonga particularly) live on their farming lands. The former indigenous houses of thatch and timber have been almost totally replaced by homes of cement and timber with iron roofs.

The people. With the exception of the inhabitants of isolated Pukapuka, who are of predominantly Samoan and Tongan descent, almost all Cook Islanders have mixed Polynesian ancestry. Intermarriage with European, Chinese, and African settlers was common in the early 19th century. The Christian denominations, particularly the Cook Islands Christian Church, account for nearly all religious affiliation.

There are two main indigenous Polynesian languages, one for the island of Pukapuka and the other (with dialectal variations) for all other islands. English is spoken by most Cook Islanders, and the literacy rate is more than 90 percent.

High literacy rate

The population is static at about 17,000 and has been so for many years. Natural increase is high but is offset by emigration to New Zealand and Australia. Today more than twice as many persons of Cook Islands ancestry live in New Zealand than in the Cook Islands. The main nonindigenous population is of European origin by way of New Zealand. There is considerable internal migration from the smaller islands to Rarotonga, the most populous island, where there is generally full employment.

The economy. The largest industry by far is tourism, mainly from New Zealand, Australia, Canada, the United States, and Europe. The second largest industry is international finance, as the Cook Islands are a major tax haven. Government plays a significant part in the economy and is the largest employer, mainly for the provision of services; production and commerce is almost all private. Taxes are moderate, foreign investment is encouraged, and foreign aid—largely from New Zealand—makes a significant contribution to the economy. The banking system comprises two Australian retail banks, a New Zealand merchant bank, a government development bank, and a number of offshore banks dealing with international finance.

No mineral resources are exploited, though phosphate exists on the floor of the Manihiki lagoon, and there are vast deposits of manganese and cobalt and some other metals on the seabed near Manihiki. Agricultural production at present consists mainly of small farming, either for home consumption or for airfreight to New Zealand in its off-season. Commercial fishing for export is done mainly by Taiwanese vessels operating out of American Samoa, but there is widespread fishing for home consumption. A new cultured-pearl industry promises to generate more income than all agriculture, fishing, and industry combined.

The small industrial sector entails clothing and shoe manufacturing and food processing, mainly for export to New Zealand. The major imports are from New Zealand, Japan, and the United States. Energy is mainly produced from imported fuels, though there is some use of solar and wind power.

Each island has a network of roads; a paved road encircles Rarotonga. Regular service by small aircraft connects all the main islands, but shipping is erratic. Externally there are direct air services to New Zealand, Australia, Hawaii, Tahiti, Fiji, and Western Samoa from the international

airport on Rarotonga. External shipping is mainly with New Zealand and the United States.

Administration and social conditions. The Cook Islands is a self-governing state. Its government is democratic, with general elections, by universal suffrage, held every five years. The two main parties are the Cook Islands Party and the Democratic Party. The constitution, adopted in 1965, has been amended several times. Local government, including the Parliament of the Cook Islands, is mainly elected, but a council of hereditary leaders, the House of Ariki (High Chiefs), advises the government on traditional matters of land ownership, custom, and the like.

Council of hereditary leaders

Education is universal and compulsory from the ages of 5 to 16, with some church schools operating in addition to those run by the government. Higher education is available at such schools as the national Teachers College and Nurses College, through apprenticeship schemes, and through an extension centre of the University of the South Pacific at Avarua; many overseas scholarships are provided by the government and by overseas donors.

Free medical services are available from a government-owned hospital or dispensary on each island. Dental treatment is also free for schoolchildren. Filariasis is common, and the diseases more commonly found in developed countries—diabetes, cardiovascular disorders, cirrhosis of the liver—are becoming more widespread. Automobile accidents are a major problem.

Cultural life. Culturally the islands are relatively homogeneous, with the culture of Pukapuka—which was settled from western, rather than eastern, Polynesia—showing most variation from the general pattern. The government plays an active role in cultural life, particularly in the sponsorship of song and dance festivals for which the islands are renowned. A small library and museum in Avarua provide additional cultural attractions. Owing to tourism and intensive interaction with neighbouring industrialized nations, much international (generally Western) culture has been incorporated into the native lifeways. Nevertheless, traditional ceremonies, such as that celebrating the first haircut of the favourite son in a family, are as vibrant as ever. A distinctive style of food preparation draws on the traditions of Europe, China, Fiji, and Tahiti. Christian tradition, some of it from the English Victorian era, is strongly manifest, with modern American-derived evangelistic services and rituals becoming more common. The major national holiday is Constitution Day, which usually gives rise to a 10-day celebration. A Tiare ("Gardenia") Festival, a float parade, and a series of song and dance competitions fill the annual calendar of festivities.

There is a national (government-owned) radio station serving all islands, a government-owned television service to Rarotonga and Aitutaki, and a private FM radio station. The only daily newspaper, formerly government-owned, was privatized in 1989.

HISTORY

Polynesians, mainly from the area now known as French Polynesia, were the only inhabitants of the Cook Islands until the 19th century. With only minor exceptions, each island was autonomous, and within each of the larger islands there were several competing tribal communities. Spanish explorers visited several islands in the northern group in the late 1500s and early 1600s but did not stay. Captain James Cook was the first European to call at most of the southern group, in 1773, 1774, and 1777.

English and Tahitian missionaries of the London Missionary Society began arriving in 1821 and were the first foreigners to settle. They established a theological college on Rarotonga and exerted a strong influence on the form of government that evolved in each of the islands over the next half century.

Fear of a French takeover, such as that which had occurred in nearby Tahiti and other islands of the Society Group, prompted some chiefs to petition the United Kingdom to declare a protectorate over the Cook Islands. The British government eventually complied in 1888, and a single federal parliament was established. This was the first time that these scattered islands had come under a united government.

Petition for protectorate status

New Zealand was keen to annex the Cook Islands, but the United Kingdom would not agree to this except on certain conditions, one of which was that the request for annexation must come from the Cook Islands. With some New Zealand persuasion, chiefs of the largest islands petitioned for annexation, which was undertaken in 1901. After 1912 the federal parliament was allowed to lapse, and no direct representation at the national level occurred until 1946, when a Legislative Council was organized. In 1957 its powers were extended and its name changed to Legislative Assembly.

In 1965 the Cook Islands moved to self-government but retained an associated-state relationship with New Zealand. Under this arrangement, the Cook Islands is a sovereign state, but New Zealand is responsible for defense, after consultation with the prime minister of the Cook Islands. (R.G.C./M.T.C.)

For later developments in the history of Cook Islands, see the BRITANNICA BOOK OF THE YEAR.

Easter Island

Easter Island (Isla de Pascua), a Chilean dependency in the eastern Pacific Ocean famous for its giant stone statues, represents the easternmost outpost of the Polynesian island world. It stands in isolation 1,200 miles (1,900 kilometres) east of lonely Pitcairn Island and 2,200 miles west of Chile. Forming a triangle 14 miles long by seven miles wide, the island has an area of 63 square miles (163 square kilometres); its highest point, Mount Terevaka, is 1,969 feet (600 metres) above sea level. To its original inhabitants the island is known as Rapa Nui ("Great Rapa") or Te Pito te Henua ("Navel of the World"); its first European visitors, the Dutch, named it Paaseiland ("Easter Island") in memory of their own day of arrival. Its mixed population is predominantly of Polynesian descent; almost all live in the village of Hanga Roa on the sheltered west coast. The island is a province that is administratively part of the region of Valparaíso, Chile.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* The small and hilly island is not part of a sunken landmass but is a typical oceanic high island formed by volcanoes rising from the seafloor. Geologic and oceanographic evidence shows that no perceptible emergence or submergence of the island's coastline has taken place since the last fall in sea level, which occurred less than 10,000 years ago. Three extinct volcanoes chiefly composed of tuff (a rock formed of compacted volcanic fragments) and joined by their own lava flows give the island its characteristic triangular shape. Parasitic tuff craters and cones (*i.e.*, craters and cones formed on the side of, or near, volcanoes after the original vent has become plugged up) are interspersed in the landscape, which is otherwise dominated by eroded lava fields. Most of these fields are thickly packed with both large and small lumps of cellular and tuffaceous lava that is either black or rusty in colour. Stoneless surface soil is sparse; it is suitable for extensive cultivation mainly in the Hanga Roa and Mataveri area in the southwest, at Vaihu and on the plain southwest of the volcano Rano Raraku, and on the prehistorically cleared Poike peninsula in the eastern corner of the island. Rain collects in the partly bog-covered crater lakes of the volcanoes Rano Kao, Rano Raraku, and Rano Aroi. One intermittent stream, fed by the Rano Aroi crater lake, flows down Mount Terevaka's slopes before disappearing into the porous soil. Water from the extremely deep crater of Rano Kao, which is about 3,000 feet wide, is piped to Hanga Roa. The coast is formed by soft, eroded, ashy cliffs, with a vertical drop of about 500 to 1,000 feet; the cliffs are intercepted by long stretches of low, hard, and rugged lava formations. There is no natural harbour, but anchorage is found off Hanga Roa on the west coast; off Vinapu and Hotu-Iti on the south coast; and off Anakena and in the Bahía la Perouse on the north coast. Notable among the few small offshore islets are Motu-Nui, Motu-Iti, and Motu-Kaokao (which figured in a local bird cult) near the southwest cape. The only true sand beach is at Anakena; most other beaches are of gravel. Caves abound,

Formed by three extinct volcanoes

many consisting of subterranean rooms joined by narrow tunnels extending far into the lava beds.

Climate. The climate is subtropical. The warmest months are January through March, when the average temperature is 73° F (23° C), and the coolest months are June through August, when the average temperature is 64° F (18° C). Average annual precipitation is about 49 inches (1,250 millimetres) but with considerable annual variation. September is the driest month, and the heaviest rainfall occurs in June and July. Winds in June and August are irregular; during the rest of the year trade winds from the east and southeast are dominant. From September through March the Peru (or Humboldt) Current, which has an average temperature of about 70° F (21° C), flows against the island.

Plant and animal life. Indigenous plants and animals are few. At the time of European arrival the toromiro tree, endemic to the island, was the only wild tree and the Carolina wolfberry the only wild shrub, the vegetation being predominantly herbaceous. The toromiro tree was overexploited by the island wood carvers, and the last local specimen died in the 1950s. The species was saved from extinction, however; the Norwegian Archaeological Expedition collected seeds and planted them in the Gothenburg Botanical Garden, and saplings from the garden were reintroduced to Easter Island in 1988. Analysis of pollen deposits has revealed that other trees and shrubs, among them the giant Chile palm, were formerly present on the island until exterminated by extensive fires occurring at the time of aboriginal human settlement. Today only 31 wild flowering plants, 14 ferns, and 14 mosses are reported. Grass and small ferns dominate the barren landscape, whereas the boggy crater lakes are thickly covered by two imported American species, the totora reed (an important building material) and *Polygonum acuminatum* (a medicinal plant). A number of cultivated species of plants were also introduced partly from America and partly from Polynesia before the arrival of Europeans; of these the principal species was the sweet potato, which was cultivated in extensive plantations and formed the staple diet. Bottle gourds, sugarcane, bananas, taro, yams, and two useful trees (*i.e.*, the Asiatic paper mulberry, with bark used for cloth manufacture, and the American *Triumfetta semitriloba,* with bark used for rope making) were of aboriginal importation, as also probably were the husk-tomato, a small variety of pineapple, and the coconut.

Reintroduction of the toromiro tree

Before the arrival of human beings, the only vertebrates were either fish or seabirds capable of long flights. The animal life on land was otherwise restricted to a very few species of isopods (an order of crustaceans), spiders, insects, worms, a snail, and a centipede. Vast quantities of flies, large cockroaches, and a small scorpion were introduced recently. A small, long-legged chicken reported to have laid blue eggs was introduced in pre-European times but later interbred with European varieties. The aboriginal edible Polynesian rat was subsequently replaced by larger European species. Sheep, horses, cattle, and pigs were introduced by the missionaries who established themselves ashore in 1864. Sheep were especially numerous for almost a century after foreign ranchers began commercial ranching in 1870; sheep ranching came to an end in the mid-1980s, but cattle ranching was enhanced. A large wild cat, living in caves, is of unknown introduction. A Chilean partridge, a quail, and a small hawk have been added to the wildlife since 1880. Sea turtles and seals are now rare curiosities, but crayfish and various coastal and deep-sea fishes abound around the coast.

The people. The island's population represents the easternmost settlement of a basically Polynesian subgroup that probably derived from the Marquesas group. The original Rapa Nui vocabulary has been lost except for some mixed Polynesian and non-Polynesian words recorded before the Tahitian dialect was introduced to the decimated population by missionaries in 1864. Today Spanish is generally spoken. In their traditions, the islanders consistently divide themselves into descendants of two distinct ethnic groups, the "Long-Ears" and the "Short-Ears" (see below). Intermarriage is common, and an influx of foreign blood has become increasingly dominant in recent years.

The economy. Whereas the aboriginal economy was based on the cultivation of sweet potato, chicken raising, and coastal fishing, the island has shifted to a cash economy based on tourism. The opening of an airport near Hanga Roa has permitted an increasing influx of tourists since the 1960s, and a few small hotels have been built in the village area, where many islanders and settlers from continental Chile also have accommodations in their homes. The ties to continental Chile are strengthened through twice-weekly flights from Santiago and the building of schools, hospitals, and a large community hall for sports and performances. A well-organized Chilean national park system provides guided tours and security for the unique archaeological monuments. Reforestation projects have been successfully initiated, including eucalyptus plantations at Vaitea and coconut groves in Anakena Bay.

Chilean national park system

HISTORY

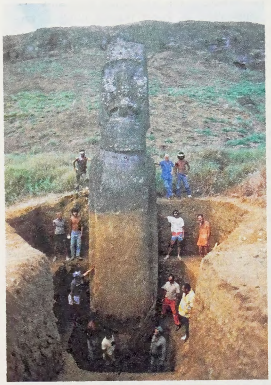
The first European to land on Easter Island was the Dutch admiral Jacob Roggeveen, who paid it a single day's visit in 1722. He and his crew found a population that they described as being of mixed physical types who worshiped huge standing statues with fires while they prostrated themselves to the rising sun. Some of them, said to be "white men," had their earlobes slit and hanging to their shoulders, a distinctly non-Polynesian custom.

An expedition dispatched by the Spanish viceroy of Peru rediscovered the island in 1770. The Spanish spent four days ashore and were the first to report that the aborigines had their own local form of script. They estimated a population of some 3,000 persons.

A civil war seems to have raged on the island before the arrival of the British navigator Captain James Cook in 1774; a decimated, poverty-stricken Polynesian population of only about 600 or 700 men and fewer than 30 women was found by the Englishmen, who also observed that the large statues were no longer venerated, most of them having been deliberately overthrown. In 1786 the French navigator Jean-François de Galaup, comte de La Pérouse, arrived and found some 2,000 people on the island; he tried in vain to introduce domestic animals. A number of sailing vessels, including whalers, visited the island from 1792 onward. By 1860 the population was about 3,000, but a major slave raid launched from Peru in 1862, followed by smallpox epidemics, reduced the population to 111 in 1877. At the end of the 19th century it began to increase once more. In 1864 Brother Eugène Eyraud, a French Catholic missionary, became the first foreigner to settle on the island; as a result, the population became converted to Christianity by 1868. Settlers from Tahiti began to raise sheep in 1870. In 1888 the island was annexed by Chile, which leased nearly all its territory for sheep raising; in 1954 the Chilean navy administration took over the sheep range. In 1965 a civilian governor was appointed by the Chilean government, and the islanders became full Chilean citizens. Within a single generation the Easter Islanders successfully responded to a complete acculturation to continental standards without losing their pride in their own ancestors and their skills and customs. Annually in February old and young of both sexes meet in contests to revive the arts and practices of the island's past, including carving, tattooing, reed-boat building, and traditional singing and dancing.

Archaeology. The island is famous for its gigantic stone statues, of which there are more than 600, and for the ruins of giant stone platforms (*ahus*) with open courtyards on their landward sides, some of which show masterly construction. Archaeological surveys were carried out in 1886, 1914, and 1934; archaeological excavations were initiated in 1955. The excavations revealed that three distinct cultural periods are identifiable on the island. The early period is characterized by *ahus* at Tahai, Vinapu, and Anakena, carbon dated to *c.* AD 700–850. The first two were admired and described by Captain Cook; the wall in Anakena remained hidden below ground until it was excavated archaeologically in 1987. The excavations in Anakena have revealed that a variety of statues were carved in the early period, among them a smaller prototype of the middle-period busts, which mainly differ from the

The stone statues



The tallest standing Easter Island stone statue (about 37 feet [12 metres] high) after being excavated by Thor Heyerdahl (top right, leaning against statue); it was subsequently partially buried again.
By courtesy of Thor Heyerdahl; photograph, Walter Leonard

latter by their rounded heads and stubby bodies. Another type was a realistic sculpture in full figure of a kneeling man with his buttocks resting on his heels and his hands on his knees, in one case with his ribs exposed, all features characteristic of pre-Inca monuments at Tiahuanaco in South America. In the middle period, *c.* 1050–1680, statues were deliberately destroyed and discarded, and all *ahus* were rebuilt with no regard for solar orientation or masonry fitting. The sole desire seems to have been to obtain strong platforms capable of supporting ever taller and heavier busts, the classical *moai* of the middle period.

Burial chambers also were constructed within the *ahus* in the middle period. The sizes of the statues made were increased until they reached stupendous dimensions; the slim and lofty busts also had huge cylindrical *pukao* (topknots) of red tuff placed on top of their slender heads. Most middle-period statues range from about 10 to 20 feet in height, but the biggest among those formerly standing on top of an *ahu* was about 32 feet tall, consisted of a single block weighing about 82 tons (74,500 kilograms), and had a *pukao* of about 11 tons balanced on its apex. The largest statue still standing partly buried in the deep silt below the quarries is about 37 feet tall, and the largest unfinished one with its back attached to the rock is about 68 feet tall. Traditions, supported by archaeology, suggest that the images represented important personalities who were deified after death.

Unfinished carvings

Statues of the middle period were all quarried from the special yellow-gray tuff found in the crater walls of Rano Raraku. Inside and outside the crater bowl numerous unfinished statues and thousands of crude stone picks are scattered about, bearing witness to a sudden interruption of the sculptors' work. The unfinished images show that each statue had its front and sides completed to a polish before the back was detached from the bedrock. The image was then slid away to be raised at random in the rubble below the quarries to have the back finished before being moved to some distant *ahu*. Eye cavities and topknots were added only after the monument was erected; recent discoveries have revealed that these concavities had inlaid eyes of white coral with a dark stone disk as pupil. From one to a dozen completed statues would stand in a row on a single *ahu*, always facing inland.

Experiments based on island traditions in 1955–56 showed that the numerous basalt picks left in the quarries were perfectly suitable for carving the hard tuff. Reenactments showed that 12 islanders were able to lift a 25-ton statue about 10 feet off the ground and to tilt it on end on top of an *ahu;* this work took 18 days with no tools other than two wooden logs that were used as levers. Stones of all sizes were wedged under the statue one by one to form a slowly rising cairn in order to lift the giant monoliths upright. Tradition claimed that the statues had "walked" across the terrain to their distant destinations, but in the experiment 180 islanders were able to pull a medium statue over the ground. A renewed experiment in 1986 revived the tradition and discovered that 15 men sufficed to move a medium-sized statue over the ground in upright position by jerking it ahead with a system of ropes.

The middle-period busts clearly evolved from a local prototype and have no counterpart elsewhere. Also peculiar to the middle period was a bird cult with attendant birdman rites that survived into the third, or late, period. Its ceremonial centre was the village of Orongo, on top of Rano Kao, which consisted of stone houses with roof vaults built as false arches. These houses and contiguous circular masonry dwellings with roof entrances are characteristic of the early and middle periods on the island; while unknown elsewhere in Polynesia they are common in the adjacent area of South America.

Traditional culture. The late-period Easter Islanders dwelt in boat-shaped pole-and-thatch houses or in caves. This period was marked by internal wars, general destruction, and cultural decadence. The *mataa,* or obsidian spearpoint, which was mass-produced, is the characteristic artifact of this period. Wood carving and small crude stone figurines replaced monumental art. Written wooden tablets covered with incised signs (called *rongo-rongo*) placed in boustrophedon (a method of writing in which the lines run alternately from right to left and from left to right) were copied from earlier specimens merely for ritual purposes; their proper reading was forgotten, and—despite many claims—modern attempts at deciphering them have failed. During this period art treasures were hidden in secret family caves, while the upright *ahu* images were successively overthrown. Silt from the abandoned quarries descended to the chests of the blind and unfinished busts standing at the foot of the volcano, rendering their overthrow impossible and thus securing for posterity the eyeless heads that have given the island its fame.

Tradition maintains that destruction began after a period of peaceful coexistence between two people of different culture and language—the Long-Ears and the Short-Ears. The latter, tired of toiling for the former, all but exterminated them in a pyre along an ancient ditch at Poike on the far northeastern coast. Carbon dating and genealogies concur in placing this event and the beginning of the late period at about AD 1680. The original construction of the artificial Poike ditch, according to carbon dating, took place in about AD 380.

The First International Science Congress convening on Easter Island in 1984 agreed on a resolution defining the island as the site of a pre-European civilization. The recent excavations, which reveal that the earliest settlers arrived with previously developed architectural concepts and a highly specialized megalithic masonry technique, support island traditions, which claim that the first ancestors arrived in an organized party of emigrants and not merely as casually wind-driven fishermen. (Th.H.)

French Polynesia

French Polynesia (French: Polynésie Française) comprises about 130 islands with a total land area of some 1,550 square miles (4,000 square kilometres) scattered over an area of the south-central Pacific between latitudes 7° to 27° S and longitudes 134° to 155° W. The islands are administered as an overseas territory of France.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* There are five distinct groups of islands, all protrusions of parallel submarine ridges trending from the northwest to the southeast.

The Society Islands are the most westerly of the group and the most important in terms of land area (40 percent) and population (85 percent). Except for a few small atolls,

the Society Islands are of the high-island type, their formation having resulted from the emergence of underwater volcanoes. The volcanic cones are highly eroded and cut up into high crests and deep, radiating valleys. The often lushly vegetated mountains drop abruptly to narrow coastal strips or directly into lagoons or the sea. The islands are protected from the force of the sea by almost completely encircling barrier reefs.

Tahiti and Moorea

The most highly populated of the Society Islands are Tahiti and its neighbour, Moorea, both situated in the eastern Windward Group, or Îles du Vent. Tahiti, formed of two ancient volcanic cones, is particularly striking because of its dramatic silhouette, which rises 7,352 feet (2,241 metres) above sea level. The mountains are empty of all human settlement, habitation and planting being entirely limited to the coastal strip and valley outlets of the island. Moorea, separated from Tahiti by a channel 8.5 miles (14 kilometres) wide, is also a high island and is encircled with very white coral sand beaches. It is well-connected to Tahiti by boat and taxi planes—a consequence of the booming tourist industry.

Some 75 miles to the west of Tahiti are the Leeward Islands, or Îles sous le Vent, made up of five high islands and four atolls. They closely resemble the Windward Islands in appearance. Raiatea, a double island group, is the largest and most densely populated of the Leeward Group. Separated by a channel that is about two miles wide, Raiatea and its northern neighbour, Vaitoare, are located on the same mountain mass and lie within a single barrier reef. Both have coastal plains suitable for growing coconut palms and raising stock. The growing of vanilla, once an important crop, declined in the 1970s. The Leeward Group's port is Uturoa, located on Raiatea. To the east of Raiatea is the picturesque island of Huahine, a volcanic structure bisected by a shallow arm of the sea.

The decline of vanilla growing

Finally, to the west of Raiatea lies the beautiful little island named Bora-Bora. It is formed from two volcanic peaks rising to 2,385 feet and 2,169 feet and dropping abruptly to the lagoon. Bora-Bora is one of the centres of the tourist trade in French Polynesia.

The Tuamotu Archipelago, lying to the east of the Society Islands, has an area of 280 square miles and consists of some 80 islands. These are low, flat islands or atolls of coral origin, surrounding a lagoon. Their size varies greatly: the largest ones, such as Rangiroa, reach 29 square miles; the smallest are made up of a few acres of land barely protruding above the surface of the sea. With only porous, coral-based soils and with no permanent streams, they have no agricultural potential aside from the everpresent coconut trees. The lagoons, however, are a source of fish, pearls, and mother-of-pearl shell. Only Rangiroa, with its airport, is in close contact with Tahiti. Elsewhere, living conditions are difficult, and many people emigrate to Tahiti. French nuclear installations are situated in the Tuamotus; test sites are on Mururoa and Fangataufa, both ceded to France in 1964.

Morphologically different, the Îles Gambier (Mangareva) lie at the southern extremity of the Tuamotu Archipelago and include four large, high islands and a few islets (14 square miles). The main island is Mangareva.

The 14 islands of the Marquesas group lie 900 miles to the northeast of Tahiti. They have a land area of 405 square miles. Some of them are high islands (over 4,000 feet), with sharp and twisting contours. Unlike the Society Islands, they are not protected from the sea by a barrier reef, with the result that they lack a coastal plain. Approaching the islands from the sea is difficult. People live exclusively in the valleys, where they engage in farming.

The Austral Islands, or Îles Tubuai, situated 450 miles south of Tahiti, make up the southernmost part of the territory. This chain of four islands, with the addition of the isolated island of Rapa in the southeast and the uninhabited Marotiri and Maria islands, covers 57 square miles. All of the islands are of volcanic origin but are relatively low (270 to 1,440 feet) and have unpronounced contours. Income is derived from agriculture (taro, arrowroot, copra, market vegetables) and pandanus plaiting.

Like the Marquesas and the Gambier-Tuamotus, the Austral Islands have poor connections with Tahiti. As elsewhere, the hard living conditions cause many people to migrate to Tahiti and Papeete.

Climate. The climate is tropical—warm and humid. A warm rainy season lasts from November to April, and a relatively cool dry season from May to October. The dispersion of the islands through 20° of latitude, however, results in regional climatic variation. Except in the Marquesas and the northern Tuamotus, rainfall is abundant, falling in violent showers. As much as 120 inches (3,050 millimetres) falls on the coastal areas. There are local variations due to differing exposures; on average, the windward coasts receive more precipitation.

Abundant rainfall

The temperature varies only slightly throughout the year. At Papeete the average annual temperature is 79° F (26° C); the high average is 91° F (33° C) in March and the low average 70° F (21° C) in August. The Austral Islands, farther south, enjoy a cooler climate; the low average can go down to 64° F (18° C) in September. The relative humidity is always high—80 to 90 percent on the average. The high areas are continually enveloped in a heavy cloud formation.

The territory is in the trade-wind zone. The dominant winds thus blow from the north and northeast, but they tend toward the southeast between May and October. There are long periods of calm (April, May, June) but with occasional typhoons. During occurrences of El Niño Current the territory is frequently in danger of typhoons.

Plant and animal life. Because of the recent origin and the isolation of the islands, there is little variety in terrestrial flora and fauna. Most of the plant species were introduced by the first Polynesians, and others were introduced by Europeans.

The vegetation's appearance varies with the ecological conditions. On the limestone soils of the atolls, it has a pronounced xerophilous (desert-plant-type) character. On the high volcanic islands it is more diversified; ferns have often conquered the hills and plateaus, whereas rain forests are established in the upper valley areas. On coastal plains coconut, breadfruit, and various fruit trees flourish.

The land fauna is especially limited, and most of the species have been introduced. No mammals are indigenous to the islands; there are feral goats, pigs, horses, cattle, and rats. A fish, called *nato,* and a variety of shrimp are found in the islands' freshwater streams. The marine life in the lagoons and surrounding seas is rich.

Patterns of settlement. On the high islands, homes are scattered through the coconut groves along the coastal roads. Villages are spaced two to three miles apart and typically have a church, government house, school, shop, pastor's home, and a few residences. The contemporary rural house is of concrete construction in a yard shaded by fruit trees, with a separate kitchen made from traditional materials (palm, bamboo) where food is prepared and eaten. On the atolls, the population is usually grouped together in villages located close to the passes through the surrounding reefs. Population and business activity tend to concentrate in the Papeete area. The town consists of the old colonial city (still the business centre), residential areas (often on the heights), and an excellent harbour.

The people. Most of the people throughout the islands may be classed as Polynesian, speaking eastern Polynesian languages (although intermixture has left few pure Polynesians). There are minorities of Chinese, Europeans, and Americans. About 95 percent of the population is Christian. More than one-half is Protestant, about 40 percent is Roman Catholic, and the remainder belongs to various other Christian denominations. The Evangelical Church of French Polynesia is the dominant church.

The economy. Many resources are used for local subsistence, including fruits, products from fishing and planting, and materials for the construction of traditional types of houses and canoes. Pigs, cattle, and chickens are raised for food. The traditional exports—copra and vanilla—have greatly declined, but this loss of revenue has been partially compensated for by the development of fishing, especially with the extension of territorial waters to 200 miles offshore in 1978. Shrimp and oysters are farmed; and the cultured-pearl industry is important, especially on the Tuamotu and Gambier island groups.

The development of fishing

Manufactured products include coconut oil, processed foods, printed textiles, construction materials, copper wares, beer and soft drinks, and sandals; traditional handicrafts and boats are made on some of the outer islands. Hydroelectric power plants on Tahiti began service in the early 1980s. Tourism has increased markedly.

Transportation facilities have been improved considerably since the early 1960s, including the construction of a modern port in Papeete, construction of an international airport at Faaa, and development of air services with some of the outlying islands: Moorea, the Leewards, the western Tuamotus, the Marquesas, and the Australs. There is scheduled shipping to other regions in the Pacific, but inter-island shipping remains largely dependent on local unscheduled schooner sailings. The road network has been increased to several hundred miles of paved or stone-surfaced roadway.

Administration and social conditions. Represented in the French parliament by two deputies and a senator and placed under French law, the territory is administered locally under the statutes of 1977 and 1984. These provide for a popularly elected territorial assembly with 41 seats, a high commissioner appointed by the French government, and a president and Cabinet selected from the assembly.

Schooling is compulsory to the end of primary school and is conducted largely in government schools, supplemented by mission schools.

Cultural life. Polynesia's cultural and artistic traditions have been misrepresented and to an extent have been reduced to a sort of folklore by the romantic image that Europeans adopted. A local ethnographic museum and a local learned society have contributed to efforts to preserve the territory's cultural heritage. The absence of newspapers in Polynesian languages, the small amount of broadcasting in the Tahitian language, and the nonrecognition of vernacular languages as official languages all threaten what survives of Polynesian culture.

For statistical data on the land and people of French Polynesia, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Archaeological evidence suggests that the Marquesas Islands may have been settled in about 200 BC from western Polynesia. In subsequent dispersions, Polynesians from the Marquesas migrated to the Hawaiian Islands in about AD 300 and reached the Society Islands by about the 9th century AD. Large chieftainships were formed on Tahiti, Bora-Bora, and Raiatea. Teriaroa, north of Tahiti, was a royal retreat, and Taputapuatea, on Raiatea, was the most sacred shrine in the islands.

European contact with the islands that later came to constitute French Polynesia was gradual. The Portuguese navigator Ferdinand Magellan sighted Pukapuka in the Tuamotu group in 1521. The southern Marquesas Islands were discovered in 1595. The Dutch explorer Jacob Roggeveen in 1722 discovered Makatea, Bora-Bora, and Maupiti. Captain Samuel Wallis in 1767 discovered Tahiti, Moorea, and Maiao Iti. The Society Islands were named after the Royal Society, which had sponsored the expedition under Captain James Cook that observed from Tahiti the 1769 transit of the planet Venus. Tubuai was discovered on Cook's last voyage, in 1777.

Sighting by Magellan

The history of the Society Island groups is virtually that of Tahiti, which was made a French protectorate in 1842 and a colony in 1880. French missionaries went to the Gambier group in 1834, and in 1844 a French protectorate was proclaimed, followed by annexation in 1881. The Austral Islands were also evangelized from Tahiti, and as late as 1888 Rimatara and Rurutu sought British protection, which was refused. They were placed under the French protectorate in 1889 and annexed in 1900. The Tuamotus were part of the kingdom of the Pomare family of Tahiti, which came originally from Fakarava. These islands were claimed as dependencies of Tahiti within the protectorate by France in 1847 and became part of the colony in 1880. In the Marquesas, Nuku Hiva was annexed to the United States in 1813 by Captain David Porter of the frigate *Essex*, but the annexation was never ratified. French occupation of the group followed the landing of forces from a French warship, requested by the chief of Tahuata (near Hiva Oa). Soon after there was a quarrel with the French; in 1842 the chiefs ceded sovereignty to France.

The islands were administered as the French Colony of Oceania. The colony was ruled by a naval government until 1885, when an organic decree provided for a French governor and Privy Council and for a General Council, representing the islands, that had some control over fiscal policies. The powers of the General Council, however, were cut back in 1899, and in 1903 it was replaced by an advisory council, the function of which was purely administrative. In 1940 the voters on the islands chose to side with the Free French Government of Charles de Gaulle, and many islanders fought alongside Allied armies during World War II. French Polynesia was made an overseas territory of France in 1946. It was provided with a territorial assembly and was allowed to elect one representative to the French National Assembly and one to the French Senate.

In 1957 the French government extended the powers of the local Territorial Assembly, and the following year Pouvanaa a Oopa, vice president of the Council of Government, announced a plan to secede from France and form an independent Tahitian republic. He was subsequently arrested; the movement collapsed, and local powers were again curtailed. France granted more local autonomy in 1977, but the pro-independence and pro-autonomy parties continued to call for popular election of the president and either more autonomy or outright independence. (F.J.W.)

Nuclear testing

In 1963 the French government began a nuclear testing program on Mururoa atoll. Mururoa and neighbouring Fangataufa were ceded to France by the territorial assembly in 1964. In response to worldwide pressure the tests were moved underground in 1975. The territory became autonomous in 1984. The following year world attention focused on the area when the French secret service blew up a yacht owned by the Greenpeace environmental group as it was preparing to lead a protest near Mururoa atoll. During 1995–96 France detonated additional nuclear devices in the territory, but protests from Polynesian, French, and foreign groups caused the government to reduce the number of explosions from eight to six, the last of which took place below Fangataufa atoll in January 1996.

(F.J.W./Ed.)

For later developments in the history of French Polynesia, see the BRITANNICA BOOK OF THE YEAR.

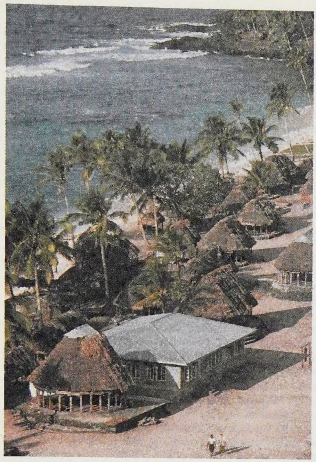
Samoa

Samoa, formerly (until 1997) Western Samoa, is among the westernmost island nations of Polynesia. It lies about 80 miles (130 kilometres) west of American Samoa, 1,800 miles northeast of New Zealand, and 2,600 miles southwest of Hawaii. Samoa, which shares the Samoan archipelago with American Samoa, consists of nine islands west of longitude 171° W—Upolu, Savai'i, Manono, and Apolima, all of which are inhabited, and the uninhabited islands of Fanuatapu, Namu'a, Nu'utele, Nu'ulua, and Nu'usafee. The total land area is 1,093 square miles (2,831 square kilometres)—about 2.5 times larger than Hong Kong. The capital is Apia, on Upolu.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* Savai'i, the largest island, covers 659 square miles and rises to a maximum elevation of 6,095 feet (1,858 metres) at Mount Silisili, a volcanic massif at the island's approximate centre. Mounts Mata'aga and Maugaloa are also imposing peaks. Upolu, the other large island, is more elongated and uneven in shape than Savai'i, but with lower average elevations. It covers an area of 432 square miles, including five offshore islets, and rises to 3,608 feet at Mount Fito. Manono and Apolima are smaller islands lying between the two main islands. Mount Matavanu on Savai'i last erupted intermittently during 1905–11. Samoa's volcanic soils support lush vegetation but are easily eroded.

Climate. The climate is tropical and humid. Precipitation varies from more than 100 inches (2,540 millimetres) on the northern and western coasts to 300 inches inland.



Cluster of traditional *fale* and a present-day rectangular house with walls and windows in the village of Papa, Savai'i, Samoa.

© Nicholas DeVore II/Bruce Coleman Inc.

Temperatures vary little, averaging 79° F (27° C) and ranging from 73° to 86° F (23° to 30° C) throughout the year. The southeast trade winds prevail, varying occasionally to northerlies during the wet season (November to April), when typhoons and other severe storms are likely to occur.

Plant and animal life. Samoa's lush vegetation includes inland rainforests and cloud forests. Large sections of the coast have been covered with taro plantations and coconut groves. The islands support limited animal life, although more than 50 species of birds are found there, at least 16 of them indigenous, including rare tooth-billed pigeons. The only native mammals are fruit bats ("flying foxes"), which are endangered, and smaller bats. Rats, cattle, and pigs have been introduced. Among the smaller animals are lizards, two snakes of the boa family, centipedes and millipedes, scorpions, spiders, and a wide variety of insects. O Le Pupu Pue National Park (1978), Samoa's first national park, occupies some 11 square miles on south-central Upolu. Soil erosion, resulting from farming steep slopes and clear-cutting forests, has produced runoff that has damaged many lagoons and coral reefs. Industrial and residential pollution has increased in and around Apia.

The people. Although some Samoan values and customs have changed markedly since European contact, particularly in Apia, Samoans have striven to preserve the *fa'a Samoa* ("Samoan way of life"); thus, many traditions and outward features of rural life have remained virtually unchanged. Samoans are mainly of Polynesian heritage, and their nation is traditionally known as the Cradle of Polynesia because Savai'i is said to be Hawaiki, the Polynesian homeland. About nine-tenths of the population are ethnic Samoans. Euronesians (people of mixed European and Polynesian ancestry) account for most of the rest, and a tiny fraction are of wholly European heritage. The Samoan language is closely related to the Maori, Tahitian, Hawaiian, and Tongan languages. A large number of Samoan words reflect maritime traditions, including names for ocean currents, winds, landforms, stars, and directions. Some verb forms indicate the relative positions of objects, including directions of movement toward or away from the speaker. English is widely spoken as a second language.

Samoan language

Samoans readily adopted Christian teachings following European contact, and even the more remote villages built churches, often of grand proportions. The Congregational Christian Church in Samoa (formerly the London Missionary Society) was dominant until the late 20th century, but it has since lost many adherents to the Mormon church. Mormon and Congregationalist groups now each account for roughly one-fourth of the population, whereas Roman Catholics account for one-fifth and Methodists one-ninth. There are also Pentecostals, Seventh-day Adventists, and other Christian groups.

Most Samoans have lived in coastal villages since the region was first settled, and about four-fifths of the population is still rural. Apia, on the northern coast of Upolu, is the nation's only town as well as the main port and centre for services and trade; it contains approximately one-fifth of Samoa's population. The birth rate has been high since the 1950s. About two-fifths of Samoans are less than 15 years old. Tens of thousands of Samoans have emigrated to New Zealand, the United States, Australia, and elsewhere.

Demographic trends

The economy. Close kinship ties within the villages traditionally bound Samoans into a collectivist society, but a cash economy developed following European contact, mainly based on agricultural exports. Agriculture accounts for two-fifths of Samoa's gross domestic product (GDP) and nearly two-thirds of the workforce; however, large quantities of food are imported. Major crops include coconuts, taro, pineapples, mangoes, and other fruits. Tourism, services, and light manufacturing have become increasingly important, as have remittances from Samoans living abroad (mainly in the United States and New Zealand), which account for as much as one-sixth of household income. Samoa has a persistently negative balance of trade. Major trading partners include New Zealand, Australia, Fiji, the United States, Japan, and American Samoa.

International flights connect the islands with American Samoa, Fiji, New Zealand, Australia, and Japan. Regular shipping services link with ports abroad. About two-fifths of Samoan roadways are paved, including many coastal highways and the major streets of Apia.

Administration and social conditions. *Government.* In 1962 Samoa promulgated its constitution as the first independent microstate in the Pacific region, and in 1970 it joined the Commonwealth. Universal suffrage for Samoans aged 21 years and older was instituted in 1990. Samoa has a parliamentary government that blends Samoan and New Zealander traditions. The constitution originally provided for a constitutional monarchy under two co-heads of state, with the provision that when one died (as happened in 1963) the other would continue as sole monarch and head of state for life, after which future heads of state would be elected by the Legislative Assembly (Fono Aoao Faitulafono) to five-year terms. The prime minister is elected by the assembly and appoints a cabinet from among its members.

Fono Aoao Faitulafono

The Legislative Assembly has 49 members. Two are directly elected by non-Samoan and mixed ethnic groups, and the remaining 47 are directly elected from among candidates who are Samoan *matai* (the leaders of extended families and villages). The justice system is headed by a Supreme Court, the judges of which also preside over the Court of Appeal. Samoa has a police force but no standing military; New Zealand is bound by treaty to provide military assistance upon request.

Education. Nearly all Samoans are literate. Education is compulsory between the ages of five and 14; however, only a small fraction of the population has completed secondary school. The University of the South Pacific has its School of Agriculture at Alafua, near Apia. The National University of Samoa was founded in 1984.

Health and welfare. Immunization programs since the late 20th century have greatly reduced the incidence of disease, particularly among children; however, there are few doctors, and the quality of hospital care is limited. Life expectancy at birth is 67 years for males and 72 for females. The leading causes of death are congestive heart failure, cancers, cerebrovascular diseases, accidents, pneumonia, and septicemia. Water shortages are common because of limited watersheds. (P.R.C./A.We./Ed.)

For statistical data on the land and people of Samoa, see

the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Polynesians traveling in outrigger canoes arrived in the Samoan islands (now Samoa and American Samoa) about 1000 BC, as indicated by Lapita pottery shards found in Mulifanua Lagoon on Upolu. Characteristics of the Samoan language indicate that the settlers probably came from Tonga. Local pottery manufacturing ceased by about AD 200, by which time Samoa had become central to much of the settlement of eastern Polynesia. Contact between Samoans, Tongans, and Fijians continued and was recorded in hundreds of legends and genealogies that were passed down through oral literature. Like other Polynesian peoples, Samoans were master navigators, boatbuilders, and fishers; every aspect of their society was related in some way to maritime life. Basic agriculture was also developed, including the cultivation of yams, taro, breadfruit, bananas, sugarcane, and coconuts. Most Samoans lived in villages ruled by councils of *matai*, and numerous fortified villages of 30 or more houses grew up along the coast. Extended blood ties linked family groups and villages, with the major families striving for supremacy and regularly plunging the islands into warfare.

European navigators, who began to visit Samoa in 1722, were at first welcomed for the technology and goods that they brought. John Williams, a member of the London Missionary Society, arrived to establish a Christian mission in 1830. He made a convert of Malietoa Vainu'upo, who had just conquered all of Samoa, and the rest of the population soon followed suit. A foreign settlement had developed around Apia Harbour by the 1850s. Samoans began to resist, however, as more settlers arrived from the United States, Great Britain, and Germany and tried to persuade their respective governments to annex Samoa. Rival *matai* played the three foreign powers against each other in pursuit of their factional wars, and they, in turn, frustrated attempts by the Samoans to establish a national government. In 1878 the United States signed a treaty allowing it to establish a naval station in Pago Pago Harbor (now in American Samoa). Great Britain and Germany signed similar agreements the following year. Warfare between the three powers in 1889 was prevented only by a great typhoon, which sank six of their warships. They subsequently signed the Berlin Act to provide for the neutrality of the islands and to avoid further conflict; however, in 1899 the United States annexed eastern Samoa, whereas Germany annexed the western part of the islands—subsequently called Western Samoa until 1997. Many Samoans resented the division, which had been made without their consent.

Mau a Pule

In Western Samoa the drive for political independence began in 1908 with the Mau a Pule, a movement led by the orator chief Lauaki Namulau'ulu. The *matai* were dissatisfied with the German governor's attempts to change the *fa'a Samoa* and centralize all authority in his hands. After the governor called in warships, Lauaki and nine of his leading supporters surrendered, whereupon they were tried and exiled to Saipan in the Mariana Islands.

Troops from New Zealand occupied Western Samoa in August 1914, meeting no resistance from the German or Samoan populations. However, the New Zealand administration was accused of negligence after more than one-fifth of Western Samoans died during the Influenza Epidemic of 1918–19, and most Samoans united against foreign rule. The League of Nations nevertheless granted New Zealand a mandate over Western Samoa in 1920. An organized political movement called the Mau ("Strongly Held View") subsequently emerged, led by Olaf Frederick Nelson, whose mother was Samoan. New Zealand outlawed the movement, claiming that Nelson and other "part-Europeans" were misleading the Samoans. New Zealand troops were sent in, and Nelson was exiled. During a Mau demonstration in December 1929, the *matai* Tupua Tamasese Lealofi III and other unarmed Mau supporters were shot and killed by New Zealand troops.

New Zealand's first Labour government, which came to power in 1935, recognized the Mau as a legal political organization. Relations between Samoans and New Zealanders improved somewhat, but many Samoans remained dissatisfied. The islands' economy improved during World War II, when a garrison of U.S. troops was stationed on Upolu and built several roads and an airport. New Zealand allowed a Western Samoan council of state and a legislative assembly to be established in the late 1940s, and a constitutional convention met in 1954. Several government reforms were carried out during the next few years, amid continuing Samoan agitation for independence.

Independence in 1962

In 1962 Western Samoa became the first Pacific nation of its size to regain its political independence. The monarch Susuga Malietoa Tanumafili II became the co-head of state in 1962 and head of state (O le Ao o le Malo) the following year. Major political figures in the late 20th century included Fiame Faumuina Mataafa, who served twice as prime minister (1962–70 and 1973–75) and Tupuola Taisi Efi, who was prime minister during 1976–82. The religious makeup of Samoa was altered markedly in the late 20th century, as many in the country joined the Mormon church. The tourist trade increased during the same period, partly because of improvements to Upolu's transportation infrastructure. Tofilau Eti Alesana, who served as prime minister during 1985–98, had mixed success with the economy and engendered controversy by attempting to censor critics of the government. In 1997 the Legislative Assembly changed the country's name from Western Samoa to Samoa, despite protests from neighbouring American Samoa. During the following year, wildfires that were started by farmers clearing land for cultivation destroyed nearly one-fourth of the forests on Savai'i.

(A.We./Ed.)

For later developments in the history of Samoa, see the BRITANNICA BOOK OF THE YEAR.

Tonga

Tonga is a constitutional monarchy and a member of the Commonwealth. Its official name is Pule'anga Fakatu'i 'o Tonga, or Kingdom of Tonga. Situated in the southwestern Pacific, it consists of 169 islands divided into three main island groups: Tongatapu in the south, Ha'apai in the centre, and Vava'u in the north. Isolated islands include Niuafo'ou, Niuatoputapu, and Tafahi in the far north and 'Ata in the far south. Tonga's total land area of 290 square miles (750 square kilometres) is dispersed between latitudes 15° to 23° S and longitudes 173° to 177° W. The capital, Nuku'alofa, on the island of Tongatapu, is approximately 1,245 miles (2,000 kilometres) northeast of Auckland, N.Z.

PHYSICAL AND HUMAN GEOGRAPHY

The land. *Relief.* The summit of volcanic undersea mountains forms the two roughly parallel chains of the Tongan islands. Most of the islands of the western chain are classified as high islands because they have been raised well above sea level by repeated volcanic activity. Four of them are still active volcanoes. Some of the islands composed of lava formed by shield volcanoes, such as Late and Kao, have a cone-shaped, hard surface that is not easily eroded. Others, such as Fonuafo'ou (Falcon Island), were formed by more explosive volcanoes, and their surfaces, composed of ash and pumice, have eroded readily.

The low islands of the eastern chain have been capped by coral polyps and foraminifera (marine organisms that have calcareous shells), which build coral rock and limestone reefs. The growth of coral counteracts the sea's erosion of the reefs and the islands enclosed by them. A protective reef surrounds Tongatapu Island; many islands in the Vava'u group lack such protection and are shrinking.

Erosion and coral growth

Tongatapu Island, a raised atoll in the Tongatapu group, is the largest island in Tonga, with an area of 99 square miles, and is the most densely populated. The highest point in Tonga, 3,380 feet (1,030 metres), is on Kao Island in the Ha'apai group. 'Eua Island (Tongatapu group) has an old volcanic ridge rising to 1,078 feet above sea level. The Vava'u group has hills ranging from 500 to 1,000 feet, and Vava'u Island has a fine, large landlocked harbour. The effects of natural erosion are particularly vivid in Vava'u.

Rainwater reacting with the carbon dioxide in vegetation acquires acidic properties and dissolves coral and limestone rock, thereby forming caves. The sheer cliffs and sandspits of Vava'u and Nuku'alofa have been caused by the constant action of the waves. There are no rivers in Tonga, but creeks are found on 'Eua and on Niuatoputapu, one of the far northern islands. The geologic subsidence of Vava'u, followed by two phases of uplift, resulted in the formation of the Ano Lagoon.

Climate. Tonga has a semitropical climate except in the northernmost islands, where truly tropical conditions prevail. Temperatures range from 60° to 70° F (16° to 21° C) in June and July to 80° F (27° C) in December and January. The mean annual humidity is 77 percent. The mean annual rainfall varies from 63 inches (1,600 millimetres) in Tongatapu to 87 inches in Vava'u; on Niuatoputapu the annual rainfall averages 101 inches. Humidity is higher as the distance from the Equator lessens. The northern islands, which are closest to the Equator, are particularly vulnerable to typhoons between December and April.

Plant and animal life. The well-drained, fertile soils of 'Eua, Kao, Tofua, Late, and the slopes and hilltops of Vava'u support original forests. 'Eua has the greatest number and variety of trees, and the ridge on the eastern side is a forest reserve. Some of the more significant are the hiapo (or paper mulberry), the fau (*Hibiscus tiliaceaus*), the coconut, the ngatata, and the toi and tavahi, which constitute a majority of the trees on 'Eua. The sandy, rocky, dry soils of the coasts and the direct exposure to strong winds and salty spray create unfavourable conditions for coastal vegetation. To conserve moisture, plants near the shore have small, waxy, or hairy leaves. Tidal sand and mud flats have swampy areas that support mangroves. Behind the mud flats, trees with buttress roots, such as the *lekileki,* sometimes grow.

Tonga's land birds include doves, rails, starlings, kingfishers, owls, cuckoos, shrikes, bulbuls, purple swamp hens, and swiftlets. The red-breasted musk parrot and the blue-crowned lory, considered by some to be the most beautiful birds in the Pacific, inhabit 'Eua. Island cliffs serve as homes for red-tailed and white-tailed tropic birds. Among the native birds of Niuafo'ou Island is the incubator bird. The common reef heron is a native shore bird. Transient species include the golden plover, wandering tattler, long-billed curlew, and bar-tailed godwit. Tongan waters attract several varieties of seabirds such as noddies, terns, frigate birds, and mutton birds. The village of Kolovai on Tongatapu Island is a home of the fruit bat (often called "flying fox" because of its appearance), which clings to large trees by day and flies at night to forage for food.

The "flying fox," or fruit bat

Human settlement. About three-quarters of the Tongan population live in the three major island groups, and nearly two-thirds live on Tongatapu Island. The urban population now accounts for more than two-fifths of the total population. Many Tongans migrate overseas, to the United States and New Zealand in particular.

The majority of the people live in villages. Traditional structures are called *fale;* they are rectangular in shape and have thatched or corrugated tin roofs and sides made of woven coconut leaves, reeds, or timber. Some Tongans reside in South Seas colonial-style wooden homes with gingerbread trim and outside walls in pastel shades.

Nuku'alofa, the name of which signifies "Abode of Love," has all the amenities of a small capital city. It is also a major port of entry and has several wharves and piers. Much of Nuku'alofa's economic activity revolves around coconuts and coconut products. Other ports and commercial centres are Neiafu in the Vava'u group and Pangai in the Ha'apai group.

The people. Tongans are closely related to Samoans and other Polynesians in culture, language, and ethnic makeup. More than 95 percent of the population claims an original Polynesian ancestry. Contact with Fiji has resulted in the transference of some Melanesian physical characteristics. Intermarriage with Europeans has become more common, especially as a result of the increasing out-migration of Tongans since the 1970s. Religion is an important aspect of Tongan society, and most Tongan families are members of a church. About two-fifths adhere to the Free Wesleyan Church, one-sixth are Roman Catholics, one-seventh are Mormon, and the remainder belong to smaller denominations. The Tongan language is taught in primary schools and is the official language, in addition to English, which is studied as a second language.

The economy. Agriculture is the mainstay of the Tongan economy, contributing nearly one-third of the gross domestic product. Coconuts, bananas, squash, cacao, coffee, ginger, black pepper, and vanilla beans constitute the main cash crops; other important crops include yams, taro, cassava, corn (maize), watermelons, pineapples, and tomatoes. All land is essentially owned by the Crown, but large estates have been divided among the nobles of Tonga. Every male peasant aged 16 or over is entitled to an allotment of 7.5 acres (three hectares) of land for cultivation. Timber production, livestock raising, and fishing also contribute to Tonga's economy.

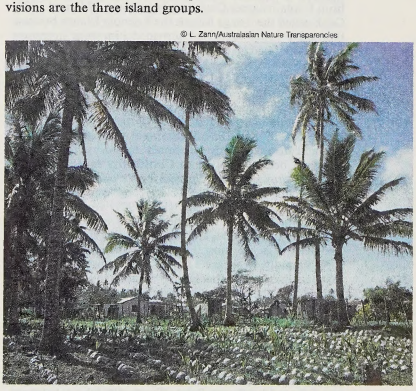
Manufactures include plastic pipe, canned goods, charcoal, corrugated iron, and small handicrafts. Crop processing has been undertaken by cooperative societies; the Tonga Feeds Manufacturing Society procures stock feed, and vanilla is processed by the Leimatu'a Vanilla Society located in the Vava'u group. Remittances from Tongans working overseas—especially in New Zealand, the United States, and Australia—and tourism have both contributed significantly to the growth of the Tongan economy.

Overseas workers

More than half of Tonga's road network is all-weather roads, almost all of which are located on the two largest islands. Tonga has no railroad. Nuku'alofa and Neiafu (Vava'u) are ports used for external shipping. Copra and bananas are exported from Pangai (Ha'apai group). Regular international air service to New Zealand, Fiji, Samoa and American Samoa, and Niue is available from Fua'amotu Airport on Tongatapu. Domestic flights are serviced by airports on 'Eua, Ha'apai, Vava'u, Niuafo'ou, and Niuatoputapu.

Administration and social conditions. Tonga's constitution, granted in 1875 by King George Tupou I and amended only slightly since, establishes a constitutional monarchy. The chief executive is the monarch, who appoints a Privy Council. The council consists of the monarch and the Cabinet. The Cabinet has a prime minister, deputy prime minister, six other ministers, and the governors of Ha'apai and Vava'u. The unicameral legislature consists of a speaker, the members of the Cabinet, nine nobles selected by the 33 nobles of Tonga, and nine representatives elected for three-year terms by literate taxpaying males and literate females aged 21 and over. Popularly elected town and district officials are the only form of local government. The most important administrative divisions are the three island groups.

© L. Zann/Australasian Nature Transparencies



Coconuts germinating on a copra plantation on Lifuka in the Ha'apai Group, Tonga.

Education is free for all Tongans, and attendance is compulsory between the ages of six and 14. The government runs primary, secondary, and vocational training schools, including a teacher-training college. Some primary and secondary schools as well as vocational institutions are run by churches. Tongans receive free dental and medical treatment, and the general health of the population is adequate. Family planning aids are disseminated with the help of the United Nations and New Zealand.

Cultural life. Although Western influence has somewhat altered traditions and culture in Tonga, certain Tongan rituals and art forms survive. For example, Tonga shares with Fiji, Samoa, and parts of French Polynesia the elaborate ritual surrounding the drinking of kava. The drink, prepared from the root of a pepper plant, has the properties of a mild narcotic.

Carving was traditionally done by men, but the craftsmanship was inferior to that of other Polynesians, such as the Maori. Carving and other traditional crafts of higher quality have been produced, however, in response to the demands of the tourist market. Women manufacture tapa cloth from bark and weave mats and baskets from several varieties of pandanus leaves. Traditional dancing is an important part of national ceremonies and local village festivities. In the popular paddle dance, called *me'etu'upaki,* dancers carry paddle-shaped boards painted or carved with abstractions of the human body. Other popular dances include the *kailao,* a war dance; the *lakalaka,* a group action song performed while standing; the *ma'ulu'ulu,* an action song performed while seated; and the *tau'olunga,* an individual dance that is accompanied by singing. An oral tradition persists in the form of proverbs, religious epics, genealogies, poetry, fables, and myths.

Tongan dance forms

For statistical data on the land and people of Tonga, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

HISTORY

Tonga was first inhabited about 3,000 years ago by Austronesian-speaking people, who made elaborately decorated Lapita ware. From at least the 10th century AD Tonga was ruled by a line of sacred kings and queens, the Tu'i Tonga. About 1470 the reigning Tu'i Tonga transferred his temporal powers to his brother under the title of Tu'i Ha'a Takalaua. A similar transfer of power about 1600 resulted in the creation of a third line of monarchs, the Tu'i Kanokupolu, who eventually became the rulers.

Although some islands were visited by the Dutch navigators Jakob Le Maire and Abel Janszoon Tasman in 1616 and 1643, respectively, effective European contact dates from Captain James Cook's visits between 1773 and 1777. Cook called the Tonga islands the Friendly Islands because the native inhabitants had provided him with necessary supplies and had given him a warm welcome. The London Missionary Society and a mission of Methodists made unsuccessful attempts to introduce Christianity to Tonga in 1797 and 1822, respectively. A renewed attempt by the Methodist mission in 1826 was successful, and a Roman Catholic mission was established by the Marists in 1842. Between 1799 and 1852 Tonga went through a period of war and disorder. This was finally ended by Taufa'ahau, who had been converted to Christianity in 1831 by the Methodist missionaries. He became Tu'i Kanokupolu and subsequently took the title King George Tupou I in 1845. During the king's long reign (1845–93), Tonga became a unified and independent nation with a modern constitution (1875), legal code, and administrative structure. With Taufa'ahau as its most important convert, Christianity spread rapidly. In separate treaties, Germany (1876), Great Britain (1879), and the United States (1888) recognized Tonga's independence. George I was succeeded by his great-grandson George II, who died in 1918. During his reign the kingdom became a British protectorate (1900) to discourage German advances. Under the treaty with Great Britain (amended in 1905), Tonga agreed to conduct all foreign affairs through a British consul, who had veto power over Tonga's foreign policy and finances. George II was followed by Queen Salote Tupou III, who ruled from 1918 to 1965. She was succeeded by her son Prince Tungi, as Taufa'ahau Tupou IV. In 1970 Tonga regained full control of domestic and foreign affairs and became a fully independent nation within the Commonwealth. In 1991 the son of King Taufa'ahau Tupou IV, Baron Vaea, was appointed prime minister; he was replaced by the king's youngest son, 'Ulukalala Lavaka Ata, in 2000. Many Tongans have agitated for democratic reforms, and reform-minded candidates won 5 of 9 available "commoner" seats in the legislative elections of 1999, the same year in which Tonga was admitted to the United Nations. However, the king and nobles have continued to exercise firm control over the government. (Si.L./Ed.)

For later developments in the history of Tonga, see the BRITANNICA BOOK OF THE YEAR.

Wallis and Futuna

Wallis and Futuna is a small overseas territory of France consisting of two Pacific island groups separated by an expanse of open ocean. The territory is in western Polynesia and is located south of Tuvalu, west of Samoa, and northeast of Fiji. The islands of Melanesia are to the west. The capital is Mata Utu, on Uvea (Wallis).

The island of Wallis takes its European name from its 18th-century British discoverer, Captain Samuel Wallis, but the indigenous name, Uvea, has a much greater antiquity. Neither Uvea nor Futuna should be confused with islands of the same or similar names elsewhere. Uvea is sometimes confused with Ouvéa in the Loyalty Islands near New Caledonia. Futuna has a namesake among the islands of the nation of Vanuatu (formerly the New Hebrides), and, indeed, folklore suggests that both the namesake and Sikaiana Island in the Solomon Islands were settled from Futuna in the pre-European era.

Uvea

PHYSICAL AND HUMAN GEOGRAPHY

The land. Uvea is a volcanic island, but it is relatively low, with a maximum elevation of only 476 feet (145 metres). Its total land area is 23 square miles (60 square kilometres). Uvea is surrounded by a barrier reef with 22 uninhabited islets; they have a collective land area of about one square mile, and the highest elevation is 200 feet. The reef is broken by passes that give boats access to the main island, and the area between the reef and the island is a sheltered fishing ground.

Futuna lies roughly 125 miles (200 kilometres) to the southwest, and it has a smaller neighbour, Alofi. Both are volcanic islands. Futuna has an area of about 25 square miles, and volcanic peaks reach a height of 2,493 feet. Alofi's land area is only 11 square miles, and its highest elevation is about 1,198 feet. Futuna and Alofi are separated by a channel of about two miles in width, and both islands are partially sheltered by fringing reefs.

Both Uvea and Futuna islands have good water supplies, as there are numerous streams and wells. Alofi is without fresh water and has no permanent settlements.

The soils of Uvea and Futuna islands are volcanic but of limited fertility, and a number of factors restrict agricultural production. Only about one-fourth of the available land is suitable for cultivation. The system of land tenure has resulted in a fragmentation of land parcels. Traditional agricultural practices alternate two or three years of cropping with long fallow periods, and there has been a cumulative deterioration of the soil.

The territory's climate is tropical with high humidity. Two seasons are fairly well delineated—a hot, rainy period from November to April during which tropical storms may occur, and a dry, cooler season with trade winds from the southeast during the rest of the year.

Basic subsistence crops have been introduced: coconuts, breadfruit, bananas, taro, cassava, yams, mangoes, and pineapples. The coconut groves on Uvea were destroyed by the rhinoceros beetle in the 1930s, but the pest did not reach Futuna and Alofi. Substantial stands of undisturbed natural forest remain only on Alofi and a small section of Uvea. Otherwise, forests have virtually disappeared, and erosion is a problem on Futuna. Reforestation schemes have been implemented since the early 1960s.

All of the territory's land mammals have been intro-

duced. Pigs are of ceremonial importance and highly valued. There are also some cattle, but the small islands lack suitable pasture land. Chickens constitute a minor part of the diet, and dogs and cats are kept as pets. Fish and other marine fauna provide the bulk of dietary protein.

Villages are dispersed on the islands, mainly on the coast. There are no real urban areas. The international airport is located at Hihifo on northern Uvea.

Languages

The people. The indigenous people are Polynesians, but there are significant differences between the languages and peoples of Uvea and of Futuna islands. It appears that Uvea was settled about 3,000 years ago, and the people were subsequently conquered by Tongans. As a consequence, cultural, historical, and linguistic ties link Uvea and the Tongan archipelago. In contrast, local folklore suggests that Futunans had their origins in Samoa; their language is related to that of Samoa, and the Futunans share many cultural traits with Samoa. While the people of the islands have long been governed by a French administration and all profess the Roman Catholic faith, they do not share a common identity. Futuna is reluctantly tied to Uvea, and its people have expressed a desire to have separate territorial status.

The population of Uvea comprises two-thirds of the territory's total, and the rest live on Futuna. The Uvean and Futunan population in Nouméa, New Caledonia, numbers slightly more than those who live on the home islands. The majority in New Caledonia are Uveans. Opportunities for employment abroad and limited options combined with population pressures at home have caused the exodus. The expatriates are an asset as long as they remain abroad, because their remittances are a major source of capital.

The number of French expatriates in Wallis and Futuna has always been small, and only a few hundred residents are of European descent. Nonetheless, French is the official language of administration.

The economy. Most of the people of Wallis and Futuna are subsistence agriculturalists. The notion of selling produce from the land is contrary to traditional custom, wherein items are bartered and not sold. Similarly, fish are caught mainly for the immediate needs of family, near kin, and neighbours. Most fishing takes place in sheltered areas within the fringing reef. Wallis and Futuna is resource-poor, and little revenue is earned from exports. A small amount of trochus shell is exported to New Caledonia.

The French administration is located on Uvea, and that island has the better infrastructure. Roads and public services are superior, and most households have running water and electricity. The administration is the island's largest employer. In contrast, Futuna is somewhat isolated, without roads or cars, and administrative positions are scarce.

The international airport on Uvea is linked by jet service to Fiji and New Caledonia, and a small plane operates between Uvea and Futuna islands. A cargo vessel occasionally travels between the islands and Nouméa.

Administration and social conditions. In late 1959, by an overwhelming majority, the people of Wallis and Futuna voted to change their political status from that of a protectorate to that of an overseas territory of the French Republic. Legislation passed two years later incorporated traditional powers into the new administrative arrangement.

The territory is divided into three districts that correspond to three traditional political divisions, or kingdoms (more accurately, paramount chieftaincies)—one that includes all of Uvea and two on Futuna. A French-appointed high administrator (*administrateur supérieur*) is the chief executive officer of the territory and serves as the president of the Territorial Council. The council includes the three kings and three others appointed by the president with the approval of the Territorial Assembly. The council decides on matters of general policy.

The Territorial Assembly has legislative power and consists of 20 members (13 from Uvea and seven from Futuna) elected for five-year terms by universal suffrage. The high administrator has fairly broad veto powers over actions of the Assembly. As an integral part of France, the territory is represented in and elects a deputy and a senator to the French national parliament.

Schools

Before the islands became an overseas territory, schools in Wallis and Futuna were operated tuition-free by the Roman Catholic mission. In 1961 the education system came under the French Ministry of National Education. Today, elementary and lower-secondary schools remain free and available in the territory; higher levels of education must be pursued in New Caledonia and metropolitan France. Free health services are also available at dispensaries on both Uvea and Futuna islands and at a hospital on Uvea.

HISTORY

The Dutch explorers Jakob Le Maire and Willem Corneliszoon Schouten sighted Futuna in 1616 during the early voyages of exploration in the Pacific. A lapse of 151 years occurred before Captain Samuel Wallis discovered Uvea in 1767. There was another lapse of more than 50 years before the whaling industry reached the area, in the 1820s, and Europeans began to make regular calls at both islands.

French Marist priests arrived as missionaries in the 1830s. They achieved considerable success within a decade and have remained an important force in the politics of the island group. Protestants never mounted a serious challenge, and the inhabitants of the islands were spared the religious conflicts that were common elsewhere in the Pacific.

As early as the 1840s, the islanders petitioned for French protection, but France was slow to respond; Wallis became a French protectorate in 1887, and Futuna in the following year. Over the next five decades, the French administration became well-entrenched and ruled the colony with a relatively firm hand.

World War II

In 1942, in order to counter the Japanese advances into the American central Pacific, the United States based 6,000 troops on Uvea, and within a short time they built a system of roads, two landing strips, and anchoring facilities in the lagoon. These developments still form the basis of the island's infrastructure.

A permanent change in the colony's status occurred in 1959 when the islanders elected to become an overseas territory of France. In subsequent elections, the people of Wallis and Futuna have demonstrated conservatism, opposing Socialist governments in France and any proposal for a change in their dependent status. (R.C.Ki.)

BIBLIOGRAPHY

Physical and human geography. Overviews of the region are provided by NORMAN DOUGLAS and NGAIRE DOUGLAS (eds.), *Pacific Islands Yearbook*, 17th ed. (1994); and FREDERICA M. BUNGE and MELINDA W. COOKE (eds.), *Oceania, a Regional Study*, 2nd ed. (1985). Large-scale maps are found in *Atlas of the South Pacific*, 2nd ed. (1986), published by the New Zealand government. The classic demographic study is NORMA MCARTHUR, *Island Populations of the Pacific* (1967, reprinted 1983). DOUGLAS L. OLIVER, *Oceania: The Native Cultures of Australia and the Pacific Islands*, 2 vol. (1989), surveys aspects of Pacific Islands societies. Aspects of Oceanic music and dance are surveyed in SELMA JEANNE COHEN (ed.), *International Encyclopedia of Dance* (1998); and ADRIENNE L. KAEPPLER and J.W. LOVE (eds.), *The Garland Encyclopedia of World Music*, vol. 9 (1998). Current writings about the area are found in *Pacific Islands Business* (monthly, Fiji); *Pacific Magazine* (bimonthly, Hawaii); and *Pacific Islands Monthly* (Fiji).

History. The best general history of the precolonial Pacific is K.R. HOWE, *Where the Waves Fall: A New South Sea Island History from First Settlement to Colonial Rule* (1984). For early periods, see JOHN TERRELL, *Prehistory in the Pacific Islands: A Study of Variation in Language, Customs, and Human Biology* (1986); PATRICK VINTON KIRCH (ed.), *Island Societies: Archaeological Approaches to Evolution and Transformation* (1986); PETER BELLWOOD, *The Polynesians: Prehistory of an Island People*, rev. ed. (1987); and JANET DAVIDSON, *The Prehistory of New Zealand* (1984). For current writings, see *Journal of Pacific History* (semiannual, Australia).

The development of Pacific trade and the establishment of exploitative industries are discussed in DOROTHY SHINEBERG, *They Came for Sandalwood* (1967), the first full-length study of trade in the Pacific; O.H.K. SPATE, *Paradise Found and Lost* (1989); and MASLYN WILLIAMS and BARRIE MACDONALD, *The Phosphateers: A History of the British Phosphate Commissioners and the Christmas Island Phosphate Commission* (1985).

Politics is studied in PETER LARMOUR and ROPATE QALO

(eds.), *Decentralization in the South Pacific: Local, Provincial, and State Government in Twenty Countries* (1985); UENTABO FAKAOFO NEEMIA, *Cooperation and Conflict: Costs, Benefits, and National Interests in Pacific Regional Cooperation* (1986), discussing national self-realization; LEONARD MASON and PATRICIA HERENIKO (eds.), *In Search of a Home* (1987), reports by island authors on problems of immigrants in urban centres; HARUHIRO FUKUI (ed.), *Political Parties of Asia and the Pacific*, 2 vol. (1985), an encyclopaedic source; FRANK P. KING (ed.), *Oceania and Beyond: Essays on the Pacific Since 1945* (1976); and DIRK ANTHONY BALLENDORF and FRANK P. KING (eds.), *Oceania Today: Towards New Directions and Political Self-Actualization* (1980). See also STEWART FIRTH, *Nuclear Playground* (1987).

Traditional cultures. ANDREW P. VAYDA (ed.), *Peoples and Cultures of the Pacific: An Anthropological Reader* (1968); ROBERT KIRK and EMÖKE SZATHMARY (eds.), *Out of Asia: Peopling the Americas and the Pacific* (1985), which explores the demography and language; and LYNDALL RYAN, *The Aboriginal Tasmanians* (1981). Cultural life, beliefs, and customs are examined in LENORE MANDERSON (ed.), *Shared Wealth and Symbol: Food, Culture, and Society in Oceania and Southeast Asia* (1986). Other works include WILLIAM A. FOLEY, *The Papuan Languages of New Guinea* (1986), a treatment of the most complex linguistic area of the world; EMILIANA AFEAKI, RON CROCOMBE, and JOHN MCCLAREN (eds.), *Religious Cooperation in the Pacific Islands* (1983); A.C. HADDON and JAMES HORNELL, *Canoes of Oceania* (1975); and BEN R. FINNEY (comp.), *Pacific Navigation and Voyaging* (1976). The arts are surveyed in SIDNEY M. MEAD (ed.), *Exploring the Visual Art of Oceania: Australia, Melanesia, Micronesia, and Polynesia* (1979); ADRIENNE L. KAEPPLER, *Artificial Curiosities: Being an Exposition of Native Manufactures Collected on the Three Pacific Voyages of Captain James Cook, R.N.* (1978); and ROLAND W. FORCE and MARYANNE FORCE, *The Fuller Collection of Pacific Artifacts* (1971). MILES M. JACKSON (ed.), *Pacific Island Studies: A Survey of the Literature* (1986).

Melanesian cultures. H.C. BROOKFIELD, *Melanesia: A Geographical Interpretation of an Island World* (1971), is a comparative survey of geography. R.J. MAY and HANK NELSON (eds.), *Melanesia, Beyond Diversity*, 2 vol. (1982), presents general information. An excellent summary is ANN CHOWNING, *An Introduction to the Peoples and Cultures of Melanesia*, 2nd ed. (1977). On material culture, B.A.L. CRANSTONE, *Melanesia: A Short Ethnography* (1961), remains useful. A fine study of a particular people is MARIE REAY, *The Kuma: Freedom and Conformity in the New Guinea Highlands* (1959). Syntheses and collections of papers on aspects of society and culture in New Guinea include PAULA BROWN, *Highland Peoples of New Guinea* (1978); PAULA BROWN and GEORGEDA BUCHBINDER (eds.), *Man and Woman in the New Guinea Highlands* (1976); ANDREW STRATHERN (ed.), *Inequality in New Guinea Highland Societies* (1982); JERRY W. LEACH and EDMUND LEACH (eds.), *The Kula: New Perspectives on Massim Exchange* (1983); R.M. GLASSE and M.J. MEGGITT (eds.), *Pigs, Pearlshells, and Women: Marriage in the New Guinea Highlands* (1969); GILBERT H. HERDT (ed.), *Rituals of Manhood: Male Initiation in Papua New Guinea* (1982), and *Ritualized Homosexuality in Melanesia* (1984); DAN JORGENSEN (ed.), *Concepts of Conception: Procreation Ideologies in Papua New Guinea* (1983); PAULA G. RUBEL and ABRAHAM ROSMAN, *Your Own Pigs You May Not Eat: A Comparative Study of New Guinea Societies* (1978); and C.A. GREGORY, *Gifts and Commodities* (1982). Other works include MICHELE STEPHEN (ed.), *Sorcerer and Witch in Melanesia* (1987); L.L. LANGNESS and JOHN C. WESCHLER (eds.), *Melanesia: Readings on a Culture Area* (1971); MARILYN STRATHERN (ed.), *Dealing with Inequality: Analysing Gender Relations in Melanesia and Beyond* (1987); P. LAWRENCE and M.J. MEGGITT (eds.), *Gods, Ghosts, and Men in Melanesia: Some Religions of Australian New Guinea and the New Hebrides* (1965); and PETER WORSLEY, *The Trumpet Shall Sound: A Study of "Cargo" Cults in Melanesia*, 2nd ed. (1968).

Micronesian cultures. A general overview is offered in WILLIAM H. ALKIRE, *An Introduction to the Peoples and Cultures of Micronesia*, 2nd ed. (1977); JOHN L. FISCHER, *The Eastern Carolines* (1957, reprinted 1966), is a regional study. Special topics are discussed in MARK R. PEATTIE, *Man'yō: The Rise and Fall of the Japanese in Micronesia, 1885–1945* (1988); DONALD F. MCHENRY, *Micronesia, Trust Betrayed: Altruism vs. Self Interest in American Foreign Policy* (1975); and DAVID NEVIN, *The American Touch in Micronesia* (1977). A good island history is DAVID HANLON, *Upon a Stone Altar: A History of the Island of Pohnpei to 1890* (1988). For further study, see MAC MARSHALL and JAMES D. NASON, *Micronesia, 1944–1974: A Bibliography of Anthropological and Related Source Materials* (1975).

Polynesian cultures. The following works treat the development of Polynesian cultures: IRVING GOLDMAN, *Ancient Polynesian Society* (1970); JESSE D. JENNINGS (ed.), *The Prehistory of Polynesia* (1979); ROBERT BOROFSKY, *Making History: Pukapukan and Anthropological Construction of Knowledge* (1987); PATRICK VINTON KIRCH, *The Evolution of the Polynesian Chiefdoms* (1984); and ANTONY HOOPER and JUDITH HUNTSMAN (eds.), *Transformations of Polynesian Culture* (1985).

Major political units of Melanesia. *Fiji:* On geography and demography, see R. GERARD WARD, *Land Use and Population in Fiji* (1965). Socioeconomic aspects are studied in CYRIL S. BELSHAW, *Under the Ivi Tree: Society and Economic Growth in Rural Fiji* (1964); H.C. BROOKFIELD, F. ELLIS, and R.G. WARD, *Land, Cane, and Coconuts* (1985); and R.F. WATTERS, *Koro: Economic Development and Social Change in Fiji* (1969). Traditional Fijian society and 20th-century change are discussed in R.R. NAYACAKALOU, *Leadership in Fiji* (1975); JOHN NATION, *Customs of Respect: The Traditional Basis of Fijian Communal Politics* (1978); and G.K. ROTH, *Fijian Way of Life*, 2nd ed. (1973). R.A. DERRICK, *History of Fiji* (1946), is a pioneering study and an excellent introduction. DERYCK SCARR, *Fiji, a Short History* (1984), gives a general account. PETER FRANCE, *The Charter of the Land: Custom and Colonization in Fiji* (1969), deals with the foundations of colonial policy. BRUCE KNAPMAN, *Fiji's Economic History, 1874–1939: Studies of Capitalist Colonial Development* (1987), examines the colonial economy. On the history of Indians in Fiji, see K.L. GILLION, *Fiji's Indian Migrants: A History to the End of Indenture in 1920* (1962), and *The Fiji Indians: Challenge to European Dominance, 1920–1946* (1977). See also ADRIAN C. MAYER, *Peasants in the Pacific: A Study of Fiji Indian Rural Society*, 2nd ed. (1973). The political situation is outlined in BRIJ V. LAL (ed.), *Politics in Fiji: Studies in Contemporary History* (1986); and the coup is discussed in BRIJ LAL, *Power and Prejudice: The Making of the Fiji Crisis* (1988).

New Caledonia: OFFICE DE LA RECHERCHE SCIENTIFIQUE ET TECHNIQUE OUTRE-MER, *Atlas de la Nouvelle-Calédonie et dépendances* (1981), gives an account of the territory's development, with detailed maps. *Comptes économiques* (annual), and INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES. FRANCE, *Résultats du recensement de la population de la Nouvelle Calédonie* (1984), are useful official publications. Political and social affairs are the subject of VIRGINIA THOMPSON and RICHARD ADLOFF, *The French Pacific Islands: French Polynesia and New Caledonia* (1971); MYRIAM DORNOY, *Politics in New Caledonia* (1984); and ALAN WARD, *Land and Politics in New Caledonia* (1982). Sociological studies include J.M. KOHLER, *Colonie ou Democratie* (1987); J.M. KOHLER and P. PILLON, *Impact de l'opération café en milieu Mélanésien* (1982); J.M. KOHLER and LOIC J.D. WACQUANT, *L'École inégale: éléments pour un de sociologie de l'école en Nouvelle-Calédonie* (1985); and K.R. HOWE, *The Loyalty Islands: A History of Culture Contacts, 1840–1900* (1977). ALAIN SAUSSOL, *L'Héritage: essai sur le problème foncier mélanésien en Nouvelle-Calédonie* (1979); and PIERRE GASCHER, *La Belle au bois dormant: regards sur l'administration coloniale en Nouvelle-Calédonie de 1874 à 1894* (1974), provide accounts of French settlement. The most comprehensive history is JOHN CONNELL, *New Caledonia or Kanaky?: The Political History of a French Colony* (1987).

Papua New Guinea: The best general work is still the book by H.C. Brookfield, cited at the beginning of the *Melanesian cultures* section. Many aspects of European exploration are included in JAMES SINCLAIR, *Kiap: Australia's Patrol Officers in Papua New Guinea* (1981). Essential historical coverage is provided in J.L. WHITTAKER *et al.* (eds.), *Documents and Readings in New Guinea History: Pre-History to 1889* (1975). J.A. BALLARD (ed.), *Policy-Making in a New State: Papua New Guinea, 1972–1977* (1981), examines political decision making; and the Melanesians' view of change is discussed in GERNOT FUGMANN (ed.), *Ethics and Development in Papua New Guinea* (1986). Among the best anthropological works are LAWRENCE S. GROSSMAN, *Peasants, Subsistence Ecology, and Development of the Highlands of Papua New Guinea* (1984); ROY A. RAPPAPORT, *Pigs for the Ancestors: Ritual in the Ecology of a New Guinea People*, new ed. (1984); and MARILYN STRATHERN, *Women in Between: Female Roles in a Male World: Mount Hagen, New Guinea* (1972). FRAISER MCCONNELL, *Papua New Guinea* (1988), is an informative bibliography.

Solomon Islands: PATRICK VINTON KIRCH and D.E. YEN, *Tikopia: The Prehistory and Ecology of a Polynesian Outlier* (1982); JUDITH A. BENNETT, *Wealth of the Solomons: A History of a Pacific Archipelago, 1800–1978* (1987); DAVID HILLIARD, *God's Gentlemen: A History of the Melanesian Mission, 1849–1942* (1978); ROGER M. KEESING and PETER CORRIS, *Lightning Meets the West Wind: The Malaita Massacre* (1980); HUGH LARACY, *Marists and Melanesians: A History of Catholic Missions in the Solomon Islands* (1976); HUGH LARACY (ed.), *Pacific Protest: The Maasina Rule Movement, Solomon Islands, 1944–1952* (1983); and DEBORAH WAITE, *Art of the Solomon Islands:*

From the Collection of the Barbier-Müller Museum (1983). For further study, see SALLY EDRIDGE (comp.), *Solomon Islands Bibliography to 1980* (1985).

Vanuatu: The standard bibliography is PATRICK O'REILLY, *Bibliographie méthodique, analytique et critique des Nouvelles-Hébrides* (1958); his *Hébridais: répertoire bio-bibliographique des Nouvelles-Hébrides* (1957), provides short sketches of some of the persons (principally European) in Vanuatu's colonial past. See also JOSÉ GARANGER, *Archaeology of the New Hebrides: Contribution to the Knowledge of the Central Islands* (1982; originally published in French, 1972), an account of the excavation of Roy Mata's burial site; MICHAEL ALLEN (ed.), *Vanuatu: Politics, Economics, and Ritual in Island Melanesia* (1981), comparative ethnographic studies; UNIVERSITY OF THE SOUTH PACIFIC. INSTITUTE OF PACIFIC STUDIES, *Vanuatu* (1980), with contributions by leading ni-Vanuatu figures; JEREMY MACCLANCY, *To Kill a Bird with Two Stones: A Short History of Vanuatu* (1981), a survey; RON ADAMS, *In the Land of Strangers: A Century of European Contact with Tanna, 1774–1874* (1984), an account of the first hundred years of European contact; and JEAN GUIART, *Un Siècle et demi de contacts culturels à Tanna, Nouvelles-Hébrides* (1956), on the origins and development of the John Frum cargo cult.

Major political units of Micronesia. *Guam:* LAURA THOMPSON, *Guam and Its People*, 3rd rev. ed. (1947, reprinted 1969), remains a valuable source. An overview of economics, politics, and industry is given in WILLIAM LUTZ, *Guam* (1987). For history, see PAUL CARANO and PEDRO C. SANCHEZ, *A Complete History of Guam* (1964); TIMOTHY P. MAGA, *Defending Paradise: The United States and Guam, 1898–1950* (1988); and ROBERT F. ROGERS, *Guam's Commonwealth Effort 1987–1988* (1988). Guam in World War II is discussed in TONY PALOMO, *An Island in Agony* (1984).

Kiribati: TONY WHINCUP, *Nareau's Nation: A Portrait of the Gilbert Islands* (1979), offers insights into modern life. ROSEMARY GRIMBLE (ed.), *Migrations, Myth, and Magic from the Gilbert Islands: Early Writings of Sir Arthur Grimble* (1972), gives an account of myth, legends, and traditions. See also GERD KOCH, *The Material Culture of Kiribati* (1986; originally published in German, 1965); ALAIMA TALU *et al.*, *Kiribati: Aspects of History* (1979); ALBERT F. ELLIS, *Ocean Island and Nauru: Their Story* (1935); and BARRIE MACDONALD, *Cinderellas of the Empire: Towards a History of Kiribati and Tuvalu* (1982), is comprehensive.

Marshall Islands: ALEXANDER SPOEHR, *Majuro: A Village in the Marshall Islands* (1949, reprinted 1973); and ROBERT C. KISTE, *The Bikinians: A Study in Forced Migration* (1974), describe the ethnology, anthropology, and social conditions.

Micronesia: For a geographic survey, see KENNETH BROWER, *Micronesia, the Land, the People, and the Sea* (1981); and WILLIAM H. STEWART, *Explorer's Atlas of Tourist Maps for the Federated States of Micronesia, Pacific Ocean: Kosrae, Ponape, Truk, and Yap* (1982). Other sources include STEVEN C. SMITH, *The Federated States of Micronesia: An Emerging Nation: An Overview for Peace Corps Volunteers* (1986); and AUSTIN RANNEY and HOWARD R. PENNIMAN, *Democracy in the Islands: The Micronesian Plebiscites of 1983* (1985).

Nauru: General surveys are found in the titles cited in the beginning section of this bibliography. See also RON CROCOMBE and AHMED ALI (eds.), *Politics in Micronesia* (1983); BARRIE MACDONALD, *In Pursuit of the Sacred Trust: Trusteeship and Independence in Nauru* (1988); and the work by M. Williams and B. Macdonald cited in the *History* section.

Northern Mariana Islands: For information on the Northern Mariana Islands, see the works by Francis X. Hezel, Frank P. King, and Dirk Anthony Ballendorf, cited above under *History*, and the works by Mark R. Peattie, Donald F. McHenry, and David Nevin, under *Micronesian cultures.*

Palau: DOUGLAS FAULKNER, *This Living Reef* (1974), is an illustrated survey of the marine environment. See also DOUGLAS OSBORNE, *The Archeology of the Palau Islands: An Intensive Survey* (1966); and GEORGE J. GUMERMAN, DAVID SNYDER, and W. BRUCE MASSE, *An Archaeological Reconnaissance in the Palau Archipelago, Western Caroline Islands, Micronesia* (1981). R.E. JOHANNES, *Words of the Lagoon: Fishing and Marine Lore in the Palau District of Micronesia* (1981); ROLAND W. FORCE and MARYANNE FORCE, *Just One House: A Description and Analysis of Kinship in the Palau Islands* (1972); and H.G. BARNETT, *Being a Palauan* (1959, reissued 1979), are anthropological accounts. DANIEL J. PEACOCK, *Lee Boo of Belau: A Prince in London* (1987), is the story of the first Palauan to visit the West. EDWARD C. BARNARD, *Naked and a Prisoner: Captain Edward C. Barnard's Narrative of Shipwreck in Palau, 1832–1833*, ed. by KENNETH R. MARTIN (1980), is a personal account of a stay in Palau.

Tuvalu: AUSTRALIAN MUSEUM. SYDNEY, *The Atoll of Funafuti, Ellice Group: Its Zoology, Botany, Ethnology, and General Structure Based on Collections Made by Mr. Charles Hedley*, 10 pt. (1896–1900), is a broad scientific survey. See also ANNE CHAMBERS, *Nanumea Report: A Socio-Economic Study of Nanumea Atoll, Tuvalu* (1975, reissued as *Nanumea*, 1984); and DONALD GILBERT KENNEDY, *Field Notes on the Culture of Vaitupu, Ellice Islands* (1931). HUGHT LARAY (ed.), *Tuvalu: A History* (1983), provides an introduction. See also the work by Barrie Macdonald cited above under *Kiribati.*

Major political units of Polynesia. *American Samoa and Western Samoa:* LOWELL D. HOLMES, *Samoan Village* (1974); MARGARET MEAD, *Coming of Age in Samoa: A Psychological Study of Primitive Youth for Western Civilization* (1928), available also in later editions; and DEREK FREEMAN, *Margaret Mead and Samoa: The Making and Unmaking of an Anthropological Myth* (1983), explore the traditional heritage and its interpretations. See also PAUL T. BAKER, JOEL M. HANNA, and THELMA S. BAKER (eds.), *The Changing Samoans: Behavior and Health in Transition* (1986); GEORGE TURNER, *Samoa, a Hundred Years Ago and Long Before* (1884, reprinted 1984); J.W. DAVIDSON, *Samoa mo Samoa: The Emergence of the Independent State of Western Samoa* (1967); and R.P. GILSON, *Samoa 1830 to 1900: The Politics of a Multi-Cultural Community* (1970). A modern survey is offered in FRED HENRY, *Samoa, an Early History* (1980). For further research, see LOWELL D. HOLMES (comp.), *Samoan Islands Bibliography* (1984). The best sources for a discussion of the political development of Western Samoa include MALAMA MELEISEA and PENELOPE SCHOEFFEL MELEISEA (ed.), *Lagaga: A Short History of Western Samoa* (1987); and MALAMA MELEISEA, *The Making of Modern Samoa* (1987).

Cook Islands: NORMAN DOUGLAS and NGAIRE DOUGLAS, *Cook Islands, a Guide* (1987), provides a general survey of the whole country; NGATUPUNA KAUTAI *et al.*, *Atiu: An Island Community* (1984), describes a particular island in greater depth. RICHARD GILSON, *The Cook Islands, 1820–1950* (1980), is a historical survey; and RON CROCOMBE (ed.), *Cook Islands Politics: The Inside Story* (1979), discusses political trends.

Easter Island: A comprehensive introduction is provided in NORWEGIAN ARCHAEOLOGICAL EXPEDITION TO EASTER ISLAND AND THE EAST PACIFIC. 1955–1956, *Reports*, 2 vol. (1961–65). History and culture are outlined in GRANT MCCALL, *Rapanui: Tradition and Survival on Easter Island* (1981). See also THOR HEYERDAHL, *Aku-Aku: The Secret of Easter Island*, trans. from Norwegian (1958, reprinted 1988), *Sea Routes to Polynesia* (1968), *Art of Easter Island* (1975), a survey of art in collections throughout the world, with more than 300 plates of photographs, and *Easter Island—a Mystery Solved* (1989).

French Polynesia: BENGT DANIELSSON, *Work and Life on Raroia: An Acculturation Study from the Tuamotu Group, French Oceania* (1956); and F. ALLAN HANSON, *Rapan Lifeways: Society and History on a Polynesian Island* (1970, reprinted 1983), study society and culture. See also DOUGLAS L. OLIVER, *Ancient Tahitian Society*, 3 vol. (1974), a classic treatment. For history, see ROBERT LANGDON, *Tahiti, Island of Love*, 5th ed. (1979); COLIN NEWBURY, *Tahiti Nui: Change and Survival in French Polynesia, 1767–1945* (1980); and PIERRE-YVES TOULELAN, *Tahiti Colonial* (1984). Other studies include WILLIAM TAGUPA, *Politics in French Polynesia, 1945–1975* (1976); VIRGINIA THOMPSON and RICHARD ADLOFF, *The French Pacific Islands* (1971); and PAUL HODÉE, *Tahiti, 1834–1984: 150 ans de vie chrétienne en eglise* (1983).

Tonga: JOHN CONNELL, *Migration, Employment, and Development in the South Pacific: Country Report* (1983), vol. 18, *Tonga*, studies the effects of internal and external migration. ELIZABETH BOTT, *Tongan Society at the Time of Captain Cook's Visits: Discussions with Her Majesty Queen Sālote Tupou* (1982), is a standard historical reference. SIONE LĀTŪKEFU, *Church and State in Tonga: The Wesleyan Methodist Missionaries and Political Development, 1822–1875* (1974), is a detailed history. A.H. WOOD, *A History and Geography of Tonga* (1943), is a brief but excellent introduction; and NOEL RUTHERFORD, *Friendly Islands: A History of Tonga* (1977), provides comprehensive coverage of life in Tonga.

Wallis and Futuna: Information on various topics can be found in general sources cited in the beginning section of this bibliography, as well as in such periodicals as *The Pacific Guide* (annual), published in England, mostly for business travelers in the area; and *The Contemporary Pacific* (semiannual), which began publication in 1989, from the University of Hawaii Press; as well as in relevant chapters of such surveys as OTIS W. FREEMAN, *Geography of the Pacific* (1951); and EMILIANA AFEAKI *et al.*, *Politics in Polynesia* (1983).

Western Samoa: For books on the geography and history of Western Samoa, see *American Samoa and Western Samoa* above.

(L.E.M./F.J.W./R.W.Fo./R.M.Ke./R.C.Ki./
B.K.M./D.L.Sh./R.T.J./H.M.La./R.Ad./D.A.B./
F.X.H./D.R.Sh./A.We./R.G.C./M.T.C./Th.H./Si.L.)

The Pacific Ocean

Of the three oceans that extend northward from the Antarctic continent, the Pacific, occupying about a third of the surface of the terrestrial globe, is by far the largest. Its area, excluding adjacent seas, encompasses about 63,800,000 square miles (165,250,000 square kilometres). It has double the area and more than double the water volume of the Atlantic—the next largest division of the hydrosphere—and its area more than exceeds that of the whole land surface of the globe. The Pacific Ocean stretches from the shores of Antarctica to the Bering Strait through 135° of latitude, or for 9,600 miles (15,500 kilometres). Its greatest longitudinal extent measures some 12,000 miles along the parallel of 5° N, between the coasts of Colombia in South America and the Malay Peninsula in Asia. The mean depth of the Pacific (excluding adjacent seas) is 14,040 feet (4,280 metres), and its greatest known depth is 36,201 feet (11,034 metres)—in the Mariana Trench—also the greatest depth found in any ocean.

The Pacific and Arctic systems mingle their waters in the Northern Hemisphere at the shallow Bering Strait, and in the Southern Hemisphere the Pacific and Atlantic mix in the relatively narrow Drake Passage between Tierra del Fuego in South America and Graham Land in Antarctica. The separation between the Pacific and Indian oceans is less distinct, but generally it is considered to lie along the line of islands extending eastward from Sumatra, through Java to Timor, thence across the Timor Sea to Cape Londonderry in Australia. To the south of Australia the boundary extends across the Bass Strait and thence from Tasmania to Antarctica.

Because of the pattern of major mountain systems of the globe, a relatively small proportion (one-seventh) of the total continental drainage enters the Pacific—*i.e.*, a total drainage area of less than about three times the total area of Australia. Of the rivers that drain into the Pacific, those of China and Southeast Asia are of the greatest importance; the basins of these rivers support more than one-quarter of the world's population.

The eastern boundary of the Pacific is associated with the American Cordilleran mountain system, which stretches from Alaska in the north to Tierra del Fuego in the south. Except for its extreme northern and southern sections, which are characterized by fjords and their numerous offlying islands, and except for the deeply indented Gulf of California, the coastal boundary is relatively regular and the continental shelf narrow. The western, or Asiatic, coastal boundary, in contrast, is irregular. Although the mountain systems there lie roughly parallel to the coast, as they do on the eastern Pacific coastlands, the western Pacific is noted for its many marginal seas. From north to south they include the Bering Sea, the Sea of Okhotsk, the Sea of Japan, the Yellow Sea, the East China Sea, and the South China Sea. Their eastern boundaries are formed by southward-jutting peninsulas or island arcs or both. It is of oceanographic significance that the great rivers of eastern Asia—including the Amur, the Huang Ho (Yellow River), the Yangtze, the Hsi and Pearl, and the Mekong—enter the Pacific indirectly by way of the marginal seas.

This article treats the physical and human geography of the Pacific Ocean, followed by discussion of the six major marginal seas of the Pacific. For discussion of the physical and chemical oceanography and marine geology of the Pacific, see OCEANS; for discussion of Pacific peoples and cultures, see PACIFIC ISLANDS. For further references, see the *Index*.

The article is divided into the following sections:

Physical and human geography 302
- Physiography 302
  - Relief
  - Principal ridges and basins
  - Bottom deposits
  - Islands
- Geology 306
- Climate 306
  - The trade winds
  - Tropical storms
  - The westerlies
  - The monsoon regime
- Temperature and salinity 306
  - Temperature
  - Salinity
- Hydrology 307
  - Surface currents
  - Deepwater circulation
  - Tides
- Economic aspects 308
  - Biological resources
  - Fisheries
  - Mineral resources
  - Trade and transportation
  - Environmental impact of human activity
- Study and exploration 309
  - Early exploration and settlement
  - European exploration
  - Exploration in the 20th century

The major marginal seas of the Pacific Ocean 310
- Bering Sea and Strait 310
- East China Sea 312
- Sea of Japan 313
- Sea of Okhotsk 314
- South China Sea 315
- Yellow Sea 316

Bibliography 317

Physical and human geography

PHYSIOGRAPHY

Relief. The Pacific basin may conveniently be divided into three major physiographic regions: the eastern, western, and central Pacific regions.

Eastern region. The eastern region, which extends southward from Alaska to Tierra del Fuego, is relatively narrow and is associated with the American cordilleran system of almost unbroken mountain chains, the coastal ranges of which rise steeply from the western shores of North and South America. The continental shelf, which runs parallel to it, is steep and comparatively narrow. Significant oceanic trenches in this region are the Middle America Trench in the North Pacific and the Peru-Chile Trench in the South Pacific.

Oceanic trenches

Western region. The seaward boundary of the western region is marked by a broken line of oceanic trenches, extending from the Aleutian Trench in the north through the Kuril and Japan trenches and southward to the Tonga and Kermadec trenches, terminating close to the northeast of North Island, New Zealand. Its structure is more complex than that of the eastern region. Characteristically associated with the ocean trenches of the western region are festoons of either peninsulas or islands or both. The islands, which include those of Japan as well as numerous smaller islands, represent the upper parts of mountain systems that rise abruptly from the deep ocean floor. The island clusters of the western Pacific form the boundaries of the several wide and deep continental seas of the region.

Central Pacific region. The central Pacific region lies between the boundaries of the eastern and western re-

gions. The largest and the most geologically stable of the structural provinces of the Earth's crust, it is characterized by expansive areas of low relief, lying at a general depth of about 15,000 feet below the surface.

Principal ridges and basins. To the east of 150° W the relief of the ocean floor is considerably less pronounced than it is to the west. In the eastern Pacific the Cocos Ridge extends southwestward from the Central American isthmus to the Galápagos Islands. To the south of the Galápagos lies the Peru Basin, which is separated by the extensive Sala y Gomez Ridge from the Southeast Pacific Basin, which, in turn, is separated from the Southwest Pacific Basin by the East Pacific Rise and indeterminate Pacific-Antarctic Ridge, which runs from the Sala y Gomez Ridge to Antarctica in the vicinity of 150° W.

Extending southward from the Tasman Basin (between New Zealand and eastern Australia) is the Macquarie Ridge, which forms a major boundary between the deep waters of the Pacific and Indian oceans. The Hawaiian Ridge extends westward from Hawaii to the 180° meridian.

The submerged parts of the series of ridges that are capped by the island archipelagoes of the western Pacific are continuous and are to be found at depths of less than about 2,000 feet. These ridges include the Aleutian ridge in the northwestern Pacific; the series of ridges extending southward through the Kurils, Bonin, the Marianas, Yap, and Palau; those extending eastward from New Guinea, including the Bismarck Archipelago, the Solomons, and Santa Cruz; and finally the ridges extending southward, from which rise the Samoan islands, Tonga, Kermadec, Chatham, and Macquarie.

Bottom deposits. Apart from the narrow coastal zone of the eastern region and the broad continental seas of the western region, the Pacific is floored with pelagic (*i.e.,* oceanic) material derived from the remains of marine plants and animals that once inhabited the waters lying above. Red or brown radiolarian ooze is found along the zone of the Pacific North Equatorial Current, east of 170° W, and on the floors of some deep Indonesian basins. A belt of diatom ooze occurs between 45° and 60° S latitude and across the North Pacific, between Japan and Alaska. Calcareous globigerina ooze occurs in the shallower parts of the South Pacific, the dissolving power of the seawater at great depths being sufficient to dissolve calcareous material to such an extent that these oozes are not generally found at depths in excess of about 15,000 feet. Silica-containing material, such as radiolarian and diatom ooze, is found at greater depths, but even these siliceous remains are dissolved at very great depths, where the characteristic deposit is red clay. Red clay, which covers no less than half of the Pacific floor, is believed to be formed of colloidal (extremely finely divided) clays derived essentially from the land.

Occurrences of ocean-floor ooze

On the abyssal plains, where sediments accumulate slowly, chemical and biological processes lead to the formation of metal-bearing coatings around objects such as the ear bones of fishes. The nodules so formed contain manganese, iron, nickel, copper, cobalt, and traces of other metals such as platinum. They cover large areas of the ocean floor in the Pacific. Similar processes form coatings, called manganese crusts, on the rock surfaces of seamounts.

Among the many different forms of land-derived muds (formed by the erosive action of rivers, tides, and currents) that floor the continental shelves and slopes of the Pacific, the yellow mud of the Yellow Sea is of particular interest. The mud is conveyed to the seabed by the Huang Ho, which drains a vast area of North China blanketed with loess, a fine-grained soil.

Islands. The islands of the western region—including the Aleutians, the Kurils, the Ryukyus, Taiwan, the Philippines, Indonesia, New Guinea, and New Zealand—are continental in character. Geologically they consist, in part, of sedimentary rocks, and their structures are similar to those of the coastal mountain ranges of the adjacent continent.

"High" and "low" islands

A geologically important boundary between the continental, or "high," islands and the numerous truly oceanic, or "low," islands of the Pacific is the Andesite Line. In the northern and western Pacific the Andesite Line follows close to seaward the trend of the island arcs from the Aleutians southward to the Yap and Palau arcs, thence eastward through the Bismarcks, the Solomons, and Santa Cruz, and thence southward through Samoa, Tonga, Chatham, and Macquarie to Antarctica. Within that boundary, islands are essentially of basalt.

The numerous oceanic islands of the Pacific are unevenly distributed. They lie, in the main, between the tropics of Cancer and Capricorn and occur in great numbers in the western Pacific. The northernmost chain of oceanic islands is associated with the Hawaiian Ridge. The Hawaiian archipelago consists of about 2,000 islands, although the term Hawaiian Islands is usually applied to the small group that lies at the eastern end of the archipelago.

The numerous small islands of Micronesia lie mainly north of the equator and to the west of the 180th meridian. Nearly all are coralline; the principal groups are the Marianas, the Marshalls, the Carolines, Kiribati (Gilbert Islands), and Tuvalu (Ellice Islands).

To the south of Micronesia lies Melanesia, which consists mostly of small coral islands. The region's physiography is dominated by a group of large continental islands, how-

Michael V. Propp



Kraternaya Bay, Yankich Island, in the Kuril Islands, Russia.



The Pacific Ocean, with depth contours and submarine features.


150°
120°
90°
Bristol Bay
Kodiak Island
Gulf of Alaska
-3,826
Vancouver Island
CASCADIA BASIN
Vancouver
Seattle/Tacoma
JUAN DE FUCA RIDGE
NORTH AMERICA
San Francisco/ Oakland
Los Angeles
ATLANTIC OCEAN
-6,298
PACIFIC
30°
Gulf of Mexico
Tropic of Cancer
Hawaiian Islands
RIDGE
HAWAIIAN RIDGE
NORTHEAST PACIFIC BASIN
MIDDLE AMERICA TRENCH
Caribbean Sea
Panama Canal
COCOS RIDGE
GUATEMALA BASIN
COLÓN RIDGE
PANAMA BASIN
Equator
0°
Galapagos Islands
CARNEGIE RIDGE
SOUTH AMERICA
PENRHYN BASIN
-5,029
TIKI BASIN
Tahiti
BAUER BASIN
GALAPAGOS RISE
PERU BASIN
PERU CHILE TRENCH
-4,525
NAZCA RIDGE
CHILE BASIN
Tropic of Capricorn
30°
ROGGEVEEN BASIN
SOUTHWEST PACIFIC BASIN
EAST PACIFIC RISE
CHILE RISE
-3,977
-4,876
SOUTHEAST PACIFIC BASIN
Strait of Magellan
PACIFIC-ANTARCTIC RIDGE
Drake Passage
150°
120°
90°

ever, including New Guinea. The principal Melanesian island groups are the Bismarck Archipelago, the Solomons, Vanuatu (New Hebrides), New Caledonia, and Fiji.

The immense area of Polynesia includes the Hawaiian Islands, the Phoenix Islands, Samoa, Tonga, the Cook Islands, the Society Islands, Tuamotu, and the Marquesas.

GEOLOGY

Evidence drawn from various geophysical fields—seismology, volcanology, gravimetry and paleomagnetism—points to the general validity of the theory of plate tectonics (see PLATE TECTONICS).

Formation of island arcs

The Pacific island arcs are believed to have originated in the lateral thrusting of the Eurasian and Australian plates toward and over the Pacific floor, the shapes of the island arcs corresponding to the intersections of the thrust planes with the spherical Earth's surface. The intense folding and faulting along the volcanic zone of the western Pacific provide unmistakable evidence of orogenic (mountain-building) forces at work. The deep basins that lie between the Asian continent and the island arcs were evidently caused by the local downfolding of the Earth's crust; the island arcs themselves, along the line of which volcanic and seismic activity are pronounced, are the result of crustal strata weakened by strong upfolding. On the oceanic side of the island arcs, intense downfolding has resulted in the formation of the deep trenches that fringe the line of the island arcs.

The character of the eastern marginal physiographic region of the Pacific suggests a lateral movement of the North American and South American plates westward over the Pacific floor. This movement causes downfolding of the ocean floor and gives rise to major fault lines (*e.g.*, the San Andreas Fault in California, U.S.) and volcanism in the coastal cordilleran mountains of North, Central, and South America. The floor of the northeastern Pacific is remarkable for its several major fracture zones, which extend east and west and which, in some instances, are identifiable over distances of thousands of miles.

Of great geologic interest are the seamounts (submerged volcanoes), guyots (flat-topped seamounts), and oceanic islands of the Pacific. The numerous tropical islands of the Pacific are mainly coralline. The three principal types of coral reef—fringing, barrier, and atoll, as well as the guyots, which rise within the Pacific from the ocean floor in latitudes north and south of the tropics—are explained partially by the slow subsidence theory advanced by the English naturalist Charles Darwin during the 19th century and partially by the theory of plate tectonics.

CLIMATE

The wind and pressure systems of the Pacific conform closely with the so-called planetary system—*i.e.,* the patterns of air pressure and the consequent wind patterns that develop in the atmosphere of the Earth as a result of its rotation (Coriolis force) and the inclination of its axis (ecliptic) toward the Sun. They are, in essence, a three-celled latitudinal arrangement of the atmospheric circulation in each hemisphere, one arrangement generally representing a mirror image of the other. The vast extent of open water in the Pacific influences wind and pressure patterns over it, and climatic conditions in the southern and eastern Pacific—where the steadiness of the trade winds and the westerlies is remarkable—are the most uniform on the globe. In the North Pacific, however, conditions are not so uniform, particularly the considerable climatic differences between the eastern and western regions in the same latitude. The rigour of the winters off the east coast of Russia, for instance, contrasts sharply with the relative mildness of winters in the region of British Columbia.

Uniformity of climate

The trade winds. The trade winds of the Pacific represent the eastern and equatorial parts of the air circulation system; they originate in the subtropical high-pressure zones that are most pronounced, respectively, over the northeast and southeast Pacific between the 30th and 40th parallels N and S. The obliquity of the ecliptic (an angle of approximately $23\frac{1}{2}°$ that is the difference between the planes of the Earth's rotation on its axis and its revolution around the Sun) limits the seasonal shifting of the Pacific trade-wind belts to about 5° of latitude. The easterly winds along the equator between the two subtropical zones form the intertropical airflow and tend to be strongest in the eastern Pacific. The equatorial region, in which the trade winds of the Northern and Southern hemispheres converge, is the region of calms or light variable breezes and is known as the doldrums.

The trade winds, especially in the eastern Pacific, convey relatively cool air toward the equator; in moving, the air comes in contact with the sea and thus becomes increasingly humid and warm, and high lapse rates (the term used to denote the rate of change of air temperature with increasing height above sea or land surface) result. The average wind speed of the Pacific trade winds is about 13 knots (15 miles per hour). The weather in the trade-wind belts is normally fine, with relatively little cloud cover; such clouds as there are characteristically take the form of broken cumulus (small piles of clouds with flat bases) at about 2,000 feet above sea level. Precipitation, usually in the form of light showers, is slight.

Off the west coasts of the American continent in the trade-wind belts, upwelling of cold, subsurface water causes the overlying air to be cooled below its dew point (the air temperature below which water vapour condenses as dew), with the consequent widespread formation of low, thick clouds. Fog in these regions is not uncommon.

Tropical storms. Although, in general, the climatic conditions of the trade-wind belts are characteristically regular and uniform, storms of great violence do originate there. In such storms, winds of exceptionally great force spiral inward toward a centre of exceedingly low atmospheric pressure. These are the tropical storms known in the western Pacific as typhoons. While the mechanism that triggers typhoons and other tropical storms is not completely understood, ideal conditions for their development occur in the western Pacific between the parallels of 5° and 25° N during late summer and early autumn, when the temperature of surface waters is at least 81° F (27° C). The regions to the east of the Philippines and in the South and East China seas are notorious for these storms, which imperil shipping and often cause severe coastal flooding accompanied by loss of life and property.

Typhoons of the western Pacific

The westerlies. Within the belts of the westerly winds, cold easterly winds from polar regions meet the warm westerly winds of the middle latitudes, causing the formation of the traveling depressions characteristic of middle latitudes. The zone of convergence, or polar front, is most strongly developed in winter when the contrast in temperature and humidity of the air between the converging flows is greatest.

The westerlies in the Southern Hemisphere are steady and strong, and their popular name, "brave west winds," is apt. The gales that accompany the depressions have given rise to the term "roaring forties," named for the latitudinal zone in which the storm winds are of greatest frequency.

The monsoon regime. The western Pacific is subject to a seasonal climatic regime that replaces the planetary system. This is the regime of the monsoon (rain-bearing winds), which is associated with the heating of the Asian landmass in summer and its intense cooling in winter. The heating of air masses over tropical Asia in summer initiates a low-pressure system, which becomes the focal point of the trade winds of both hemispheres. The doldrums, therefore, do not exist in the western Pacific during northern summer because of the immense flow of maritime air into the Asian low-pressure zone. The cooling of the continent in winter results in the development of the Asian high-pressure system, which leads to a strengthening of the trade winds of the Northern Hemisphere.

As a result of seasonal changes in pressure and wind circulation, marked seasonal contrast between continental and maritime influences—the first associated with drought and cold and the second with moisture and heat—is to be found in the whole of the western Pacific from the Sea of Japan southward.

TEMPERATURE AND SALINITY

Temperature. The oceans tend to be stratified, the principal factor being temperature; the bottom waters of

the deep parts are intensely cold, with temperatures only slightly above freezing.

The surface zone, where temperature variations are perceptible, is between 300 and 1,000 feet thick. It is more compressed in the temperate Eastern Pacific, along the coasts of North and Central America, where cold water appears at a shallower depth compared to the central and western Pacific.

Ocean temperatures in the North Pacific tend to be higher than those in the South Pacific because the ratio of land to sea areas is larger in the Northern Hemisphere and because Antarctica also influences water temperature.

The mean position of the thermal equator (the line on the Earth on which the highest average air temperatures are found; the line migrates latitudinally with the changing angular distance from the equator of the Sun) in the Pacific, although it lies in the Northern Hemisphere, is nearer to the geographic equator than in the Atlantic and Indian oceans.

There is a pronounced difference in temperature and salinity between the surface and deep zones of the Pacific. The deep zone, which contains about 80 percent of the ocean's volume, has relatively stable temperature and salinity patterns; its average temperature is 38.3° F (3.5° C).

Salinity. The salinity patterns of the surface waters of the Pacific are influenced largely by wind and by rainfall and evaporation patterns. The waters within the belt of calms and variable winds near the equator have lower salinities than those in the trade-wind belts. In the equatorial belt, relatively large amounts of rain fall and little evaporation occurs both because of low wind speeds and because of the generally cloudy skies; salinity in the equatorial belt runs as low as 34 parts per thousand.

Variation

The highest surface salinities in the open Pacific occur in the southeastern area, where they reach 37 parts per thousand; in the corresponding trade-wind belt in the North Pacific the maximum salinity seldom reaches 36 parts per thousand. Pacific waters near Antarctica have salinities of less than about 34 parts; the lowest salinities—less than about 32 parts—occur in the extreme northern zone of the Pacific.

The heavy rainfall of the western Pacific, associated with the monsoons of the region, gives rise to relatively low salinities. Seasonal variations there, as well as in the eastern Pacific, where seasonal changes in surface currents occur, are significant.

HYDROLOGY

Surface currents. Pacific trade winds drive surface waters toward the west to form the North and South Equatorial currents, the axes of which coincide, respectively, with the parallel of 15° N and the equator. Squeezed between the Equatorial currents is a well-defined countercurrent, the axis of which is always north of the equator and which extends from the Philippines to the shores of Ecuador.

The major part of the North Equatorial Current swings northward in the vicinity of the Philippines to form the warm Kuroshio (also called the Japan Current). To the east of Japan the Kuroshio swings eastward to form the Kuroshio Extension. The branching of this current in the region of 160° E results in the movement known as the North Pacific Current. The surface waters of the Bering Sea circulate in a counterclockwise direction. The southward extension of the Kamchatka Current forms the cold Oya Current, which flows to the east of the Japanese island of Honshu to meet the warm Kuroshio waters in the vicinity of 36° N. The cold, southeast-flowing California Current forms the eastern segment of the returning branch of the North Equatorial Current system.

The main part of the South Equatorial Current divides into three large branches as it flows westward. The two westernmost branches, on reaching the east coast of Australia, swing south to form the East Australian Current, which, becoming the Tasman Current, turns back to the northeast and dissipates west of New Zealand. The easternmost branch flows first to the south, roughly along the 180° meridian, before turning back to the east in the vicinity of 50° S as the warm South Pacific Current; between 80° and 90° W this flow turns northward and then westward as the Mentor Current, the waters eventually returning to the South Equatorial Current. Flowing between Antarctica and the South Pacific Current is the cold Antarctic Circumpolar Current, which constitutes the other portion of the South Pacific circulation system; as it reaches the South American coast in the region of 45° S, one branch flows northward along the coast to form the Peru (Humboldt) Current, and a second branch flows southward to pass through the Drake Passage.

Periodically—usually at intervals of three to four years—an anomaly of ocean currents and climatology called El Niño occurs in the South Pacific. This event is associated with the appearance of unusually warm ocean conditions off the tropical coast of South America and with changes in tropical atmospheric patterns (called the Southern Oscillation) that can adversely affect fishing, agriculture, and rainfall patterns along the west coast of South America. During particularly strong events, El Niño can cause weather anomalies in the equatorial and southern Pacific and in northeastern South America, Asia, and North America.

El Niño

Deepwater circulation. Observations of temperature and salinity at different levels in the ocean reveal well-defined layers, each forming a water mass distinguished by its own temperature and salinity characteristics.

It appears that the most important influence on the vertical circulation of the Pacific is the cold water generated around the Antarctic continent. This dense circumpolar water sinks and then spreads northward to form the bottom layer of the greater part of the Pacific. It is thought that cold, deep water flows northward in the western Pacific in a relatively well-defined current from the vicinity of Antarctica to Japan. Branches from this deep main stream convey cold water eastward and then poleward in both hemispheres.

Deepwater circulation is influenced by the descent of surface water at zones of convergence of neighbouring water flows. In the Pacific Tropical Convergence, which coincides with the Equatorial Countercurrent, water sinks to a depth of about 300 feet before it spreads laterally. The Pacific Subtropical Convergences are located between the parallels of 35° and 40° N and S. Water that sinks at the convergences spreads laterally at increasing depths as the distance from the equator increases. The Antarctic Convergence lies in the zone of the southern westerly winds. A corresponding Arctic Convergence is prominent in the northeastern Pacific.

To compensate for downward-moving water, some water rises at zones of divergence, particularly along the so-called cold-water coasts of both North and South America, where upwelling of cold water is a well-marked phenomenon.

Tides. In contrast to the tides of the Atlantic—which are almost always semi-diurnal (*i.e.,* twice-daily) occurrences—those of the Pacific include many instances of diurnal (daily) and mixed tides. In the diurnal type of tidal oscillation, only a single high water and a single low water occur each tidal day (which lasts for about 24 hours and 50 minutes). Tides of this type occur in the Gulfs of Tonkin and Thailand in Southeast Asia, the Java Sea in Indonesia, and Bismarck and Solomon seas north and east of the island of New Guinea. Mixed tides, in which both diurnal and semidiurnal oscillations appear, are characterized by large inequalities in successive high (or low) water heights. This type of tide is prevalent along much of the Pacific coast.

At certain places in the South Pacific the natural period of oscillation of the sea accentuates the solar tidal oscillation. At these locations the time of the AM (or PM) high (or low) water, instead of getting later each day by about 50 minutes (as is generally the case), occurs at approximately the same time for several days in succession. The tide at Tahiti, for example, follows the Sun and not the Moon—the time of high water occurring, day after day, at about midnight and noon, and that of low water at about 6 AM and 6 PM.

In general, tidal ranges within the Pacific are small. That at Tahiti is about one foot; at Honolulu it is about two feet; at Yokohama, it seldom exceeds five feet; and at Cape

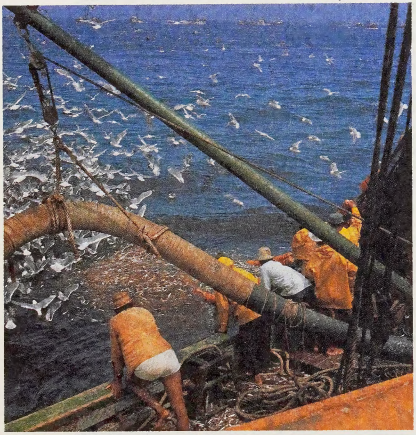
Tidal ranges

Horn it is never more than about six feet. In the upper reaches of the Gulf of California and Korea Bay, however, very large tidal ranges of more than 40 feet are quite common.

ECONOMIC ASPECTS

Biological resources. The Pacific has the most varied complement of plants (*i.e.,* algae) and animals of the world's oceans. The circumglobal mixing of water in the southern and, to a much more limited extent, northern polar reaches of the Pacific permit the intermingling of flora and fauna from other oceanic regions, while temperate and tropical surface waters of the Pacific are more likely to have indigenous biotas. On the rocky, cold-water coasts of North and South America, for example, are found vast forestlike kelp beds made up of brown algae of the genus *Laminaria,* with individual plants often reaching heights of 100 feet or more. They harbour a rich animal complement of invertebrates and fishes approaching a faunal variety that vies with that of tropical rain forests. Where upwelling and other current conditions add nutrients to the offshore surface waters of these same reaches of the Pacific, dense concentrations of plankton-feeding fishes thrive, predominantly those of the herring family and its relatives. Examples include the Japanese sardine and the Peruvian anchovy, both of which are among the largest single-species fishing catches in the world.

Robert Harding Picture Library Ltd.



Fishing for anchovies off the coast of Peru.

In the North Pacific the circulation patterns and runoff from the land create conditions in which demersal, or bottom-living, species abound. The North Pacific hake and the Alaska pollack are prominent examples. Salmon, likewise, thrive in the North Pacific, proliferating there in five species of the genus *Oncorhynchus,* as compared to the single species, *Salmo salar,* of the Atlantic.

In the warm, tropical region—roughly between the North and South Equatorial Current systems—the wealth of marine animals, especially, increases dramatically. The variety of animal life is greater in the western Pacific, where the warm monsoonal climate and variegated landforms have promoted evolution of the unique Indo-Pacific marine forms. The western Pacific also has the richest and most extensive coral reefs of any ocean, with some six times more species of fish associated with them than with the coral reefs of the Caribbean Sea in the Atlantic. The tropical sea passages between the Pacific and the Indian oceans have also given the latter ocean a rich reef fauna; and Indo-Pacific mollusks have reached copious evolutionary diversification, with the giant clam, *Tridacna gigas,* a spectacular example. Another example of the Pacific's richness in species is found among the tunas: six species (one of them endemic) roam the tropical reaches of the Pacific, furnishing more than half of the world's tuna catch.

Whales

Whales are a prominent and spectacular component of the Pacific marine biota. The habits of many species include regular, long-distance migrations from cold-water feeding to warm-water breeding and calving grounds, thus predisposing them to global distribution.

Fisheries. Because of the biological richness of certain reaches of the Pacific and because of the large human populations in many of the nations bordering it, the catches there are substantially larger than those in the other oceans and comprise three-fifths of the total world catch. Among Pacific nations, Japan and Russia have the largest fisheries in the world as measured by tonnage caught, but China, the United States, Peru, Chile, South Korea, and Indonesia are also among the world's major fishing nations. The fishing industries in all of these countries are enormous, and all are based at least in part on fisheries in the Pacific. A number of species—including sardines, herring, anchovies, hake, pollack, and shrimp—have been fished in some cases up to, if not beyond, the limits of their sustainable yields.

Salmon fisheries are important in the United States, Japan, Russia, and Canada, while the fishing quests for tuna have particular significance for the small island nations of the Pacific. The fishing for these species is done mostly by the technologically advanced vessels of Japan, the United States, South Korea, and Taiwan. Many of the tuna are found within the 200-mile exclusive economic zones of the island nations, giving them the opportunity to extract fees for fishing rights and to develop tuna fisheries and processing facilities.

Increasing population pressure and economic and often industrial development in several coastal regions of the Pacific have led to overfishing and impairment of habitat for a number of coastal species. Aquaculture is increasingly supplementing the natural supply. Shrimp, edible and pearl oysters, salmon, sea bream, mullet, and groupers are among the Pacific fauna that have been raised successfully.

In the tropical Pacific, prized corals have long been harvested from great depths: pink coral species come mainly from the western Hawaiian atolls, while black coral is extracted from the western islands and shallow seamounts and from Malaysian and Indonesian waters.

Mineral resources. Minerals are extracted from the seawater itself, from offshore alluvial deposits, or from within the continental shelf and may be metallic or nonmetallic, with hydrocarbon fuels being the most valuable among the latter. Prominent in the Pacific also are deepwater metallic mineral deposits of potential economic significance. Fresh water itself is also obtained from the ocean by various methods of desalination, as is done in Japan and Hong Kong.

Minerals from seawater and alluvial deposits. Common salt (sodium chloride) is the most important mineral obtained directly from seawater. Mexico leads the Pacific nations in salt extraction from the sea, mostly by solar evaporation. Bromine extracted from seawater is used in the food, dye, pharmaceutical, and photo industries. The United States and Japan lead in its production among Pacific riparian nations. Magnesium, recovered by an electrolytic process, is used in industrial metal alloys, especially with aluminum; Japan and California are the main sites for its extraction.

Sand and gravel

Also of importance are the sand and gravel that are extracted from the shallow sea bottom. Japan is the major Pacific producer, but Alaska, Hawaii, and other American Pacific-coast states also rely heavily on offshore sand and gravel. Sand and gravel mining from the seabed is important in nearly all other Pacific countries.

Large submarine deposits of phosphate rock are found in the Pacific off the coasts of Peru, eastern Australia, and California and on the Chatham Rise east of New Zealand. Smaller deposits also occur in lagoons of some of the Pacific Islands.

Deep-sea minerals. Metal-bearing deposits on the deep-sea floor, consisting of nodules, crusts, and accumulations of metallic sulfides from deep vents, are of potential economic interest. In the 1970s and '80s it was hoped that mining the nodules—which contain quantities of manganese, iron, copper, nickel, titanium, and cobalt, as well as small traces of other metals—might be a way to contribute to the wealth of developing nations. Economic considerations and concern over management of mining operations, however, led to cessation of the costly exploration and development of underwater mining technology.

Marine sulfide ores, containing iron, copper, cobalt, zinc, and traces of other metallic elements, are deposited in large amounts by the actions of deepwater hydrothermal vents, such as occur in the Pacific off the Galápagos Islands and on the Juan de Fuca and Gorda ridges.

Submarine hydrocarbons. Deposits of oil and gas under the seafloor are the most valuable and sought-after fuels of the modern world economy. Shallow seas and small ocean basins, such as the South and East China seas, are thought to have a high potential for large reserves. Among the countries bordering the Pacific Ocean and its marginal seas, the proportion of production from submarine reserves varies widely, from less than half in Indonesia and Japan to nearly all in Australia and Malaysia.

Main sites for exploration

The principal areas in the southwestern Pacific for offshore oil and gas exploration are in the South China Sea—the waters off Vietnam and off Hai-nan island in China and on the continental shelf northwest of Palawan island in the Philippines—but they also include the area off Natuna Island and some areas off the Sumatran coast in Indonesia. In the northwestern Pacific the main areas lie to the northwest of the island of Kyushu in Japan, in the southern portion of the Yellow Sea and in the Po Hai (Gulf of Chihli), and in regions off Sakhalin Island and the Kamchatka Peninsula. Oil and gas wells have been drilled in the Bering Sea in the north and areas off the southern coast of California in the eastern Pacific. In the southern Pacific, hydrocarbon production and exploration is taking place in the Gippsland Basin off the southeastern coast of Australia; in New Zealand's Cook Strait, its area in the Tasman Sea, and off South Island; and in areas off Fiji.

Trade and transportation. Since the end of World War II there has been remarkable growth in trade between countries on the western Pacific rim—most notably Japan, South Korea, and Taiwan—and North America, particularly the United States. Trade has also expanded between North America and Hong Kong and such Southeast Asian countries as Singapore, Thailand, Malaysia, and, to a lesser degree, Indonesia and the Philippines and in the western Pacific in the north between Japan and South Korea and Australia and Southeast Asia in the south. In addition, trade patterns in the United States have shifted, with Pacific nations now accounting for some two-fifths of overall trade. Los Angeles has surpassed New York City as the port with the country's largest trade volume, in terms of value, and the nearby port of Long Beach has also become one of the nation's largest.

Thus, the Pacific Ocean now contains some of the world's most important trade routes. Most of the exports moving from west to east and from north to south are high value-added manufactured goods. Conversely, most of the exports moving from east to west and from south to north are raw materials.

Outside of the United States, Japan is the largest Pacific trading nation, and the most important commodity flows in the Pacific are to and from Japan. Japanese imports, mostly raw materials, far exceed its exports in tonnage, but in monetary terms Japanese exports give the country an overall positive trade balance. Japan's imports include iron ore and coal from Australia and from India via the Indian Ocean, grain from the American Gulf Coast via the Panama Canal, and crude petroleum from the Persian Gulf via the Indian Ocean and from Indonesia. Its exports—principally motor vehicles, machinery, and precision and electronic equipment—are distributed virtually worldwide, although the largest quantities go to the United States.

Among other Pacific trading nations, Australia and New Zealand are principally exporters of raw materials, while South Korea, Taiwan, and Hong Kong are highly dependent on trade and are also large importers of raw materials. The small nations of the Pacific Islands contribute only a small fraction to overall Pacific trade, but most of them are heavily dependent on trade, particularly of imports of such basic materials as foodstuffs and petroleum. Generally, the small islands are net importers rather than net exporters, with only Papua New Guinea exporting more than it imports.

With the increased importance of the Pacific in worldwide trade has come a corresponding growth in the size of its transportation infrastructure. Japan, South Korea, China, and the Philippines rank high in ship ownership, and Japan, South Korea, and Taiwan are among the world's major shipbuilding nations. In addition to the ports of Los Angeles and Long Beach, the other major ports in the eastern Pacific are those in San Francisco Bay and in the Puget Sound region. In the western Pacific, Japan's ports on Tokyo and Ōsaka bays dominate trade; other major ports include Pusan in South Korea, Shanghai in China, Kao-hsiung in Taiwan, Hong Kong, and Sydney in Australia. Singapore, though on the fringe of the Pacific Ocean, dominates traffic to and from Southeast Asia and is the principal link between the Pacific and Indian oceans.

Major ports

Environmental impact of human activity. The Pacific Ocean is able to absorb, dilute, and disperse large quantities of human-generated wastes, but the increasing pace of economic activity—particularly the transport of crude oil and other hazardous substances—has led to measurable levels of pollution in some nearshore waters, especially close to ports and large coastal cities. At the same time, the loads of pollutants in waters close to land have increased—primarily raw sewage, industrial waste products such as heavy metals, and river-borne fertilizers and pesticides. Thus, although the large expanse of open ocean in the Pacific has not been seriously affected, serious problems have arisen in some more confined bodies of water. In a number of instances, populations of commercially important fish and crustaceans have been depleted or made unfit for use by the pollution of confined waters.

STUDY AND EXPLORATION

Early exploration and settlement. The Pacific Islands are thought to have been peopled by influxes from both mainland and archipelagic Southeast Asia. The earliest migrations were to what is now generally referred to as Melanesia. From there generations of voyagers ranged northward into eastern Micronesia and eastward into Polynesia. The peopling of Polynesia, by means of long-range seagoing voyages carried out in larger, seaworthy sailing canoes, constitutes a remarkable feat of navigational skill. These appear to have begun between 3,000 and 4,000 years ago and to have lasted until about 1,000 years ago, when the Maori (a Polynesian people) settled New Zealand. The longest voyages of the Polynesians involved the settlement of Hawaii, first from the Marquesas Islands and later from Tahiti.

European exploration. Pacific islanders had long inhabited their homelands before the European "discovery" of the Pacific in the 16th century. European exploration can be divided into three phases: Spanish and Portuguese, Dutch, and English and French. The Spanish and Portuguese period began with the voyages in the 1520s of Ferdinand Magellan and, after his death, of his crew members. Later discoveries included the Solomon Islands, the Marquesas, and possibly New Guinea, all by the Spaniard Álvaro de Mendaña de Neira; Vanuatu by the Portuguese Pedro Fernándes de Quirós; and the Torres Strait by the Spaniard Luis Váez de Torres.

During the Dutch period—roughly the 17th century—Jakob Le Maire and Willem Corneliszoon Schouten discovered inhabited islands in the northern Tuamotus, islands in the Tonga group, and Alofi and Futuna islands. The best-known of the Dutch explorers, Abel Janszoon Tasman, discovered New Zealand, islands in the Tonga group, the northeastern sector of the Fiji group, and islands in the Bismarck Archipelago.

Exploration and discovery in the Pacific reached its

zenith with the British and French in the 18th century. Four Englishmen—John Byron, Samuel Wallis, Philip Carteret, and James Cook—and the Frenchman Louis Antoine de Bougainville were preeminent. Byron explored the northern Marianas and discovered islands in the Tuamotuan, Cook, and Tokelauan archipelagoes. Wallis discovered islands in the Tahiti group, while Carteret sighted Pitcairn Island and explored vast areas of the southern Pacific. Bougainville sailed to Tahiti, Samoa, Vanuatu, New Guinea, and the Solomons.

Voyages of James Cook

European Pacific exploration and discovery culminated in the three voyages of Captain James Cook (1768–79). On his first voyage to Tahiti he discovered Raiatea, Vaitoare (Tahaa), Huahine, and Bora-Bora and surveyed the coasts of New Zealand and the east coast of Australia. On his second voyage he sailed south of 70° S, charted Tonga and Easter Island, and discovered New Caledonia. His third voyage included exploration of the North Pacific and Bering Strait, but he was killed in 1779 in the Hawaiian Islands, discovered earlier in the voyage. Cook's voyages left little land to be discovered in the Pacific, and his maps and charts were so accurate that many have not been substantially revised. The remaining island groups and larger landmasses were mapped in the 19th century.

In 1831–36 the English naturalist Charles Darwin sailed on the British naval vessel *Beagle,* collecting the information that he later would use in his writings. Growing interest in the physical and biological properties of the ocean paved the way for the great oceanographic voyages made later in the century. The best-known of these took place in the 1870s, initiated by the British Challenger Expedition, followed by the voyage of the USS *Tuscarora* in the northern Pacific and that of the German research vessel *Gazelle.*

Exploration in the 20th century. Among the most important 20th-century expeditions to the Pacific were those of the American ship *Carnegie* (1928–29), the Danish vessel *Dana II* (1928–1930), the Swedish ship *Albatross* (1947–48), and the Danish ship *Galathea* (1950–52). In the late 1950s the Soviet vessel *Vityaz* carried out detailed soundings of the deep trenches of the western Pacific, and in 1960 the American bathyscaphe *Trieste* descended to the bottom of the Mariana Trench.

Since 1960, numerous shorter cruises have been undertaken; they have been highly effective because of advances in instrumentation and knowledge of the basic characteristics of the ocean. In addition, satellites carrying remote sensing equipment have made it possible to determine with great accuracy the surface characteristics of the Pacific. These two factors have made lengthy cruises unnecessary. Exploration and research of the ocean bottom in the Pacific has likewise benefited from technological advances and has yielded the discovery of manganese nodules and other potentially valuable bottom resources. In addition, knowledge of marine life near active vents associated with centres of tectonic activity and of undersea volcanism has been greatly expanded.

Much of 20th-century oceanographic exploration in the Pacific and elsewhere has been for economic reasons, particularly for the recovery of offshore hydrocarbons in relatively shallow waters; but technology has made possible exploration and exploitation in increasingly deeper waters. Deepwater oil-drilling technology has also been employed to shed much light on the nature of the Earth's crust and the underlying mantle. (C.H.C./Jo.E.B./J.R.Mo.)

The major marginal seas of the Pacific Ocean

BERING SEA AND STRAIT

The Bering Sea (Russian: Beringovo More) is the northernmost part of the Pacific Ocean, separating the continents of Asia and North America at their closest point. To the north it connects with the Arctic Ocean through the Bering Strait (Proliv Beringa), at the narrowest point of which the two continents are about 53 miles (85 kilometres) apart. The boundary between the United States and Russia passes through the sea and the strait.

The Bering Sea roughly resembles a triangle with its apex to the north and its base formed by the 1,100-mile-long arc of the Alaska Peninsula in the east; the Aleutian Islands, which constitute part of the U.S. state of Alaska, in the south; and the Komandor (Commander) Islands in the west. Its area is about 890,000 square miles (2,304,000 square kilometres), including its islands. The maximum width from east to west is about 1,490 miles and from north to south about 990 miles.

The Bering Strait is a relatively shallow passage averaging 100 to 165 feet (30 to 50 metres) in depth. During the Ice Age the sea level fell by several hundred feet, making the strait into a land bridge between the continents of Asia and North America, over which a considerable migration of plants and animals occurred.

The strait as a bridge

In addition to the Aleutian and Komandor groups, there are several other large islands in both the sea and strait. These include Nunivak, St. Lawrence, and Nelson islands in Alaskan waters and Karagin Island in Russian waters.

Physical features. *Physiography.* The Bering Sea may be divided into two nearly equal parts: a relatively shallow area along the continental and insular shelves in the north and east and a much deeper area in the southwest. In the shelf area, which is an enormous underwater plain, the depths are, in most cases, less than 500 feet. The deep part in the southwestern portion of the sea is also a plain, lying at depths of 12,000 to 13,000 feet and divided by separate ridges into three basins: the larger Aleutian Basin to the north and east, the Bowers Basin to the south, and the Komandor Basin to the west. The sea's deepest point, 13,442 feet (4,097 metres), is in the Bowers Basin.

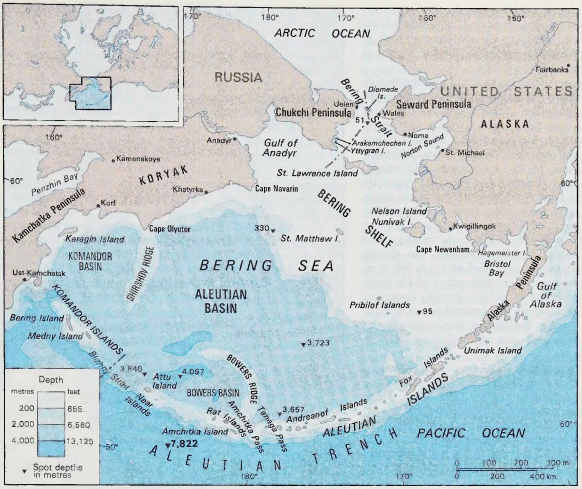
The continental crust is more than 12 miles thick along the shallow shelves and in the Aleutian Islands. The thickness decreases in the slope areas, and in the deep part of the sea the crust is 6 to 9 miles thick.

Enormous quantities of sedimentary material enter the sea annually from the land as a result of erosion of the shore. Plant and animal life at the surface also produce sedimentary material, but very little reaches the bottom, and consequently most of the sediment on the floor of the sea is from the land. Along with a great deal of silica, the bottom ooze holds a large quantity of boulders, pebbles, and gravel torn from the shores by ice and carried out to sea. In the southern part, the sediments are rich in material of volcanic origin.

Climate. Although the Bering Sea is situated in the same latitude as Great Britain, its climate is much more severe. The southern and western parts are characterized by cool, rainy summers with frequent fogs and comparatively warm, snowy winters. Winters are extreme in the northern and eastern portions, with temperatures of −31° to −49° F (−35° to −45° C) and high winds. The summers in the north and east are cool, with comparatively low precipitation. Snow persists on the Koryak coast for as long as 8 months and on the Chukchi Peninsula for nearly 10 months, with a snow cover one to two feet thick. The annual precipitation in the southern part of the sea is more than 40 inches (1,000 millimetres), mainly in the form of rainfall, while in the northern part the precipitation is less than half as much and is mainly snow.

Mean annual air temperatures range from −14° F (−10° C) in the northern areas to about 39° F (4° C) in the southern parts. Water temperatures on the surface average from 34° F (1° C) in the north to 41° F (5° C) in the south. The period without frosts lasts for about 80 days in the northern part of the sea, where snow is common even in the summer and maximum temperatures are only 68° F (20° C). In the southern area there are nearly 150 days without frost, and the temperature seldom falls much below freezing. January and February are the coldest months, July and August the warmest. Severe storms caused by strong centres of low atmospheric pressure occasionally penetrate the southern part of the sea.

Hydrology. Practically all of the Bering Sea water comes from the Pacific Ocean. The salinity of the surface water is relatively low, 31 to 33 parts per thousand; in the deeper parts of the sea the salinity increases to 35 parts per thousand near the bottom. In winter the northern portion of the sea is covered with ice, and even in summer the water below the surface retains a subfreezing temper-



The Bering Sea and the Bering Strait.

ature. The structure of the Bering Sea waters in general is subarctic, characterized by the presence in summer of a cold intermediate layer with warmer waters above and below. During the summer the surface water is heated, but a considerable layer of water that was cooled during the winter remains cold and is known as the cold intermediate layer. The maximum thickness of this intermediate layer is about 475 feet in the northern part of the sea and as much as 280 feet in the south. Underneath this layer is one that is slightly warmer, below which lie the colder bottom waters. In the northern and eastern shallow regions of the sea, only two upper layers develop: surface water and a cooler intermediate layer.

Warm oceanic waters from the south enter the Bering Sea through the numerous straits of the Fox Islands, through the Amchitka and Tanaga passes, and to a great extent through the Blizhny Strait between Attu and Medny islands. The Attu, Tanaga, and Transverse currents carry the warm water to the northwest. The Transverse Current, proceeding along the Asian continental slope in the direction of Cape Navarin, branches in two: one branch forms the Lawrence Current moving northward, and the other joins the Anadyr Current, which in turn gives birth to a powerful Kamchatka Current that governs the southward movement of the Bering Sea waters along the Asian coasts. Near the Alaska coast the general direction of the water is to the north, a factor responsible for the less severe ice conditions in that part of the sea as compared with the western part. Some of the Bering Sea water passes through the Bering Strait into the Arctic Ocean, but the bulk of it returns to the Pacific. The deep Bering Sea waters rise gradually to the surface and return to the Pacific as surface waters. Thus, the Bering Sea is an important factor in the general circulation of the northern part of the Pacific Ocean waters. The rise to the surface of oceanic waters rich in nutrient salts gives the sea a high biological productivity.

Transverse Current

Economic aspects. *Biological and mineral resources.* The existence in the Bering Sea of the cold intermediate layer separating the deep waters, which are rich in nutrient salts, from the upper photic layer (*i.e.,* the layer exposed to sunlight) results in two growths of floating plant life during the year. The first growth occurs in the spring after the mixing of waters in winter, and the second during the autumnal mixing, when the cold surface waters descend and the deeper waters come to the surface while there is still sufficient sunlight for plant growth.

This floating plant life consists of some 160 species, of which the most common are diatom algae. The largest concentration of diatoms have been found in the shallow part of the sea. Diatoms are the principal producers of organic matter, and they are consumed by small copepods (microscopic crustaceans), which in turn become the food of fish and mammals. On the continental shelf there are vast quantities of mollusks, echinoderms (particularly sea urchins and starfish), and barnacles. Also abundant on the shelves are sponges, marine worms, and crustaceans. In the southern regions, down to depths of 100 or 130 feet, populations of giant brown algae grow like forests on the rocky bottom. There are about 200 species of algae, some reaching lengths of 200 to 300 feet.

The Bering Sea has more than 300 species of fish, including 50 deep-sea species, of which 25 are caught commercially. The most important among them are salmon, herring, cod, flounder, halibut, and pollack. The islands are breeding grounds for the fur seal and the sea otter. The northern areas are inhabited by the walrus, seal, and sea lion. Several whale species, notably gray whales, migrate to Bering waters to feed during the summer. Intensive fishing in the last half of the 20th century has drastically reduced some of the most valuable fish species, and this has led to greater exploitation of less commercially valuable species.

Commercially valuable fish species

Oil and gas deposits are believed to exist under the Bering Shelf and along the margin of the Kamchatka Peninsula. The extent of potential reserves, however, is unknown.

Navigation. The Bering Sea is considered to be one of the most difficult bodies of water to navigate. Winter storms are frequent and severe, often coating the superstructures of ships with ice. Wave heights may exceed 40 feet. Added to these hazards are powerful tidal currents in many parts of the sea and fog, rain, and floating ice in the north. In winter the northern area is covered by ice fields about 4 or 5 feet thick, with hummocks in some places more than 100 feet high. At its maximum extent in April, the ice reaches as far south as Bristol Bay and the Kamchatka coasts. Melting begins in May, and by July there is

no ice in the sea except for drift ice in the Bering Strait. Nonetheless, the sea contains important shipping routes for the Soviet Far East, including the eastern terminus at Provideniya on the Chukchi Peninsula for the northern sea route to Arkhangelsk in the west.

Vitus Bering

Study and exploration. The Bering Strait and the Bering Sea were first explored by Russian ships under Semyon Dezhnyov, in 1648. They are named for Vitus Bering, a Danish captain who was taken into Russian service by Peter the Great, in 1724. He sailed into the strait four years later but did not see the Alaskan coast, although he discovered the islands of St. Lawrence and Diomede. In 1730 the strait was charted for the first time by Mikhail Gvozdev and Ivan Fyodorov. Bering sailed again in 1733, leading a large expedition from St. Petersburg along the northern coast of Siberia, and he reached the Gulf of Alaska in the summer of 1741. He reconnoitred the southwestern coast of mainland Alaska, the Alaska Peninsula, and the Aleutians, but misfortune befell him, and he perished in that year along with many of his men. In 1780, Russian merchants founded a private company to trade in fur-bearing animals in northwestern America. A geographic study of the Bering Sea was made at the end of the 18th century and was supplemented later by hydrographic studies.

Deep-sea studies were begun in 1827 by British explorers. Extensive work was also done by a U.S. group aboard the American research vessel *Albatross* in 1893–1906. Since then the sea has been systematically studied by Soviet, American, and Japanese investigators. Some of the most detailed study was undertaken by the Soviet vessel *Vityaz* in a series of expeditions undertaken in the 1950s and '60s. (A.P.L./A.V.A./K.T.B.)

EAST CHINA SEA

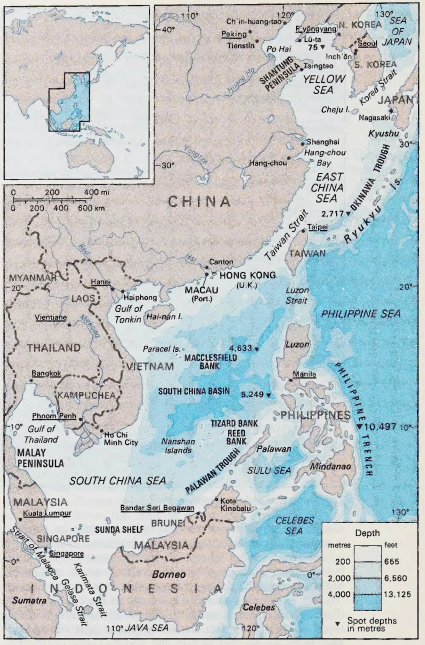
The East China Sea, or Eastern Sea, of the western Pacific Ocean—known in Chinese as the Tung Hai (Pinyin: Dong Hai)—borders the East Asian mainland and extends northeastward from the South China Sea, to which it is connected by the shallow Taiwan Strait between Taiwan and mainland China. The sea extends to the east to the chain of the Ryukyu Islands; north to Kyushu, which is the southernmost of Japan's main islands; northwest to Cheju Island off South Korea; and hence west to China. This northern boundary separates the East China Sea from the Yellow Sea.

Physical features. *Physiography.* The East China Sea, which has an area of 290,000 square miles (752,000 square kilometres), is largely shallow, with almost three-fourths of the area less than 650 feet (200 metres) and an average depth of only 1,145 feet. The deeper part is the Okinawa Trough, extending alongside the Ryukyu Islands, with a large area more than 3,300 feet deep and a maximum depth of 8,912 feet (2,717 metres). The western edge of the sea is a continuation of the shelf that extends between the South China Sea and the Yellow Sea. A large number of islands and shoals dot the eastern boundary as well as the area near the Chinese mainland.

The shallow shelf areas are covered with sediments from the bordering landmasses deposited mainly by the Yangtze and other rivers near the northern part of the sea. Coarser sediments of sand occur farther out, and rocks, muds, and oozes are also found in scattered areas. Seismic profiling indicates that the geologic structure beneath the ocean floor consists of nearly parallel folds, with rock ridges near the northern limits of the East China Sea, near the edge of the continental shelf, and along the Ryukyus. These have afforded barriers for sediment brought down by the great Huang Ho (Yellow River) and also the Yangtze. Between the ridges the sediment is up to one mile thick.

Most of the shelf belongs to the stable Neo-Cathaysian Geosyncline (or Cathaysian Platform), dating back at least 300 million years or longer. The Okinawa Trough appears to be perhaps 10 million years old. The Ryukyus are a double-island chain, with several volcanic islands on the East China Sea side. Many of the volcanoes are still active. Epicentres of earthquakes are found along the axis of the Okinawa Trough and the Ryukyu Island arc.

Climate. Weather is dominated by the monsoon wind system, the result of differential heating between land and water. In summer the Asian landmass is much warmer than the sea; in the winter it is much colder, particularly in the Plateau of Tibet. Summer heating of air masses over Asia builds areas of low pressure and creates the monsoonal winds, which in this season blow predominantly from the southeast. This brings in warm, moist air from the western Pacific, producing a rainy summer season that is accompanied by typhoons. In winter the situation is reversed: winds blow predominantly from the north, bringing with them cold, dry air from the continent.



The East China, South China, and Yellow seas.

Sea currents and tides

Hydrology. Winds also influence water circulation of the Kuroshio (Japan Current), the north-flowing western extension of the warm North Equatorial Current. Some of the Kuroshio enters the eastern part of the East China Sea to form the Tsushima Current, which flows north into the Sea of Japan, while the main part of the current diverts eastward back out into the Pacific south of Kyushu and flows east of Japan. Strengthened by monsoon winds, the Kuroshio is at its widest and fastest in summer, and the axis is displaced well into the East China Sea. This warmed surface water varies from 86° F (30° C) in the south to 77° F (25° C) in the north. In winter, northerly monsoon winds modify the circulation, and the north-flowing Kuroshio, though still important, is reduced in strength, while southerly flowing coastal currents are strengthened. This brings in colder water, with temperatures of 41° F (5° C) in the north to 73° F (23° C) in the south. Some warm water also enters the sea via the Luzon and Taiwan straits.

Because of the constricting nature of the adjoining Yellow Sea and the funnel shape of some of the inlets on the mainland, tidal ranges are especially high along the coast of China. For example, the spring tide range, which

is highest in summer and winter, is as much as 23 feet at San-sha Bay and 30 feet at Hang-chou Bay.

Economic aspects. The East China Sea is a region of high marine-life productivity, and China, Japan, and North and South Korea actively fish in the area. Most of the fishing is done by small local boats, although larger trawlers are also used. Tuna, mackerel, shrimps, sardines, milkfish, sea breams, croakers, shellfish, and seaweeds are the main resources harvested.

Oil and natural gas deposits have been discovered under the sea's continental shelf. These discoveries have led to disputes among the bordering countries over the control of areas with potential hydrocarbon reserves, particularly the regions around deep trenches, straits, rocks, and uninhabited islands. A small amount of China's oil production comes from offshore wells in the East China Sea.

In addition to the local shipping traffic in and out of Chinese and Korean ports, the East China Sea serves as the main shipping route from the South China Sea to Japanese and other North Pacific ports. The main ports of the East China Sea are Shanghai in China, Nagasaki in Japan, and Chi-lung in Taiwan.

Study and exploration. Although the East China Sea has been known and traveled for centuries, the major oceanographic research on it has been done since 1950. Most of the research has been carried out by the Chinese, sometimes in cooperation with the Japanese. In addition to gaining a broader scientific knowledge of the sea, these efforts have resulted in the discovery of the warm current coming from the South China Sea and of nearby offshore oil and natural gas reserves. (E.C.La.F.)

SEA OF JAPAN

The Sea of Japan (Japanese: Nihon-kai; Russian: Yaponskoye More) is a marginal sea of the western Pacific Ocean. It is bounded by Japan and Sakhalin Island to the east and by Russia and Korea on the Asian mainland to the west. (The Koreans refer to the sea as the "East Sea," or Tonghae.) Its area is 377,600 square miles (978,000 square kilometres). It has a mean depth of 5,748 feet (1,752 metres) and a maximum depth of 12,276 feet (3,742 metres).

Physical features. *Physiography.* The sea is almost elliptical, having its major axis from southwest to northeast. To the north it is approximately bounded by latitude 51°45′ N, while to the south it is bounded by a line drawn from the Japanese island of Kyushu westward through the Gotō Islands of Japan to the South Korean island of Cheju and then northward to the Korean Peninsula.

The sea itself lies in a deep basin, separated from the East China Sea to the south by the Tsushima and Korea straits and from the Sea of Okhotsk to the north by the La Perouse (or Sōya) and Tatar straits. To the east it is also connected with the Inland Sea of Japan by the Kanmon Strait and to the Pacific by the Tsugaru Strait.

The Sea of Japan is a classic semienclosed sea, since its connections with adjacent bodies of water are greatly restricted by the narrow straits. Inflow of water takes place primarily through the eastern and western channels of the Korea Strait; the inflow of water into the Sea of Japan through the narrow and shallow Tatar Strait is negligible, while through the Tsugaru and La Perouse straits the water flows out of the Sea of Japan.

Component basins

Underwater the sea is separated into the Japan Basin to the north, the Yamato Basin to the southeast, and the Tsushima Basin to the southwest. While a narrow continental shelf fringes Siberia and the Korean Peninsula, on the Japanese side of the sea there are wider continental shelves, as well as groups of banks, troughs, and basins lying offshore. The banks lying off the coasts of Japan are divided into groups, which include Okujiri Ridge, Sado Ridge, Hakusan Banks, Wakasa Ridge, and Oki Ridge.

Geology. Yamato Ridge consists of granite, rhyolite, andesite, and basalt, with boulders of volcanic rock scattered on the seabed. Geophysical investigation has revealed that, while the ridge is of continental origin, the Japan Basin and the Yamato Basin are of oceanic origin.

Bottom deposits in the Sea of Japan indicate that sediments of continental origin, such as mud, sand, gravel, and fragments of rock, exist down to depths of 650 to 1,000 feet; hemipelagic sediments (*i.e.,* half of oceanic origin), mainly consisting of blue mud rich in organic matter, are found down to depths of about 1,000 to 2,600 feet; and deeper pelagic sediments, consisting of red mud, are found down to the sea's greatest depths.

The four straits connecting the sea either to the Pacific Ocean or to marginal seas were formed in recent geologic periods. The oldest of these are the Tsugaru and Tsushima straits, the formation of which interrupted the migration of elephants into the Japanese islands at the end of the Tertiary Period (about 1.6 million years ago); the most recent is La Perouse Strait, which was formed about 60,000 to 11,000 years ago and which closed the route used by the mammoths whose fossils have been found in Hokkaido.

Climate. The Sea of Japan influences the climate of Japan because of its relatively warm waters; evaporation is especially noticeable in winter, when an enormous quantity of water vapour rises in the region between the cold, dry polar air mass and the warm, moist tropical air mass. From December to March the prevailing northwest monsoon wind carries cold and dry continental polar air masses over the warmer waters of the sea, resulting in snow along the mountainous western coasts of Japan. In summer the southerly tropical monsoon blows from the North Pacific onto the Asian mainland, causing dense fog when its warm and moist winds blow over the cold currents that prevail over the northern part of the sea at that season. The winter monsoon brings rough seas and causes coastal erosion along the western coasts of Japan.

The northern part of the sea, especially off the Siberian coast as well as in the Tatar Strait, freezes in winter; as a result of convection, melted ice feeds the cold currents in that part of the sea in spring and summer.

Prevailing currents

Hydrology. The waters of the sea generally circulate in a counterclockwise pattern. A branch of the Kuroshio (Japan Current), the Tsushima Current, together with its northern branch, the East Korea Warm Current, flows north, bringing warmer and more saline water before flowing into the Pacific through the Tsugaru Strait as the Tsugaru Current, as well as into the Sea of Okhotsk through the La Perouse Strait as the Sōya Current. Along the coast of the Asian mainland three cold currents—the Liman, North Korea, and Central (or Mid-) Japan Sea cold currents—bring cooler, relatively fresh water southward.

The enclosed nature of the Sea of Japan and its three deep basins facilitate the formation of distinct water masses. In the summer months, when the surface is heated by higher air temperatures, a distinct thermocline separates the surface water from the middle water, which is characterized by relatively high temperatures and salinities. The middle water's origin is in the intermediate water layers of the Kuroshio off the coast of Kyushu that enter the Sea of Japan via the Tsushima Current during the winter and spring. A third water mass, the deep water, forms in February and March by the cooling of the water surface in the northern part of the sea; it is of very uniform character, with temperatures ranging narrowly between 32° and 33° F (0° and 0.5° C). The formation of these water masses, particularly the middle water, influences the concentration of dissolved oxygen, which is generally very high; and this contributes to the productive fisheries of the sea.

Economic aspects. Fisheries and mineral deposits form the main economic resources of the Sea of Japan. Pelagic (oceanic) fishes include mackerel, horse mackerel, sardines, anchovies, herring, fishes of the salmon and trout family, sea bream, and squid; the demersal (sea-bottom) category includes cod, Alaskan pollack (bluefish), and Atka mackerel. Seals and whales are also to be found, as well as such crustaceans as shrimps and crabs. The fishing grounds are for the most part on the continental shelves and their adjacent waters.

Herring, sardines, and bluefin tuna have traditionally been caught, but since World War II the fisheries have gradually been depleted. Squid fishing is carried on in the central part of the sea, salmon fishing in the shoal areas of the north and southwest, and crustacean trapping in the deeper parts. The sea is heavily fished by fleets from Japan, Russia, and North and South Korea.

Mineral resources on or in the sea bottom include mag-



The Seas of Japan and Okhotsk.

netite sands as well as natural gas and petroleum deposits off Japan and Sakhalin Island.

Trade across the Sea of Japan is only moderate, since most of Japan's trade is with countries not bordering the sea. Consequently, the most important Japanese ports are located on its Pacific coast. Important ports of South Korea are Pusan, Ulsan, and P'ohang, located on the southeast coast of the country, but most of the shipping in and out of these ports is also destined for countries not bordering the sea. Primary Russian ports are Vladivostok, Nakhodka, and Vostochny. Vladivostok's traffic is primarily with other Russian ports, while Nakhodka and Vostochny are international ports. Trade between countries around the sea, however, has increased, spurred by the growth of the South Korean economy and by the development of trade agreements with Russia.

Study and exploration. The waters of the Sea of Japan have been traveled for centuries, and historically they have served to protect Japan from foreign invasions. European exploration in the region began in the 18th century. In the 1780s in one of the first voyages both to search for new trade opportunities and to collect scientific data, the Frenchman Jean-François de Galaup, Count de La Pérouse, traveled northward through the Sea of Japan and the strait that was named for him. Robert Broughton in 1796 also combined exploration with science on his track through the Tatar Strait and then south along the coast of the Russian Far East and around the Korean Peninsula. European exploration

Other Russian voyages include those of Adam Johann Krusenstern in the *Nadezhda,* who explored the sea (1803–06) during his circumnavigation of the globe; the *Pallada* (1853–54), which monitored surface temperatures and made tidal observations; and the *Vityaz* (1886–89), which added to oceanographic knowledge of the sea. Also of note are the American North Pacific Exploring (1853–54) and British Challenger (1872–76) expeditions.

Major research cruises by Japan began in 1928 and continued until the start of the Sino-Japanese War in 1937. After World War II the American *Challenger* seismic study of the sea (1951) and the work of the *Baird* added to scientific knowledge and stimulated Japan to reinstitute exploration and research. Most of the current oceanographic work in the sea is characterized by the use of advanced technology, by international cooperation, and by a variety of research projects. (M.U./J.R.Mo.)

SEA OF OKHOTSK

The Sea of Okhotsk (Russian: Okhotskoye More), which borders the northwestern Pacific Ocean, is bounded on the west and north by the east coast of Asia from Cape Lazarev to the mouth of the Penzhina River, on the east and southeast by the Kamchatka Peninsula and the Kuril Islands, on the south by the northern coast of the Japanese island of Hokkaido, and on the southwest by Sakhalin Island. Except for the small area touching Hokkaido, the sea is completely enclosed by Russian territory. Its area covers 611,000 square miles (1,583,000 square kilometres), and it has a mean depth of about 2,818 feet (859 metres). The sea's maximum depth is 11,063 feet (3,372 metres).

Physical features. *Physiography.* For the most part, the continental shore is high and rocky and interspersed with the mouths of large, regional rivers that flow into the sea—the Amur, Tugur, Uda, Okhota, Gizhiga, and Penzhina. The shores of Hokkaido and northern Sakhalin Island are predominantly lower. The Gulfs of Aniva and Terpeniya are found on the southeastern coast of Sakhalin. Nearly all of the other main islands—Shantar, Zavyalov, Spafaryev, Yam, and Tyuleny—are situated close to the shore; only Ion Island is in the open sea.

The formations of the Okhotsk basin have evolved in the Quaternary Period (the past 1.6 million years). The geologic history of the sea shows that the forming of the present hollow was accompanied by repeated glacial transgressions and regressions. The sea bottom has a slope from north to south and southwest. The northern and northwestern parts constitute a continental shallow up to 650 feet in depth, while the remaining area (about 70 percent of the total) is a continental slope from 650 to 5,000 feet deep, on which isolated underwater heights, depressions, Geologic history

and troughs stand out. The deepest part of the sea, the Kuril Basin west of the Kuril Islands, reaches 8,200 feet and deeper.

Large quantities of continental sediment flow into the sea, primarily from the Amur River. Other sources of sediment include coastal abrasion and volcanic activity. Bottom deposits in the Kuril Basin consist of a clay-diatom silt, whereas approaching the shore there are fine, silt-covered sands, coarse sands, and pebbles mixed together with mussel shells.

Climate. The Sea of Okhotsk is the coldest sea of East Asia; in winter the climate and thermal regime over much of the region differ only slightly from those found in the Arctic. The northeastern, northern, and western regions of the sea experience severe weather during the winter, because of the influence of the Asian continent; from October through April these areas experience very cold air temperatures, are constantly covered with ice, and have very little precipitation. In short, a continental climate pervades these parts of the sea. To the south and southeast the proximity of the Pacific results in a milder marine climate. The coldest months in the sea are January and February; the warmest are July and August. In the northeastern part the average monthly air temperature during February is −4° F (−20° C), while in August the average is 54° F (12° C). To the north and west of the sea, average monthly air temperature is −11° F (−24° C) in February and 57° F (14° C) in August. In the southern and southeastern parts, the average monthly air temperature is 19° F (−7° C) in February and 64° F (18° C) in August. The yearly precipitation averages 16 inches (400 millimetres) in the north, 28 inches in the west, and about 41 inches in the south and southeast.

Hydrology. The water of the Sea of Okhotsk consists of continental drainage, precipitation, and waters flowing from the Pacific Ocean through the straits of the Kuril Islands and from the Sea of Japan through the La Perouse (Sōya) Strait. During the summer months the sea is warmed to a depth of 100–165 feet. The water temperature on the surface rises to 46° to 54° F (8° to 12° C), and the salinity drops to 32.5 parts per thousand and lower. Deeper water has an average temperature of 29° to 30° F (−1.8° to −1° C) and salinity of up to 34 parts per thousand. The thickness of the cold-water layer fluctuates from a few feet in the southeastern part of the sea to 245–525 feet in the northwest.

Water currents

The general movement of water in the sea is counterclockwise. Water flows from the Sea of Japan into the Sea of Okhotsk, accounting for the comparative warmth of its southwestern part. Warm water is also carried into the sea by Pacific currents. Because of the influence of these currents, the waters of the eastern half of the sea are warmer than those of the western part. For the most part, the currents flow clockwise around the Kuril Islands; in the northern half of the straits they flow into the sea, but in the southern half they return into the Pacific. Penzhin Bay has the strongest tides (42.3 feet), while the weakest tides occur at southeastern Sakhalin (2.6 feet). Ice cover appears at the end of October and reaches its greatest extent in March. In the coastal areas it extends to the shore, but in the open sea there is floating ice. The ice vanishes in June, with the exception of the Sakhalin gulfs and the region around Shantar Island, where ice floes are not at all uncommon in July and sometimes they occur even in August.

Economic aspects. *Biological and mineral resources.* The Sea of Okhotsk is one of the most biologically productive of the world's seas. The river drainage, the intense intermingling of waters by straits and wind, and the upwelling of deep, nutrient-laden ocean waters are all favourable to marine life. In those months when it is warm enough, there occurs an extraordinarily rapid spread of life. The flora is represented by algae and seaweed and the fauna by crayfish, sea mussels, crabs, sea urchins, polyps, and various types of fish. Salmon, herring, pollack, flounder, cod, capelin, and smelts (or frostfish), as well as crab and shrimp, are all commercially important. The sea is also inhabited by marine mammals—whales, seals, and sea lions.

Almost the entire sea fell under the supervision of the Soviet Union in 1977 when a 200-mile exclusive economic zone was established. This produced favourable conditions for the development of fisheries and for mineral exploitation. The sea now supplies a large portion of the catches in eastern Russia. Also, deposits of oil and natural gas have been discovered on the sea's northern shelf.

Navigation. Regular navigation connecting the ports of eastern Russia is conducted through the sea. On the continental coast the most important of these ports are Magadan in Nagayeva Bay and Okhotsk. Korsakov on Sakhalin Island and Severo-Kurilsk and Yuzhno-Kurilsk on the Kuril Islands are also important. Ice floes are an impediment to sea navigation during the winter, and dense fog is a hindrance during the summer. Strong currents and submerged rocks are other perils of the area. The Sea of Okhotsk region plays an important role in the economic development of eastern Russia.

Study and exploration. The Sea of Okhotsk was the first Pacific body of water in which Russian explorers ventured. In 1787 the French navigator Jean-François de Galaup, Count de La Pérouse, sailed northward through the strait named for him and crossed the sea to the Kamchatka Peninsula. The Russian explorer Adam Johann Krusenstern also sailed through the sea to the peninsula in the early 19th century. The expedition of the Soviet research vessel *Vityaz* in 1949 marked the beginning of the modern stage of oceanographic research in the sea. Continuing studies have been carried out by Russian scientists.

Russian exploration

(T.Y.S./A.V.A./K.T.B.)

SOUTH CHINA SEA

The South China Sea—known in Chinese (Wade-Giles and Pinyin) as Nan Hai, or the South Sea—is that part of the western Pacific Ocean that borders the Southeast Asian mainland. It is bounded on the northeast by the Taiwan Strait (by which it is connected to the East China Sea), on the east by Taiwan and the Philippines, on the southeast and south by Borneo, the southern limit of the Gulf of Thailand, and the east coast of Malay Peninsula, and on the west and north by the Asian mainland. The southern boundary of the sea is a rise in the seabed between Sumatra and Borneo, and the northern boundary stretches from the northernmost point of Taiwan to the coast of Fukien province, China, in the Taiwan Strait. It embraces an area of about 1,423,000 square miles (3,685,000 square kilometres), with a mean depth of 3,478 feet (1,060 metres). For map coverage of the South China Sea, see above *East China Sea.*

Physical features. *Physiography.* The major feature is a deep, rhombus-shaped basin on the eastern part, with reef-studded shoals rising up steeply within the basin to the south (Reed and Tizard banks and the Nanshon Island area) and northwest (Paracel Islands and Macclesfield banks). The deep portion, called the China Sea Basin, has a maximum depth of 16,452 feet (5,042 metres) and an abyssal plain with a mean depth of 14,100 feet.

Along the eastern side of the basin, the continental shelf falls off sharply near the Philippine islands of Luzon and Palawan and forms the Palawan Trough near the latter island. Along the northwest side of the basin to the mainland is a broad, shallow shelf as wide as 150 miles (240 kilometres). This shelf includes the Gulf of Tonkin and the Taiwan Strait and the large islands of Hai-nan and Taiwan.

To the south, off Vietnam, the shelf narrows and connects with the Sunda Shelf, one of the largest in the world, which covers the area between Borneo, Sumatra and the Malay Peninsula and includes the southern part of the South China Sea, the Gulf of Thailand, and the Java Sea. This broad trough is about 130 feet deep at its periphery and up to 330 feet in its central part. On the bottom of the shelf is a network of submerged river valleys that converge into the Sunda Depression and then into the China Sea Basin.

The Sunda Shelf is covered with littoral sediments contributed by submerged valleys. The inner zone of mud is characteristic of the continental shelf near the Mekong and Red river deltas, while the sediment of the deeper

parts of the South China Sea is mainly composed of clay. A characteristic part of the sediments in both deep and shallow water is volcanic ash. This is found in layers, derived from large volcanic eruptions in the East Indies, notably the enormous eruption of Krakatoa in 1883, when ash was transported through the entire area by both wind and currents.

Connecting channels

The South China Sea has connecting channels. The Taiwan Strait on the north is about 100 miles wide, with a depth of about 230 feet. The main deep channel connecting the South China Sea with the Pacific Ocean, the Luzon Strait, lies between Taiwan and the Philippines and has a depth of about 8,500 feet. Shallow channels are found on the east along the Philippine island chain and on the south between Borneo and Sumatra. The western connection to the Indian Ocean is the long Strait of Malacca. At its narrowest part it is 19 miles wide and about 100 feet deep.

The South China Sea is the largest marginal sea of the western Pacific. Some 1 to 60 million years ago, it was rifted and collapsed as a result of seafloor spreading. The China Sea Basin is believed to have dropped 2.5 miles, leaving residual plateaus studded with numerous coral reefs, islets, and banks, some of which are drowned atolls.

Climate. Weather in the sea is tropical and largely controlled by monsoons. In summer, monsoonal winds blow predominantly from the southwest; in winter, winds blow from the northeast. Annual rainfall varies from about 80 inches (2,000 millimetres) to more than 120 inches around the southern basin. Summer typhoons are frequent.

Hydrology. Monsoons control the sea-surface currents as well as the exchange of water between the South China Sea and adjacent bodies of water. In August the surface flow into the South China Sea is from the south through the Karimata and Gelasa (Gasper) straits. Near the mainland the general flow is northeasterly, passing out through the Taiwan and Luzon straits. There is a weak countercurrent on the eastern side of the sea. In February the flow is generally to the southwest; the strongest flow occurs in summer along the bulging part of Vietnam, with speeds of up to three knots (nautical miles per hour) generated by the strong southwestern monsoon.

The near-surface waters are relatively warm (about 84° F [29° C] in the summer) because of the low latitude and a tendency for the equatorial current to feed warm water into the area. In early summer, wind from the southwest not only moves the surface water to the northeast but causes it to be displaced off the coast. As a result, upwelling areas having colder surface temperatures and higher nutrient content are found off central Vietnam. In winter the general surface temperature is cooler, ranging from about 70° F (21° C) in the north to 81° F (27° C) in the south.

The major rivers draining into the South China Sea are the tributaries of the Pearl River delta south of Canton, including the Hsi River; the Red River at Haiphong, Vietnam; and the Mekong River, near Ho Chi Minh City (Saigon), Vietnam. The wet summer season causes the Mekong to triple its annual average flow, and it causes an even greater relative change in the flow of the Red.

Economic aspects. The South China Sea is rich in marine life. Contributing to this abundance are the extensive runoff of nutrient-laden waters from land and the upwellings of water in certain areas of the sea. The sea is heavily fished, however, and is the main source of animal protein for the densely populated Southeast Asian area. Most abundant are the various species of tuna, mackerel, croaker, anchovy, shrimp, and shellfish. Nearly the entire catch is consumed locally, either fresh or preserved.

Fishing

Large reserves of oil and natural gas have been discovered under the floor of the South China Sea. The main locations for hydrocarbon production are located north of Borneo, east of the Malay Peninsula, and northwest of Palawan.

The South China Sea contains some of the world's most important shipping lanes. The main route to and from Pacific and Indian ocean ports is through the Strait of Malacca and the South China Sea. Generally, oil and minerals move north, and food and manufactured goods move south. Some areas in the central South China Sea are not well sounded, and nautical charts bear the notation "dangerous ground."

Study and exploration. The South China Sea has long been known and traveled. For centuries, however, pirates prowling its waters made passage through the sea hazardous. The sea has been fairly thoroughly studied in the 20th century. A number of cooperative programs have been conducted since World War II, including an expedition by Thailand, the United States, and South Vietnam in 1959–61 that investigated marine life, water structures, and circulation. Satellite photography has been particularly useful. One such study of currents showed that Taiwan acts as a ship's prow, deflecting the north moving currents around the island. Further study demonstrated that these points of deflection are upwell areas, which led to the establishment of commercial fishing there. (E.C.LaF.)

YELLOW SEA

The Yellow Sea (Chinese: Huang Hai) is a large inlet of the western Pacific Ocean lying between the Chinese mainland on the west and north and the Korean Peninsula on the east. It is situated to the north of the East China Sea, which it bounds on a line running from the mouth of the Yangtze River to Cheju Island off South Korea. It measures about 600 miles (960 kilometres) from north to south and about 435 miles from east to west. In the northwest part of the sea, northwest of a line between the Liaotung Peninsula to the north and the Shantung Peninsula to the south, is the Po Hai (Gulf of Chihli). The area of the Yellow Sea proper (excluding the Po Hai) is about 161,000 square miles (417,000 square kilometres), its mean depth is 131 feet (40 metres), and its maximum depth is 338 feet (103 metres).

For map coverage of the Yellow Sea, see above *East China Sea.*

Physical features. *Physiography and geology.* The Yellow Sea, including the Po Hai and Korea Bay, forms a partly enclosed, flat, and shallow marine embayment. Most of the sea, which is deeper than the Po Hai, consists of an oval-shaped basin with depths of about 200 to 260 feet.

Floor of the sea

The floor of the Yellow Sea is a geologically unique, shallow portion of the continental shelf that was submerged only after the last Ice Age (*i.e.,* within the past 10,000 years). The seafloor slopes gently from the Chinese mainland and more rapidly from the Korean Peninsula to a north-south-trending seafloor valley with its axis close to the Korean Peninsula. This axis represents the path of the meandering Huang Ho (Yellow River) when it flowed across the exposed shelf during times of lowered sea levels and emptied sediments into the Okinawa Trough. The Yellow Sea derives its name from the colour of the silt-laden water discharged from the major Chinese rivers emptying into it. The sea annually receives an immense quantity of sediments, mostly from the Huang Ho and Yangtze River, both of which have formed large deltas. Relict sandy sediments occupy the northern part of the Yellow Sea, the nearshore northern Po Hai, the offshore old Huang Ho delta, and the central part of the south Yellow Sea. The sandy layer is covered with silty and muddy sediments derived from the large rivers of China and Korea since the last glacial period. The dividing line between silt derived from China and sand derived from Korea nearly coincides with the seafloor valley.

Climate. Generally, the climate is characterized by very cold, dry winters and wet, warm summers. From late November to March, a strong northerly monsoon prevails, which in the Po Hai is sometimes accompanied by severe blizzards. Typhoons occur in summer, and in the colder season there are occasional storms. Air temperatures range from 50° to 82° F (10° to 28° C) and precipitation from about 20 inches (500 millimetres) in the north to 40 inches in the south. Sea fog is frequent along the coasts, especially in the upwelling cold-water areas.

Hydrology. The warm current of the Yellow Sea is a part of the Tsushima Current, which diverges near the western part of the Japanese island of Kyushu and flows at less than 0.5 knot (nautical mile per hour) northward into the middle of the sea. Along the continental coasts south-

ward-flowing currents prevail, which strengthen markedly in the winter monsoon period when the water is cold, turbid, and of low salinity.

The tidal range is considerable (13 to 26 feet) along the shallower west coast of the Korean Peninsula, with a maximum spring tide of almost 27 feet. Along the coasts of China it amounts to about 3 to 10 feet (except the Po Hai, where it exceeds 10 feet). In the Yellow Sea the tides are semi-diurnal (*i.e.,* they rise twice daily). The tidal system rotates in a counterclockwise direction. The speed of the tidal current is less than one knot in the middle of the sea, but near the coasts and in the straits and channels stronger currents of more than two or three knots are recorded.

Temperature and salinity

The innermost coastal sections of the Po Hai freeze in winter, and drift ice and ice fields hinder navigation in parts of the Yellow Sea. Surface temperature ranges from freezing level in winter in the Po Hai to summer temperatures of from 72° to 82° F (22° to 28° C) in the shallower parts. Thus, the annual range is very large—from 40° to 50° F (4° to 10° C). In winter the temperature and salinity in the sea are homogeneous from surface to bottom. In spring and summer the upper layer is warmed and diluted by the fresh water from rivers, while the deeper water remains cold and saline. This deep layer of cold water stagnates and moves slowly south in summer. Around this mass of water, especially at its southern tip, concentrations of commercial bottom fishes are found. The dominant salinity in the region is relatively low: in the Po Hai it is 30 to 31 parts per thousand, while in the Yellow Sea proper it is 31 to 33 parts per thousand. In the southwest monsoon season (June to August) the increased rainfall and runoff cause a further reduction in salinity in the upper layer in summer.

Economic aspects. The Yellow Sea, like the East China Sea, is famous for its fishing grounds. The rich demersal (bottom-dwelling) fish resources have been exploited by Chinese, South and North Korean, and Japanese trawlers for years. Although the overall annual catch has grown, the catch by the Japanese has decreased, while those of the Chinese and South Koreans have increased. The main species caught are sea bream, croakers, lizard fish, prawns, cutlass fish, horse mackerel, squids, and flounders; all species, however, are overfished, and the catch of particularly valuable species has declined.

Oil exploration has been successful in the Chinese and North Korean portions of the Yellow Sea. In addition, the sea has become more important with the growth in trade among its bordering nations. The main Chinese ports are Shanghai, Lü-ta, Tientsin, Tsingtao, and Ch'in-huang-tao; the main South Korean port is Inch'ŏn, the outport for Seoul; and that for North Korea is Namp'o, the outport for P'yŏngyang.

Study and exploration. The Yellow Sea has been studied fairly extensively in the 20th century, but its geologic, physical, and biological properties are incompletely understood. There is, for example, no comprehensive report on the sea's fisheries, despite the awareness of stock depletion. There has been some cooperation between nations—China began joint oil explorations with foreign oil companies in 1979—but the political tension in the region since World War II has inhibited the sharing of scientific information. Thus, it has been difficult to evaluate the state of the sea's marine environment. (M.U./M.J.V.)

BIBLIOGRAPHY

General works: Comprehensive summaries of information about the Pacific Ocean and its seas are found in ALASTAIR COUPER (ed.), *The Times Atlas and Encyclopaedia of the Sea,* 2nd ed. (1989), with excellent maps and text covering the ocean environment, resources, ocean trade, strategic use of the seas, law of the sea, problems of pollution, and scientific research; and MARTYN BRAMWELL (ed.), *The Rand McNally Atlas of the Oceans* (1977, reprinted 1987), with specific sections on bathymetry, structural geology, reefs and islands circulation, mineral resources, and living resources. See also M. GRANT GROSS, *Oceanography, a View of the Earth,* 5th ed. (1990), offering topical rather than regional coverage, but with numerous examples based on the Pacific Ocean; RHODES W. FAIRBRIDGE (ed.), *The Encyclopedia of Oceanography* (1966), with articles on the Pacific and its seas; and the articles pertaining to the Pacific in the periodical *Oceanus* (quarterly), particularly vol. 32, no. 4 (Winter 1989/90), which is devoted to the Pacific.

Physical environment: OTIS W. FREEMAN, *Geography of the Pacific* (1951), surveys the physical features of the ocean; and ALAN E.M. NAIRN and FRANCIS G. STEHLI (eds.), *The Ocean Basins and Margins,* vol. 7, *The Pacific Ocean,* 2 vol. (1985–88), is comprehensive and highly detailed. ZVI BEN-AVRAHAM (ed.), *The Evolution of the Pacific Ocean Margins* (1989); and H.W. MENARD, *Marine Geology of the Pacific* (1964), explore marine geology and geophysics. See also BRUNO WAUTHY, *Physical Ocean Environment in the South Pacific Commission Area* (1986), a short but valuable summary of the oceanography of the central and South Pacific, prepared as a regional report under the United Nations Environment Programme.

Economic aspects: A detailed review of Pacific Ocean minerals, including those contained in seawater and those on or beneath the ocean floor, is offered in G.P. GLASBY, "Marine Minerals in the Pacific," *Oceanography and Marine Biology* 24:1–64 (1986). Information on economic and other resources of the ocean areas is also found in such works as *Asia & Oceania,* vol. 4 of the *Worldmark Encyclopedia of the Nations,* 7th ed., 5 vol. (1988); FILLMORE C.F. FARNEY, *Marine Mineral Resources* (1990); and *FAO Yearbook: Fishery Statistics.* JOHN P. CRAVEN, *The Management of Pacific Marine Resources* (1982), offers well-written, scholarly coverage of resources and their management, or in some cases mismanagement; details on fisheries, shipping, scientific research, and maritime law are given in EDWARD MILES *et al.,* *The Management of Marine Regions: The North Pacific* (1982); and PETER N. NEMETZ (ed.), *The Pacific Rim* (1987), studies commerce, economic conditions, foreign investments, and resources in the Pacific Rim countries. Another valuable overview of economic conditions, international business enterprises, marine resources, and regionalism is ENRI-JACOB KOLDE, *The Pacific Quest: The Concept and Scope of an Oceanic Community* (1976); and HENRY S. ALBINSKI *et al.,* *The South Pacific* (1989), studies the "New Pacific" and the economic, political, and social problems of island nations that have achieved independence since World War II.

Study and exploration: ANDREW SHARP, *The Discovery of the Pacific Islands* (1960, reprinted 1985), summarizes the work of 121 explorers of the area and their contributions to charting the Pacific. J.C. BEAGLEHOLE, *The Exploration of the Pacific,* 3rd ed. (1966), and *The Life of Captain James Cook* (1974), are classics on the subject. ERNEST S. DODGE, *Beyond the Capes: Pacific Exploration from Captain Cook to the Challenger, 1776–1877* (1971), provides a thorough coverage of this important period. (Jo.E.B./J.R.Mo.)

Marginal seas: For the Bering Sea, see P.L. BEZRUKOV (ed.), *Geographical Description of the Bering Sea: Bottom Relief and Sediments,* trans. from Russian (1964); DAVID M. HOPKINS (ed.), *The Bering Land Bridge* (1967); A.P. LISITSYN, *Recent Sedimentation in the Bering Sea* (1969; originally published in Russian, 1966); L.K. COACHMAN, K. AAGAARD, AND R.B. TRIPP, *Bering Strait: The Regional Physical Oceanography* (1975); and relevant sections in YVONNE HERMAN (ed.), *Marine Geology and Oceanography of the Arctic Seas* (1974). Materials on the Sea of Okhotsk are published primarily in Russian, although English-language articles can be found in *Oceanus.* (A.V.A./K.T.B.)

For the East and South China seas, see DENNIS E. HAYES (ed.), *The Tectonic and Geologic Evolution of Southeast Asian Seas and Islands,* 2 vol. (1980–83); JOHN M. WAGEMAN, THOMAS W.C. HILDE, and K.O. EMERY, "Structural Framework of East China Sea and Yellow Sea," *American Association of Petroleum Geologists Bulletin* 54(9):1611–1643 (1970); CHOON-HO PARK, *East Asia and the Law of the Sea* (1983); and DOUGLAS M. JOHNSTON, *Environmental Management in the South China Sea: Legal and Institutional Developments* (1982). The periodical *Progress in Oceanography* (eight times a year) includes, among other relevant writings, selections of papers presented at the workshops of the Japan and East China Seas Study group, which take place from time to time in Japan. (E.C.LaF.)

On the Sea of Japan, see MARK J. VALENCIA (ed.), *International Conference on the Sea of Japan* (1989), which discusses oceanography, fisheries, petroleum and minerals, trade and shipping, and environment; and GLENN T. TREWARTHA, *Japan: A Geography* (1965), which offers useful information on the climatic influence of the Sea of Japan. The Sea of Japan and other regions of the northwest Pacific are discussed in S. CHIKUNI, *The Fish Resources of the Northwest Pacific* (1985); and maritime policies are addressed in TSUNEO AKAHA, *Japan in Global Ocean Politics* (1985). (J.R.Mo.)

Much of the information on the Yellow Sea is published in Chinese. Useful English-language works include Y.A. PARK, O.H. PILKEY, and S.W. KIM (eds.), *Proceedings of Korea–U.S. Seminar and Workshop of Marine Geology and Physical Processes of the Yellow Sea* (1984); and MARK J. VALENCIA (ed.), *International Conference on the Yellow Sea* (1987). (M.J.V.)

The Art of Painting

The art of painting is the expression of ideas and emotions, with the creation of certain aesthetic qualities, in a two-dimensional visual language. The elements of this language—its shapes, lines, colours, tones, and textures—are used in various ways to produce sensations of volume, space, movement, and light on a flat surface. These elements are combined into expressive patterns in order to represent real or supernatural phenomena, to interpret a narrative theme, or to create wholly abstract visual relationships. An artist's decision to use a particular medium, such as tempera, fresco, oil, watercolour, ink, gouache, encaustic, or casein, as well as the choice of a particular form, such as mural, easel, panel, miniature, manuscript illumination, scroll, screen or fan, panorama, or any of a variety of modern forms, is based on the sensuous qualities and the expressive possibilities and limitations of those options. The choices of the medium and the form, as well as the artist's own technique, combine to realize a unique visual image.

Earlier cultural traditions—of tribes, religions, guilds, royal courts, and states—largely controlled the craft, form, imagery, and subject matter of painting and determined its function, whether ritualistic, devotional, decorative, entertaining, or educational. Painters were employed more as skilled artisans than as creative artists. Later the Far East and Renaissance Europe saw the emergence of the fine artist, with the social status of scholar and courtier, who signed his work, who decided its design and often its subject and imagery, and who established a more personal—if not always amicable—relationship with his patron.

During the 19th century painters in Western societies began to lose their social position and secure patronage. Some artists countered the decline in patronage support by holding their own exhibitions and charging an entrance fee. Others earned an income through touring exhibitions of their work. The need to appeal to a marketplace had replaced the similar (if less impersonal) demands of patronage, and its effect on the art itself was probably similar as well. Generally, artists can now reach an audience only through commercial galleries and public museums, although their work may be occasionally reproduced in art periodicals. They may also be assisted by financial awards or commissions from industry and the state. They have, however, gained the freedom to invent their own visual language and to experiment with new forms and unconventional materials and techniques. For example, some painters have combined other media, such as sculpture, with painting to produce three-dimensional abstract designs. Other artists have attached real objects to the canvas in collage fashion or used electricity to operate coloured kinetic panels and boxes. Conceptual artists frequently express their ideas in the form of a proposal for an unrealizable project, while performance artists are an integral part of their own compositions. The restless endeavour to extend the boundaries of expression in Western art produces continuous international stylistic changes. The often bewildering succession of new movements in painting is further stimulated by the swift interchange of ideas by means of international art journals, travelling exhibitions, and art centres.

This article is concerned with the elements and principles of design in painting and with the various mediums, forms, imagery, subject matter, and symbolism employed or adopted or created by the painter. For the history of painting in the West, see PAINTING, THE HISTORY OF WESTERN. For treatment of painting as practiced in non-European cultures, see AFRICAN ARTS; AMERICAN PEOPLES, ARTS OF NATIVE; CENTRAL ASIAN ARTS; EAST ASIAN ARTS; EGYPTIAN ARTS AND ARCHITECTURE, ANCIENT; FOLK ARTS; ISLĀMIC ARTS; OCEANIC ARTS; PREHISTORIC PEOPLES AND CULTURES; SOUTH ASIAN ARTS; and SOUTHEAST ASIAN ARTS. For the conservation and restoration of paintings on canvas, wall paintings, paintings on paper, and paintings on ivory, see ART CONSERVATION AND RESTORATION. For a discussion of the philosophy of art, see AESTHETICS and PHILOSOPHIES OF THE BRANCHES OF KNOWLEDGE. For the history of critical reactions to painting from antiquity to the turn of the 21st century, see ART CRITICISM, THE HISTORY OF. For a discussion of the role of painting and other arts in religion, as well as of the use of religious symbols in art, see RELIGIOUS SYMBOLISM AND ICONOGRAPHY. For the methods employed in the display of art, see MUSEUMS. For more details on specific types of painting, such as lacquerwork, frescoes, and murals, and their use in design, see DECORATIVE ARTS AND FURNISHINGS. For information on other arts related to painting, see such articles as DRAWING and PRINTMAKING.

For further coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, section 628, and the *Index*.

The article is divided into the following sections:

Elements and principles of design 318
  Elements of design
  Principles of design
  Design relationships between painting and other visual arts
Techniques and methods 324
Mediums 325
  Tempera
  Fresco
  Oil
  Watercolour
  Ink
  Gouache
  Encaustic
  Casein
  Synthetic mediums
  Other mediums
  Mixed mediums
Forms of painting 329
  Mural painting
  Easel and panel painting
  Miniature painting
  Manuscript illumination and related forms
  Scroll painting
  Screen and fan painting
  Panoramas
  Modern forms
Imagery and subject matter 331
  Kinds of imagery
  Kinds of subject matter
Symbolism 333
Bibliography 334

ELEMENTS AND PRINCIPLES OF DESIGN

The design of a painting is its visual format: the arrangement of its lines, shapes, colours, tones, and textures into an expressive pattern. It is the sense of inevitability in this formal organization that gives a great painting its self-sufficiency and presence.

Formal interplay of colours and shapes

The colours and placing of the principal images in a design may be sometimes largely decided by representational and symbolic considerations. Yet it is the formal interplay of colours and shapes that alone is capable of communicating a particular mood, producing optical sensations of space, volume, movement, and light and creating forces

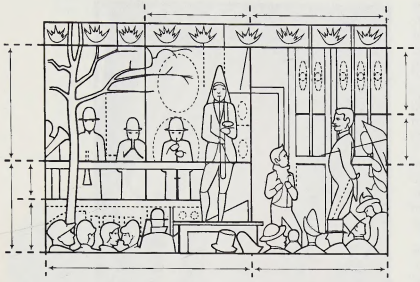








Figure 1: *Linear design.*
(Top) The linear design of Georges Seurat's oil painting "La Parade," 1887–88, composed of ovals and rectangles arranged in an overall grid pattern, based on the Golden Mean system of proportions. In the Metropolitan Museum of Art, New York City. 1 m × 1.5 m. (Bottom) The interwoven, linear pattern of Leonardo da Vinci's panel painting "Virgin and Child with St. Anne," *c.* 1501–12. In the Louvre, Paris. 1.68 m × 1.3 m.
By courtesy of (top right) the Metropolitan Museum of Art, New York, bequest of Stephen C. Clark, 1961; photograph, (bottom right) Giraudon—Art Resource

of both harmony and tension, even when a painting's narrative symbolism is obscure.

Elements of design. *Line.* Each of the design elements has special expressive qualities. Line, for example, is an intuitive, primeval convention for representing things; the simple linear imagery of young children's drawings and prehistoric rock paintings is universally understood. The formal relationships of thick with thin lines, of broken with continuous, and of sinuous with jagged are forces of contrast and repetition in the design of many paintings in all periods of history. Variations in the painted contours of images also provide a direct method of describing the volume, weight, spatial position, and textural characteristics of things. The finest examples of this pictorial shorthand are found in Japanese ink painting, where an expressive economy and vitality of line is closely linked to a traditional mastery of calligraphy.

In addition to painted contours, a linear design is composed of all of the edges of tone and colour masses, of the axial directions of images, and of the lines that are implied by alignments of shapes across the picture. The manner in which these various kinds of line are echoed and repeated animates the design (Figure 1). The artist, whether acting consciously or intuitively, also places them in relationship to one another across the picture, so that they weave a unifying rhythmic network throughout the painting (Figure 1).

Apart from the obvious associations of some linear patterns with particular actions—undulating lines suggesting buoyant movement, for instance—emotive sensations are produced by certain linear relationships. Thus, lines moving upward express feelings of joy and aspiration, those directing the eye downward evoke moods of sadness or defeat, while lines at angles opening to the right of a design are more agreeable and welcoming than those spreading outward to the left.

Shape and mass. Shape and mass, as elements of design, include all areas of different colour, tone, and texture, as well as individual and grouped images.

Children instinctively represent the things they see by geometrical symbols. Not only have sophisticated modern artists, such as Paul Klee and Jean Dubuffet, borrowed this primitive imagery, but the more arresting and expressive shapes and masses in most styles of painting and those to which most people intuitively respond will generally be found to have been clearly based on such archetypal forms. A square or a circle will tend to dominate a design and will therefore often be found at its focal centre—the square window framing Christ in Leonardo da Vinci's "Last Supper," for example, the hovering "sun" in an Adolph Gottlieb abstract, or the halo encircling a Chris-

Response to archetypal forms

tian or Buddhist deity. A firmly based triangular image or group of shapes seems reassuring, even uplifting, while the precarious balance implied by an inverted triangular shape or mass produces feelings of tension. Oval, lozenge, and rectangular forms suggest stability and protection and often surround vulnerable figures in narrative paintings.

There is generally a cellular unity, or "family likeness," between the shapes and masses in a design similar to the visual harmony of all units to the whole observed in natural forms—the gills, fins, and scales in character with the overall shape of a fish, for example.

The negative spaces between shapes and masses are also carefully considered by the artist, since they can be so adjusted as to enhance the action and character of the positive images. They can be as important to the design as time intervals in music or the voids of an architectural facade.

Colour. In many styles and periods of painting, the functions of colour are primarily decorative and descriptive, often serving merely to reinforce the expression of an idea or subject communicated essentially in terms of line and tone. In much of modern painting, however, the full-spectrum range of pigments available has allowed colour to be the primary expressive element.

The principal dimensions of colour in painting are the variables or attributes of hue, tone, and intensity. Red, yellow, and blue are the basic hues from which all others on the chromatic scale can be made by mixtures. These three opaque hues are the subtractive pigment primaries and should not be confused with the behaviour of the additive triads and mixtures of transparent, coloured light. Mixtures of primary pairs produce the secondary hues of orange, violet, and green. By increasing the amount of one primary in each of these mixtures, the tertiary colours of yellow-orange, orange-red, red-violet, violet-blue, blue-green, and green-yellow, respectively, are made. The primary colours, with their basic secondary and tertiary mixtures, can be usefully notated as the 12 segments of a circle (Figure 2). The secondary and tertiary colour segments between a pair of parent primaries can then be seen to share a harmonious family relationship with one another—the yellow-orange, orange, and orange-red hues that lie between yellow and red, for example.

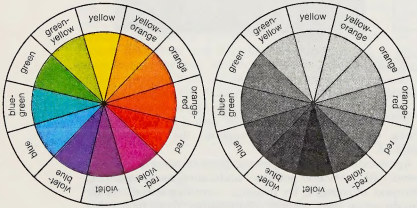


Figure 2: *Colour.*
(Left) Colour wheel made up of the primary colours and their basic secondary and tertiary mixtures. (Right) Colour wheel with approximate, inherent tonal values.

Local hues are the inherent and associative colours of things. In everyday life, familiar things are described by particular colours, and these often are identified by reference to familiar things; the green of grass and the grass green of paint, for instance. Although, as the Impressionists demonstrated, the inherent colours of forms in the real world are usually changed by effects of light and atmosphere, many of the great primitive and classical styles of representational painting are expressed in terms of local hues.

Tonal pattern of a painting

Tone is a colour's relative degree, or value, of lightness or darkness. The tonal pattern of a painting is shown in a monochrome reproduction. A painting dominated by dark colours, such as a Rembrandt, is in a low tonal key, while one painted in the pale range of a late Claude Monet is said to be high keyed. The tonal range of pigments is too



Figure 3: An example of the early oil method of (left) colour glazing a (right) monochrome painting.

narrow for the painter to be able to match the brightest lights and deepest darks of nature. Therefore, in order to express effects of illumination and dense shadow, he must lower the overall tonal key of his design, thus intensifying the brightness value of his lightest pigment colours.

The Greco-Roman, Renaissance, and Neoclassical method of representing volume and space in painting was by a system of notated tonal values, the direction of each plane in the design being indicated by a particular degree of lightness or darkness. Each tonal value was determined by the angle at which a plane was meant to appear to turn away from an imaginary source of light. The tonal modelling, or shading, of forms was often first completed in a monochrome underpainting. This was then coloured with transparent washes of local hues, a technique similar to that of colour tinting a black-and-white photograph (Figure 3).

Each hue has an intrinsic tonal value in relation to others on the chromatic scale; orange is inherently lighter than red, for instance, and violet is darker than green (Figure 2). Any reversal of this natural tonal order creates a colour discord. An optical shock is therefore produced when orange is juxtaposed with pink (a lighter tone of red) or pale violet is placed against dark green. Such contrasts as these are deliberately created in paintings for the purpose of achieving these dramatic and disturbing effects.

The intensity of a colour is its degree of purity or hue saturation. The colour of a geranium, therefore, is said to be more intense, more highly saturated with pure orange-red than is mahogany. The pigment vermilion is orange-red at maximum intensity; the brown earth pigment burnt sienna is grayer and has a lower degree of orange-red saturation.

Intense hues are termed chromatic colours. The achromatic range is made up of hues reduced in intensity by the addition of white, making the tints, or pastel colours, such as cream and pink; or of black, producing the shades, or earth colours, such as mustard and moss green; or of both white and black, creating the neutralized hues, or colour-tinged grays, such as oatmeal and charcoal.

An achromatic colour will seem more intense if it is surrounded by neutralized hues or juxtaposed with its complementary colour. Complementaries are colour op-

posites. The complementary colour to one of the primary hues is the mixture of the other two; the complementary to red pigment, for example, is green—that is, blue mixed with yellow. The colour wheel (Figure 2) shows that the tertiaries also have their colour opposites, the complementary to orange-red, for instance, being blue-green. Under clear light the complementary to any chroma, shade, or tint can be seen if one "fixates," or stares at, one colour intently for a few seconds then looks at a neutral, preferably white, surface. The colour afterimage will appear to glow on the neutral surface (Figure 4). Mutual enhancement

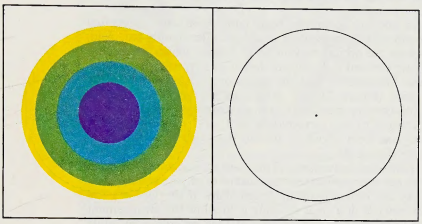


Figure 4: Coloured afterimages: if a person stares for about 30 seconds at the coloured disk under a clear light and then fixes upon the empty space of the adjacent circle, coloured afterimages will appear.

Enhancement of colour intensity

of colour intensity results from juxtaposing a complementary pair, red becoming more intensely red, for instance, and green more fiercely green when these are contiguous than either would appear if surrounded by harmonious hues. The 19th-century physicist Michel-Eugène Chevreul referred to this mutual exaltation of opposites as the law of simultaneous contrast. Chevreul's second law, of successive contrast, referred to the optical sensation that a complementary colour halo appears gradually to surround an intense hue. This complementary glow is superimposed on surrounding weaker colours, a gray becoming greenish when juxtaposed with red, reddish in close relationship with green, yellowish against violet, and so on (Figure 5).

Hues containing a high proportion of blue (the violet to green range) appear cooler than those with a high content of yellow or red (the green-yellow to red-violet range). This difference in the temperature of hues in a particular painting is, of course, relative to the range and juxtaposition of colours in the design. A green will appear cool if surrounded by intense yellow, while it will seem warm against blue-green (Figure 5). The optical tendency for warm colours to advance before cold had been long exploited by European and Oriental painters as a method of suggesting spatial depth. Changes in temperature and intensity can be observed in the atmospheric effects of nature, where the colours of distant forms become cooler, grayer, and bluish, while foreground planes and features appear more intense and usually warmer in colour.

The apparent changes in a hue as it passes through zones of different colour has enabled painters in many periods to create the illusion of having employed a wide range of pigment hues with, in fact, the use of very few. And, although painters had applied many of the optical principles of colour behaviour intuitively in the past, the publication of research findings by Chevreul and others stimulated the Neo-Impressionists and Post-Impressionists and the later Orphist and Op Art painters to extend systematically the expressive possibilities of these principles in order to create illusions of volume and space and vibrating sensations of light and movement. Paul Cézanne, for example, demonstrated that subtle changes in the surface of a form and in its spatial relationship to others could be expressed primarily in facets of colour, modulated by varying degrees of tone, intensity, and temperature and by the introduction of complementary colour accents.

While the often-complex religious and cultural colour symbologies may be understood by very few, the emotional response to certain colour combinations appears to be almost universal. Optical harmonies and discords seem to affect everyone in the same way, if in varying degrees. Thus, an image repeated in different schemes of colour will express a different mood in each change (Figure 6).

Texture. Pointillism was a term given to the Neo-Impressionist system of representing the shimmer of atmospheric light with spots of coloured pigment. This technique produced an overall granular texture. As an element of design, texture includes all areas of a painting enriched or animated by vibrating patterns of lines, shapes, tones, and colours, in addition to the tactile textures created by the plastic qualities of certain mediums. Decorative textures may be of geometrical repeat patterns, as in much of Indian, Islāmic, and medieval European painting and other art, or of representations of patterns in nature, such as scattered leaves, falling snow, and flights of birds.

Perceptual and conceptual levels of understanding spatial relationships

Volume and space. The perceptual and conceptual methods of representing volume and space on the flat surface of a painting are related to the two levels of understanding spatial relationships in everyday life.

Perceptual space is the view of things at a particular time and from a fixed position. This is the stationary window view recorded by the camera and represented in the later periods of ancient Greek and Roman paintings and in most Western schools of painting since the Renaissance. Illusions of perceptual space are generally created by use of the linear perspectival system, based on the observations that objects appear to the eye to shrink and parallel lines and planes to converge as they approach the horizon, or viewer's eye level (Figure 7).

Young children and primitive artists, however, do not understand space in this way and represent it conceptually. Their paintings, therefore, show objects and surroundings independently of one another and from the views that best present their most characteristic features. The notion of scale in their pictures is also subjective, the relative size of things being decided by the artist either by their degree of emotional significance for him or by their narrative importance in the picture (interest perspective).

The conceptual, polydimensional representation of space has been used at some period in most cultures. In much of ancient Egyptian and Cretan painting, for example, the head and legs of a figure were shown in profile, but the eye and torso were drawn frontally. And in Indian,

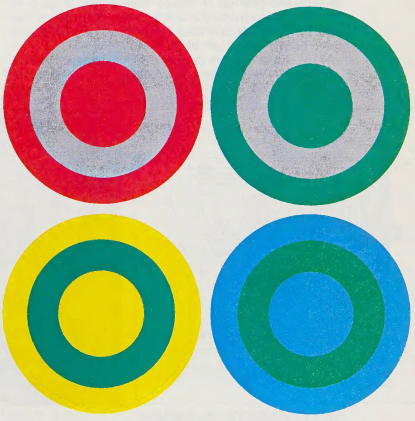


Figure 5: *Optical colour change.*
(Top) By complementary action, the same gray pigment will appear greenish when adjacent to red but reddish if adjacent to green. (Bottom) A green hue will seem cool if surrounded by yellow but warm when surrounded by blue-green.

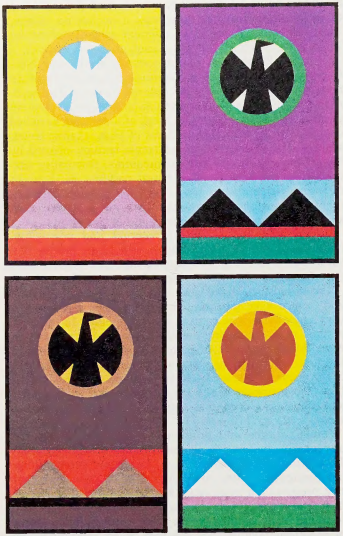


Figure 6: Emotive colour relationships: an identical pattern of shapes may express a different emotional mood through each colour variation.

Islāmic, and pre-Renaissance European painting, vertical forms and surfaces were represented by their most informative elevation view (as if seen from ground level), while the horizontal planes on which they stood were shown in isometric plan (as if viewed from above) (Figure 7). This system produces the overall effect that objects and their surroundings have been compressed within a shallow space behind the picture plane.

By the end of the 19th century Cézanne had flattened the conventional Renaissance picture space (Figure 7), tilting horizontal planes so that they appeared to push vertical forms and surfaces forward from the picture plane and toward the spectator (Figure 8). This illusion of the picture surface as an integrated structure in projecting low relief was developed further in the early 20th century by the Cubists. The conceptual, rotary perspective of a Cubist painting shows not only the components of things from different viewpoints but presents every plane of an object and its immediate surroundings simultaneously. This gives the composite impression of things in space that is gained by having examined their surfaces and construction from every angle.

In modern painting, both conceptual and perceptual methods of representing space are often combined. And, where the orbital movement of forms—which has been a basic element in European design since the Renaissance—was intended to hold the spectator's attention within the frame (Figure 7), the expanding picture space in late 20th-century mural-size abstract paintings directs the eye outward to the surrounding wall, and their shapes and colours seem about to invade the observer's own territory (Figure 9).

Time and movement. Time and movement in painting are not restricted to representations of physical energy, but they are elements of all design. Part of the viewer's full experience of a great painting is to allow the arrangement of lines, shapes, and accents of tone or colour to guide the eye across the picture surface at controlled tempos and rhythmic directions. These arrangements contribute overall to the expression of a particular mood, vision, and idea.

Kinetic sensations on a flat surface

Centuries before cinematography, painters attempted to produce kinetic sensations on a flat surface. A mural of 2000 BC in an Egyptian tomb at Beni Hasan, for instance, is designed as a continuous strip sequence of wrestling holds and throws, so accurately articulated and notated that it might be photographed as an animated film cartoon. The gradual unrolling of a 12th-century Japanese hand scroll produces the visual sensation of a helicopter flight along a river valley, while the experience of walking to the end of a long, processional Renaissance mural by Andrea Mantegna or Benozzo Gozzoli is similar to that of having witnessed a passing pageant as a standing spectator.

In the Eastern and Western narrative convention of continuous representation, various incidents in a story were depicted together within one design, the chief characters in the drama easily identified as they reappeared in different situations and settings throughout the painting. In Byzantine murals and in Indian and medieval manuscript paintings, narrative sequences were depicted in grid patterns, each "compartment" of the design representing a

(Centre) Copyright reserved

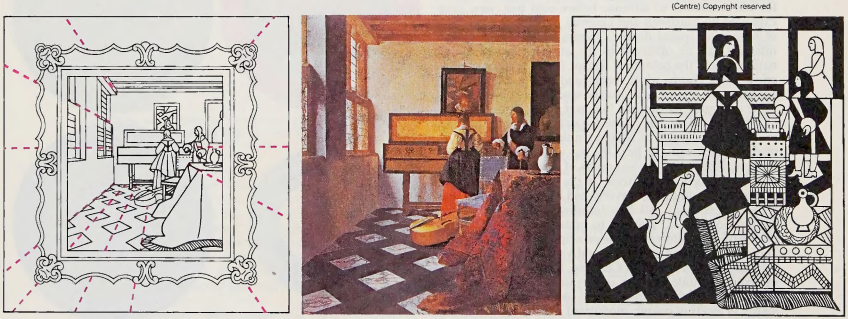


Figure 7: *Perceptual and conceptual space.*
(Left) Perceptual space: the illusion of an interior space as it might be recorded by a camera and as it is represented by the optical laws of Renaissance perspective. (Centre) Example based on Jan Vermeer's oil painting "The Music Lesson," c. 1660. In Buckingham Palace, London. 74 cm × 64 cm. (Right) Conceptual space: how the same subject might be represented by polydimensional and "interest" perspectival systems.

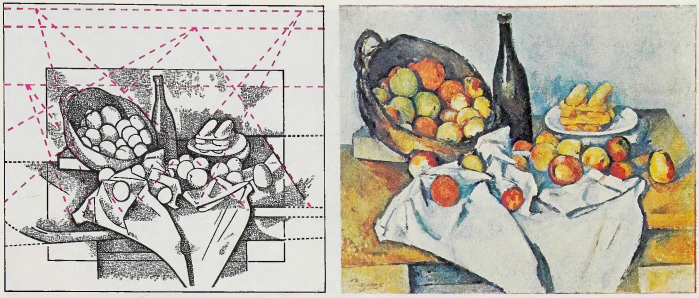


Figure 8: *Simultaneous viewpoints.*
(Left) Paul Cézanne's method of representing things in a painting as if seen from different directions and at varying eye levels over a period of time. In contrast to the receding planes of Vermeer's design (Figure 7), the illusion is created, by tone and colour contrasts, that planes and forms are advancing toward the spectator. (Right) Example based on Cézanne's oil painting "The Basket of Apples," 1890–94. In the Art Institute of Chicago. 65 cm × 81 cm.
(Right) By courtesy of the Art Institute of Chicago, Helen Birch Bartlett Memorial Collection

visual chapter in a religious story or a mythological or historical epic.

The Cubists aimed to give the viewer the time experience of moving around static forms in order to examine their volume and structure and their relationships to the space surrounding them. In paintings such as "Nude Descending a Staircase," "Girl Running on a Balcony," and "Dog on Leash," Marcel Duchamp and Giacomo Balla combined the Cubist technique of projected, interlocking planes with the superimposed time-motion sequences of cinematography. This technique enabled the artists to analyze the structural mechanics of forms, which are represented as moving in space past the viewer.

Principles of design. Because painting is a two-dimensional art, the flat pattern of lines and shapes is an important aspect of design, even for those painters concerned with creating illusions of great depth. And, since any mark made on the painting surface can be perceived as a spatial statement—for it rests upon it—there are also qualities of three-dimensional design in paintings composed primarily of flat shapes. Shapes in a painting, therefore, may be balanced with one another as units of a flat pattern and considered at the same time as components in a spatial design, balanced one behind another. A symmetrical balance of tone and colour masses of equal weight creates a serene and sometimes monumental design, while a more dynamic effect is created by an asymmetrical balance.

Geometrical shapes and masses are often the basic units in the design of both "flat patterns," such as Byzantine and Islāmic paintings, and "sculptural compositions," such as Baroque and Neoclassical figure tableaux. The flat, overlapping squares, circles, and triangles that create the pattern of a Romanesque mural, for example, become the interlocking cubic, spherical, and pyramidal components that enclose the grouped figures and surrounding features in a Renaissance or a Neoclassical composition (Figure 10).

An emphasis upon the proportion of the parts to the whole is a characteristic of Classical styles of painting. The Golden Mean, or Section, has been used as an ideal proportion on which to base the framework of lines and shapes in the design of a painting. The Renaissance mathematician Lucas Pacioli defined this aesthetically satisfying ratio as the division of a line so that the shorter part is to the longer as the longer is to the whole (approximately 8 to 13). His treatise (*Divina proportione*) influenced Leonardo da Vinci and Albrecht Dürer. The Neo-Impressionists Georges Seurat (Figure 1) and Paul Signac based the linear pattern of many of their compositions upon the principle of this "divine proportion." Golden Mean proportions can be discovered in the design of many other styles of painting, although often they may have been created more by intuitive judgment than by calculated measurement.

Tension in paintings

Tension is created in paintings, as it is experienced in everyday life, by the anticipation of an event or by an unexpected change in the order of things. Optical and psychological tensions occur in passages of a design, therefore, when lines or shapes almost touch or seem about to collide, when a harmonious colour progression is interrupted by a sudden discord, or when an asymmetrical balance of lines, shapes, tones, or colours is barely held.

Contrasts in line, shape, tone, and colour create vitality; rectilinear shapes played against curvilinear, for instance, or warm colours against cool. Or a painting may be composed in contrasted overall patterns, superimposed in counterpoint to one another—a colour scheme laid across contrasting patterns of lines and tones, for example (Figure 11).

Design relationships between painting and other visual arts. The philosophy and spirit of a particular period in painting usually have been reflected in many of its other visual arts. The ideas and aspirations of the ancient cultures, of the Renaissance, Baroque, Rococo, and Neoclassical periods of Western art and, more recently, of the 19th-century Art Nouveau and Secessionist movements were expressed in much of the architecture, interior design, furniture, textiles, ceramics, dress design, and handicrafts, as well as in the fine arts, of their times. Following the Industrial Revolution, with the redundancy of handcraft-

By courtesy of the trustees of the Tate Gallery, London; photograph, A.C. Cooper Ltd.

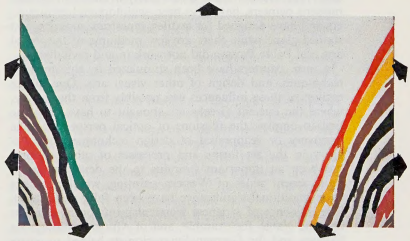


Figure 9: Opening out the picture space: the movement of shapes in Morris Louis' acrylic painting "Alpha-Phi," 1961, directs the spectator's eye outside the picture surface. In the Tate Gallery, London. 2.59 m × 4.58 m. In contrast, Vermeer's design (Figure 7) is contrived to hold the spectator's attention within the frame.



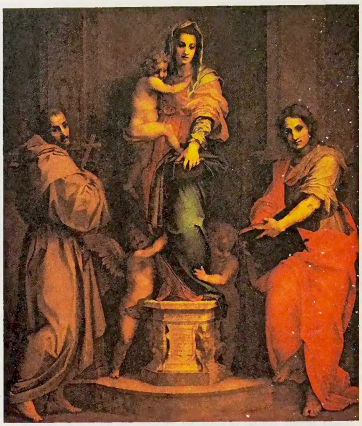


Figure 10: *Principles of design.*
(Left) The flat pattern design of the anonymous Spanish panel painting "Virgin and Child," 12th century. In the Museo Arqueológico Artístico Episcopal, Vich, Spain. (Right) The three-dimensional design of interlocking pyramids of Andrea del Sarto's panel painting "Madonna of the Harpies," 1517. In the Uffizi, Florence. 2.07 m × 1.78 m.
(Left) Archivo Mas, Barcelona, (right) SCALA—Art Resource

manship and the loss of direct communication between the fine artist and society, idealistic efforts to unite the arts and crafts in service to the community were made by William Morris in Victorian England and by the Bauhaus in 20th-century Germany. Although their aims were not fully realized, their influences, like those of the short-lived de Stijl and Constructivist movements, have been far-reaching, particularly in architectural, furniture, and typographic design.

Michelangelo and Leonardo da Vinci were painters, sculptors, and architects. Although no artists since have excelled in so wide a range of creative design, leading 20th-century painters have expressed their ideas in many other mediums. In graphic design, for example, Pierre Bonnard, Henri Matisse, and Raoul Dufy produced posters and illustrated books; André Derain, Fernand Léger, Marc Chagall, Mikhail Larionov, Robert Rauschenberg, and David Hockney designed for the theatre; Joan Miró, Georges Braque, and Chagall worked in ceramics; Braque and Salvador Dalí designed jewelry; and Dalí, Hans Richter, and Andy Warhol made films. Many of these, with other modern painters, have also been sculptors and printmakers and have designed for textiles, tapestries, mosaics, and stained glass, while there are few mediums of the visual arts that Pablo Picasso did not work in and revitalize.

In turn, painters have been stimulated by the imagery, techniques, and design of other visual arts. One of the earliest of these influences was possibly from the theatre, where the ancient Greeks are thought to have been the first to employ the illusions of optical perspective. The discovery or reappraisal of design techniques and imagery in the art forms and processes of other cultures has been an important stimulus to the development of more recent styles of Western painting, whether or not their traditional significance have been fully understood. The influence of Japanese woodcut prints on Synthetism and the Nabis, for example, and of African sculpture on Cubism and the German Expressionists helped to create visual vocabularies and syntax with which to express new visions and ideas. The invention of photography introduced painters to new aspects of nature, while eventually prompting others to abandon representational painting altogether. Painters of everyday life, such as Edgar Degas, Henri de Toulouse-Lautrec, Édouard Vuillard, and Bonnard, exploited the design innovations of camera cutoffs, close-ups, and unconventional viewpoints in order to give the spectator the sensation of sharing an intimate picture space with the figures and objects in the painting.

Influence of photography

TECHNIQUES AND METHODS

Whether a painting reached completion by careful stages or was executed directly by a hit-or-miss alla prima method (in which pigments are laid on in a single application) was once largely determined by the ideals and established techniques of its cultural tradition. For example, the medieval European illuminator's painstaking procedure, by which a complex linear pattern was gradually enriched with gold leaf and precious pigments, was contemporary with the Sung Chinese Zen practice of immediate, calligraphic brush painting, following a contemplative period of spiritual self-preparation. More recently, the artist has decided the technique and working method best suited to his aims and temperament. In France in the 1880s, for instance, Seurat might be working in his studio on drawings, tone studies, and colour schemes in preparation for a large composition at the same time that, outdoors, Monet was endeavouring to capture the effects of afternoon light and atmosphere, while Cézanne analyzed the structure of the Montagne (mountain) Sainte-Victoire with deliberated brush strokes, laid as irrevocably as mosaic tesserae (small pieces, such as marble or tile).

The kind of relationship established between artist and patron, the site and subject matter of a painting commission, and the physical properties of the medium employed may also dictate working procedure. Peter Paul Rubens, for example, followed the businesslike 17th-century custom of submitting a small oil sketch, or *modella,* for his client's approval before carrying out a large-scale commission. Siting problems peculiar to mural painting, such as spectator eye level and the scale, style, and function of a building interior, had first to be solved in preparatory drawings and sometimes with the use of wax figurines or scale models of the interior. Scale working drawings are essential to the speed and precision of execution demanded by quick-drying mediums, such as buon' fresco (see below *Mediums: Fresco*) on wet plaster and acrylic resin



Figure 11: *Visual counterpoint in design.*
(Left) Linear arabesques drawn across (centre) patterns of tone and colour. (Right) Example based on Henri de Toulouse-Lautrec's oil paint on cardboard painting "At the Nouveau Cirque: Five Stuffed Shirts," 1891. In the Philadelphia Museum of Art. 119 cm × 87 cm.
(Right) By courtesy of the Philadelphia Museum of Art, John D. Mcilhenny Fund

on canvas. The drawings traditionally are covered with a network of squares, or "squared-up," for enlarging on the surface of the support. Some modern painters prefer to outline the enlargement of a sketch projected directly onto the support by epidiascope (a projector for images of both opaque and transparent objects).

In Renaissance painters' workshops, pupil assistants not only ground and mixed the pigments and prepared the supports and painting surfaces but often laid in the outlines and broad masses of the painting from the master's design and studies.

Preservation of a painting

The inherent properties of its medium or the atmospheric conditions of its site may themselves preserve a painting. The wax solvent binder of encaustic paintings (in which after application, the paint is fixed by heat [see below *Mediums*], for example) both retains the intensity and tonality of the original colours and protects the surface from damp. And, while prehistoric rock paintings and buon' frescoes are preserved by natural chemical action, the tempera pigments thought to be bound only with water on many ancient Egyptian murals are protected by the dry atmosphere and unvarying temperature of the tombs. It has, however, been customary to varnish oil paintings, both to protect the surface against damage by dirt and handling and to restore the tonality lost when some darker pigments dry out into a higher key. Unfortunately, varnish tends to darken and yellow with time into the sometimes disastrously imitated "Old Masters' mellow patina." Once cherished, this amber-gravy film is now generally removed to reveal the colours in their original intensity. Glass began to replace varnish toward the end of the 19th century, when painters wished to retain the fresh, luminous finish of pigments applied directly to a pure white ground. The air-conditioning and temperature-control systems of modern museums make both varnishing and glazing unnecessary, except for older and more fragile exhibits.

The frames surrounding early altarpieces, icons, and cassone panels (painted panels on the chest used for a bride's household linen) were often structural parts of the support. With the introduction of portable easel pictures, heavy frames not only provided some protection against theft and damage but were considered an aesthetic enhancement to a painting, and frame making became a specialized craft. Gilded gesso moldings (consisting of plaster of paris and sizing that forms the surface for low relief) in extravagant swags of fruit and flowers certainly seem almost an extension of the restless, exuberant design of a Baroque or Rococo painting. A substantial frame also provided a proscenium (in a theatre, the area between the orchestra and the curtain) in which the picture was isolated from its immediate surroundings, thus adding to the window view illusion intended by the artist. Deep, ornate frames are unsuitable for many modern paintings, where the artist's intention is for his forms to appear to advance toward the spectator rather than be viewed by him as if through a wall aperture. In contemporary Minimal paintings, no effects of spatial illusionism are intended; and, in order to emphasize the physical shape of the support itself and to stress its flatness, these abstract, geometrical designs are displayed without frames or are merely edged with thin protective strips of wood or metal.

MEDIUMS

By technical definition, mediums are the liquids added to paints to bind them and make them workable. They are discussed here, however, in the wider meaning of all the various paints, tools, supports, surfaces, and techniques employed by painters. The basis of all paints is variously coloured pigment, ground to a fine powder. The different expressive capacities and characteristic final surface texture of each medium are determined by the vehicle with which it is bound and thinned, the nature and surface preparation of the support, and the tools and technique with which it is handled.

Sources of pigments

Pigments are derived from various natural and artificial sources. The oldest and most permanent pigments are the blacks, prepared from bone and charcoal, and the clay earths, such as raw umber and raw sienna, which can be changed by heating into darker, warmer browns. In early periods of painting, readily available pigments were few. Certain intense hues were obtainable only from the rarer minerals, such as cinnabar (orange-red vermilion), lapis lazuli (violet-blue ultramarine), and malachite (green). These were expensive and therefore reserved for focal accents and important symbolic features in the design. The opening of trade routes and the manufacture of synthetic substitutes gradually extended the range of colours available to painters.

Tempera. A tempera medium is dry pigment tempered with an emulsion and thinned with water. It is a very ancient medium, having been in constant use in most world cultures, until in Europe it was gradually superseded, during the Renaissance, by oil paints. Tempera was the original mural medium in the ancient dynasties of Egypt, Babylonia, Mycenean Greece, and China and was used to decorate the early Christian catacombs. It was employed on a variety of supports, from the stone stelae (or commemorative pillars), mummy cases, and papyrus rolls of ancient Egypt to the wood panels of Byzantine icons and altarpieces and the vellum leaves of medieval illuminated manuscripts.

True tempera is made by mixture with the yolk of fresh eggs, although manuscript illuminators often used egg white and some easel painters added the whole egg.

Other emulsions have been used, such as casein glue with linseed oil, egg yolk with gum and linseed oil, and egg white with linseed or poppy oil. Individual painters have experimented with other recipes, but few of these have proved successful; all but William Blake's later tempera paintings on copper sheets, for instance, have darkened and decayed, and it is thought that he mixed his pigment with carpenter's glue.

Distemper is a crude form of tempera made by mixing dry pigment into a paste with water, which is thinned with heated glue in working or by adding pigment to whiting, a mixture of fine-ground chalk and size. It is used for stage scenery and full-size preparatory cartoons for murals and tapestries. When dry, its colours have the pale, mat, powdery quality of pastels, with a similar tendency to smudge. Indeed, damaged cartoons have been retouched with pastel chalks.

Egg tempera is the most durable form of the medium, being generally unaffected by humidity and temperature. It dries quickly to form a tough film that acts as a protective skin to the support. In handling, in its diversity of transparent and opaque effects, and in the satin sheen of its finish, it resembles the modern acrylic resin emulsion paints.

Traditional process of tempera painting

Traditional tempera painting is a lengthy process. Its supports are smooth surfaces, such as planed wood, fine set plaster, stone, paper, vellum, canvas, and modern composition boards of compressed wood or paper. Linen is generally glued to the surface of panel supports, additional strips masking the seams between braced wood planks. Gesso, a mixture of plaster of paris (or gypsum) with size, is the traditional ground. The first layer is of gesso grosso, a mixture of coarse, unslaked plaster and size. This provides a rough, absorbent surface for ten or more thin coats of gesso sotile, a smooth mixture of size and fine plaster previously slaked in water to retard drying. This laborious preparation results, however, in an opaque, brilliant white, light-reflecting surface, similar in texture to hard, flat icing sugar.

The design for a large tempera painting traditionally was executed in distemper on a thick paper cartoon. The outlines were pricked with a perforating wheel so that when the cartoon was laid on the surface of the support, the linear pattern was transferred by dabbing, or "pouncing," the perforations with a muslin bag of powdered charcoal. The dotted contours traced through were then fixed in paint. Medieval tempera painters of panels and manuscripts made lavish use of gold leaf on backgrounds and for symbolic features, such as haloes and beams of heavenly light. Areas of the pounced design intended for gilding were first built up into low relief with gesso duro, the harder, less absorbent gesso compound used also for elaborate frame moldings. Background fields were often textured by impressing the gesso duro, before it set, with small, carved, intaglio wood blocks to create raised, pimpled, and quilted repeat patterns that glittered when gilded. Leaves of finely beaten gold were pressed onto a tacky mordant (adhesive compound) or over wet bole (reddish-brown earth pigment) that gave greater warmth and depth when the gilded areas were burnished.

Colours were applied with sable brushes in successive broad sweeps or washes of semitransparent tempera. These dried quickly, preventing the subtle tonal gradations possible with watercolour washes or oil paint; effects of shaded modelling had therefore to be obtained by a crosshatching technique of fine brush strokes. According to the Italian painter Cennino Cennini, the early Renaissance tempera painters laid the colour washes across a fully modelled monochrome underpainting in terre vert (olive-green pigment), a method developed later into the mixed mediums technique of tempera underpainting followed by transparent oil glazes.

The luminous gesso base of a tempera painting, combined with the accumulative effect of overlaid colour washes, produces a unique depth and intensity of colour. Tempera paints dry lighter in value, but their original tonality can be restored by subsequent waxing or varnishing. Other characteristic qualities of a tempera painting, resulting from its fast drying property and disciplined technique, are its steely lines and crisp edges, its meticulous detail and rich linear textures, and its overall emphasis upon a decorative flat pattern of bold colour masses.

The great Byzantine tradition of tempera painting was developed in Italy in the 13th and 14th centuries by Duccio di Buoninsegna and Giotto. Their flattened picture space, generously enriched by fields and textures of gold leaf, was extended by the Renaissance depth perspectives in the paintings of Giovanni Bellini, Piero della Francesca, Carlo Crivelli, Sandro Botticelli, and Vittore Carpaccio. By that time, oil painting was already challenging the primacy of tempera, Botticelli and some of his contemporaries apparently adding oil to the tempera emulsion or overglazing it in oil colour.

20th-century revival of tempera techniques

Following the supremacy of the oil medium during succeeding periods of Western painting, the 20th century saw a revival of tempera techniques by such U.S. artists as Ben Shahn, Andrew Wyeth, and George McNeil and by the British painter Edward Wadsworth. It would probably have been the medium also of the later hard-edge abstract painters had the new acrylic resin paints not proved more easily and quickly handled.

Fresco. Fresco (Italian: "fresh") is the traditional medium for painting directly onto a wall or ceiling. It is the oldest known painting medium, surviving in the prehistoric cave mural decorations and perfected in 16th-century Italy in the buon' fresco method.

The cave paintings are thought to date from about 20,000–15,000 BC. Their pigments probably have been preserved by a natural sinter process of rainwater seeping through the limestone rocks to produce saturated bicarbonate. The colours were rubbed across rock walls and ceilings with sharpened solid lumps of the natural earths (yellow, red, and brown ochre). Outlines were drawn with black sticks of wood charcoal. The discovery of mixing dishes suggests that liquid pigment mixed with fat was also used and smeared with the hand. The subtle tonal gradations of colour on animals painted in the Altamira and Lascaux caves appear to have been dabbed in two stages with fur pads, natural variations on the rock surface being exploited to assist in creating effects of volume. Feathers and frayed twigs may have been used in painting manes and tails.

These were not composite designs but separate scenes and individual studies that, like graffiti drawings, were added at different times, often one above another, by various artists. Paintings from the Magdalenian period (*c.* 10,000 BC) exhibit astonishing powers of accurate observation and ability to represent movement. Women, warriors, horses, bison, bulls, boars, and ibex are depicted in scenes of ritual ceremony, battle, and hunting. Among the earliest images are imprinted and stencilled hands. Vigorous meanders, or "macaroni" linear designs, were traced with fingers dipped in liquid pigment.

Fresco secco. In the fresco secco, or lime-painting, method, the plastered surface of a wall is soaked with slaked lime. Lime-resistant pigments are applied swiftly before the plaster sets. Secco colours dry lighter than their tone at the time of application, producing the pale, mat, chalky quality of a distempered wall. Although the pigments are fused with the surface, they are not completely absorbed and may flake in time, as in sections of Giotto's 14th-century S. Francesco murals at Assisi. Secco painting was the prevailing medieval and early Renaissance medium and was revived in 18th-century Europe by artists such as Giovanni Battista Tiepolo, François Boucher, and Jean-Honoré Fragonard.

Buon' fresco. Buon', or "true," fresco is the most durable method of painting murals, since the pigments are completely fused with a damp plaster ground to become an integral part of the wall surface. The stone or brick wall is first prepared with a brown trullisatio scratch coat, or rough-cast plaster layer. This is then covered by the arricciato coat, on which the linear design of the preparatory cartoon is pounced (see above *Tempera*) or engraved by impressing the outlines into the moist, soft plaster with a bone or metal stylus. These lines were usually overworked in reddish sinopia pigment. A thin layer of fine plaster is then evenly spread, allowing the linear design to show

Arricciato coat

through. Before this final intonaco ground sets, pigments thinned with water or slaked lime are applied rapidly with calf-hair and hog-bristle brushes; depth of colour is achieved by a succession of quick-drying glazes. Being prepared with slaked lime, the plaster becomes saturated with an aqueous solution of hydrate of lime, which takes up carbonic acid from the air as it soaks into the paint. Carbonate of lime is produced and acts as a permanent pigment binder. Pigment particles crystallize in the plaster, fusing it with the surface to produce the characteristic lustre of buon' fresco colours. When dry, these are mat and lighter in tone. Colours are restricted to the range of lime-resistant earth pigments. Mineral colours such as blue, affected by lime, are applied over earth pigment when the plaster is dry.

The intonaco coat is laid only across an area sufficient for painting before the plaster sets. The joins between each successive "day piece" are sometimes visible. Alterations must be made by immediate washing or scraping; minor retouching to set plaster is possible with casein or egg tempera, but major corrections necessitate breaking away the intonaco and replastering. The swift execution demanded stimulates bold designs in broad masses of colour with a calligraphic vitality of brush marks.

No ancient Greek buon' frescoes now exist, but forms of the technique survive in the Pompeian villas of the 1st century AD and earlier, in Chinese tombs at Liao-yang, Manchuria, and in the 6th-century Indian caves at Ajantā. Among the finest buon' fresco murals are those by Michelangelo in the Sistine Chapel and by Raphael in the Stanze of the Vatican. Other notable examples from the Italian Renaissance can be seen in Florence: painted by Andrea Orcagna in the Museo dell'Opera di Sta. Croce, by Gozzoli in the chapel of the Palazzo Medici-Riccardi, and by Domenico Ghirlandajo in the church of Sta. Maria Novella. Buon' fresco painting is unsuited to the damp, cold climate of northern countries, and there is now some concern for the preservation of frescoes in the sulfurous atmosphere of even many southern cities. Buon' fresco was successfully revived by the Mexican mural painters Diego Rivera, José Orozco, and Rufino Tamayo.

Sgraffito. Sgraffito (Italian *graffiare,* "to scratch") is a form of fresco painting for exterior walls. A rough plaster undercoat is followed by thin plaster layers, each stained with a different lime-fast colour. These coats are covered by a fine-grain mortar finishing surface. The plaster is then engraved with knives and gouges at different levels to reveal the various coloured layers beneath. The sintered-lime process binds the colours. The surface of modern sgraffito frescoes is often enriched with textures made by impressing nails and machine parts, combined with mosaics of stone, glass, plastic, and metal tesserae.

Sgraffito has been a traditional folk art in Europe since the Middle Ages and was practiced as a fine art in 13th-century Germany. It has been recently revived in northern Europe.

Oil. Oil paints are made by mixing dry pigment powder with refined linseed oil to a paste, which is then milled in order to disperse the pigment particles throughout the oil vehicle. According to the 1st-century Roman scholar Pliny the Elder, whose writings the Flemish painters Hubert and Jan van Eyck are thought to have studied, the Romans used oil colours for shield painting. The earliest use of oil as a fine-art medium is generally attributed to 15th-century European painters, such as Giovanni Bellini and the van Eycks, who glazed oil colour over a glue-tempera underpainting. It is also thought probable, however, that medieval manuscript illuminators had been using oil glazes in order to achieve greater depth of colour and more subtle tonal transitions than their tempera medium allowed.

Earliest use as a fine-art medium

Oils have been used on linen, burlap, cotton, wood, hide, rock, stone, concrete, paper, cardboard, aluminum, copper, plywood, and processed boards, such as masonite, pressed wood, and hardboard. The surface of rigid panels is traditionally prepared with gesso and that of canvas with one or more coats of white acrylic resin emulsion or with a coat of animal glue followed by thin layers of white-lead oil primer. Oil paints can be applied undiluted to these prepared surfaces or can be used thinned with pure gum turpentine or its substitute, white mineral spirit. The colours are slow drying; the safest dryer to speed the process is cobalt siccative.

An oil glaze is a transparent wash of pigment, traditionally thinned with an oleoresin or with stand oil (a concentrate of linseed oil). Glazes can be used to create deep, glowing shadows and to bring contrasted colours into closer harmony beneath a unifying tinted film. Scumbling is the technique of scrubbing an undiluted, opaque, and generally pale pigment across others for special textural effects or to raise the key of a dark-coloured area.

Hog-bristle brushes are used for much of the painting, with pointed, red sable-hair brushes generally preferred for outlines and fine details. Oils, however, are the most plastic and responsive of all painting mediums and can be handled with all manner of tools. The later works of Titian and Rembrandt, for example, appear to have been executed with thumbs, fingers, rags, spatulas, and brush handles. With these and other unconventional tools and techniques, oil painters create pigment textures ranging from delicate tonal modulations to unvarying, mechanical finishes and from clotted, impasto ridges of paint to barely perceptible stains.

The tempera-underpainting-oil-glaze technique was practiced into the 17th century. Artists such as Titian, El Greco, Rubens, and Diego Velázquez, however, used oil pigments alone and, employing a method similar to pastel painting, applied them directly to the brownish ground with which they had tinted the white priming. Contours and shadows were stained in streaks and washes of diluted paint, while lighter areas were created with dry, opaque scumbles, the tinted ground meanwhile providing the halftones and often remaining untouched for passages of local or reflected colour in the completed picture. This use of oil paint was particularly suited to expressing atmospheric effects and to creating chiaroscuro, or light and dark, patterns. It also encouraged a bravura handling of paint, where stabs, flourishes, lifts, and pressures of the brush economically described the most subtle changes of form, texture, and colour according to the influence exerted by the tinted ground through the varying thicknesses of overlaid pigment. This method was still practiced by the 19th-century painters, such as John Constable, J.M.W. Turner, Eugène Delacroix, and Honoré Daumier. The Impressionists, however, found the luminosity of a brilliant white ground essential to the alla prima technique with which they represented the colour intensities and shifting lights of their plein air (open air) subjects. Most oil paintings since then have been executed on white surfaces.

The rapid deterioration of Leonardo's 15th-century "Last Supper," which was painted in oils on plaster, possibly deterred artists from using the medium directly on a wall surface. And the likelihood of eventual warping also prohibited using the large number of braced wood panels required to make an alternative support for an extensive mural painting in oils. Canvas, however, can be woven to any length, and, since an oil-painted surface is elastic, mural paintings could be executed in the studio and rolled and restretched on a wooden framework at the site or marouflaged (fastened with an adhesive) directly onto the wall surface. In addition to the immense studio canvases painted for particular sites by artists such as Jacopo Tintoretto, Paolo Veronese, Delacroix, Pierre-Cécile Puvis de Chavannes, and Monet, the use of canvas has made it possible for mural-size, modern oil paintings to be transported for exhibition to all parts of the world.

Deterioration of Leonardo's "Last Supper"

The tractable nature of the oil medium has sometimes encouraged slipshod craftsmanship. Working over partly dry pigment or priming may produce a wrinkled surface. The excessive use of oil as a vehicle causes colours to yellow and darken, while cracking, blooming, powdering, and flaking can result from poor priming, overthinning with turpentine, or the use of varnish dryers and other spirits. Colour changes may also occur through the use of chemically incompatible pigment mixtures or from the fading of fugitive synthetic hues, such as the crimson lakes used by Pierre-Auguste Renoir.

Watercolour. Watercolours are pigments ground with gum arabic and gall and thinned with water in use. Sable

and squirrel ("camel") hair brushes are used on white or tinted paper and card.

Three hundred years before the late 18th-century English watercolourists, Albrecht Dürer had anticipated their technique of transparent colour washes in a remarkable series of plant studies and panoramic landscapes. Until the emergence of the English school, however, watercolour became a medium merely for colour tinting outlined drawings or, combined with opaque body colour to produce effects similar to gouache (see below) or tempera, was used in preparatory studies for oil paintings.

The chief exponents of the English method were Thomas Girtin, John Sell Cotman, John Robert Cozens, Richard Parkes Bonington, David Cox, and Constable. Their contemporary J.M.W. Turner, however, true to his unorthodox genius, added white to his watercolour and used rags, sponges, and knives to obtain unique effects of light and texture. Victorian watercolourists, such as Birket Foster, used a laborious method of colour washing a monochrome underpainting, similar in principle to the tempera-oil technique. Following the direct, vigorous watercolours of the French Impressionists and Postimpressionists, however, the medium was established in Europe and America as an expressive picture medium in its own right. Notable 20th-century watercolourists have been Wassily Kandinsky, Paul Klee, Dufy, and Georges Rouault; the U.S. artists Thomas Eakins, Maurice Prendergast, Charles Burchfield, John Marin, Lyonel Feininger, and Jim Dine; and the English painters John and Paul Nash, Eric Ravilious, Edward Bawden, Edward Burra, and Patrick Procktor.

The "pure" watercolour technique

In the "pure" watercolour technique, often referred to as the English method, no white or other opaque pigment is applied, colour intensity and tonal depth being built up by successive, transparent washes on damp paper. Patches of white paper are left unpainted to represent white objects and to create effects of reflected light. These flecks of bare paper produce the sparkle characteristic of pure watercolour. Tonal gradations and soft, atmospheric qualities are rendered by staining the paper when it is very wet with varying proportions of pigment. Sharp accents, lines, and coarse textures are introduced when the paper has dried. The paper should be of the type sold as "handmade from rags"; this is generally thick and grained. Cockling is avoided when the surface dries out if the dampened paper has been first stretched across a special frame or held in position during painting by an edging of adhesive tape.

Ink. Ink is the traditional painting medium of China and Japan, where it has been used with long-haired brushes of wolf, goat, or badger on silk or absorbent paper. Oriental black ink is a gum-bound carbon stick that is ground on rough stone and mixed with varying amounts of water to create a wide range of modulated tones or applied almost dry, with lightly brushed strokes, to produce coarser textures. The calligraphic brush technique is expressive of Zen Buddhist and Confucian philosophies. brush-stroke formulas for the spiritual interpretation of nature in painting dictating the use of the lifted brush tip for the "bone," or "lean," structure of things and the spreading belly of the hairs for their "flesh," or "fat," volumes. The Far Eastern artist poises the brush vertically above the paper and controls its rhythmic movements from the shoulder. Distant forms represented in landscapes painted on silk were sometimes brushed on from the reverse side.

In the Western world, ink has been used rather more for preparatory studies and topographical and literary illustrations than as a medium for easel paintings. Western artists have generally combined ink washes with contours and textures in quill or steel pen. Among the finest of these are by Rembrandt, Nicolas Poussin, Francisco Goya, Samuel Palmer, Constable, and Édouard Manet. Claude Lorrain, Turner, and Daumier and, in the 20th century, Braque, Picasso, Reginald Marsh, Henri Michaux, and John Piper are some of those who have exploited its unique qualities. Modern artists also use ballpoint and felt pens.

Gouache. Gouache is opaque watercolour, known also as poster paint and designer's colour. It is thinned with water for applying, with sable- and hog-hair brushes, to white or tinted paper and card and, occasionally, to silk. Honey, starch, or acrylic is sometimes added to retard its quick-drying property. Liquid glue is preferred as a thinner by painters wishing to retain the tonality of colours (which otherwise dry slightly lighter in key) and to prevent thick paint from flaking. Gouache paints have the advantages that they dry out almost immediately to a mat finish and, if required, without visible brush marks. These qualities, with the capacities to be washed thinly or applied in thick impasto and a wide colour range that now includes fluorescent and metallic pigments, make the medium particularly suited to preparatory studies for oil and acrylic paintings. It is the medium that produces the suede finish and crisp lines characteristic of many Indian and Islāmic miniatures, and it has been used in Western screen and fan decoration and by modern artists such as Rouault, Klee, Dubuffet, and Morris Graves.

Medium particularly suited to preparatory studies

Encaustic. Encaustic painting (from the Greek: "burnt in") was the ancient method, recorded by Pliny, of fixing pigments with heated wax. It was probably first practiced in Egypt about 3000 BC and is thought to have reached its peak in Classical Greece, although no examples from that period survive. Pigments, mixed with melted beeswax, were brushed onto stone or plaster, smoothed with a metal spatula, and then blended and driven into the wall with a heated iron. The surface was later polished with a cloth. Leonardo and others attempted unsuccessfully to revive the technique. North American Indians used an encaustic method whereby pigments mixed with hot animal fat were pressed into a design engraved on smoothed buffalo hide.

A simplified encaustic technique uses a spatula to apply wax mixed with solvent and pigment to wood or canvas, producing a ridged, impasto surface. This is an ancient and most durable medium. Coptic mummy portraits from the 1st and 2nd centuries AD retain the softly blended, translucent colouring typical of waxwork effigies. In the 19th century, Vincent van Gogh also used this method to give body to his oil pigment; the Neo-Impressionist artist Louis Hayet applied encaustic to paper, and it was used by the U.S. painter Jasper Johns for his Pop art canvases. Coloured wax crayons have also been used by modern painters such as Picasso, Klee, Arshile Gorky, and Hockney (see above *Mediums: Fresco*).

Casein. Casein, or "cheese painting," is a medium in which pigments are tempered with the gluey curd of cheese or milk precipitate. For handling, an emulsion of casein and lime is thinned with water. The active element of casein contains nitrogen, which forms a soluble caseate of calcium in the presence of lime. It is applied in thin washes to rigid surfaces, such as cardboard, wood, and plastered walls.

Casein colours dry quickly, although lighter in tone than when first applied. Since they have more body than egg-tempera paints, they can be applied with bristle brushes to create impasto textures not unlike those of oils. Casein paints were used in ancient Rome. They are now available ready-made in tubes and have been used by such modern artists as Robert Motherwell and Claes Oldenburg.

Casein is also an ingredient of some charcoal and pastel fixatives and was a traditional primer for walls and panels.

Synthetic mediums. Synthetic mediums, developed by industrial research, range from the Liquitex fabric dyes used on canvas by the U.S. abstract painter Larry Poons to the house enamel paints employed at times by Picasso and Jackson Pollock.

The most popular medium and the first to challenge the supremacy of oils is acrylic resin emulsion, since this plastic paint combines most of the expressive capabilities of oils with the quick-drying properties of tempera and gouache. It is made by mixing pigments with a synthetic resin and thinning with water. It can be applied to any sufficiently toothed surface with brush, roller, airbrush, spatula, sponge, or rag. Acrylic paints dry quickly, without brush marks, to form a mat, waterproof film that is also elastic, durable, and easily cleaned. They show little colour change in drying, nor do they darken in time. While they lack the surface textural richness of oil or encaustic, they can be built up with a spatula into opaque impastos or thinned immediately into transparent colour glazes. Polyvinyl acetate (PVA) is applied for priming, although it is claimed that acrylic paints can be safely applied directly

Acrylic resin emulsion

onto unprepared raw canvas or cotton. The wide range of intense hues is extended by fluorescent and metallic pigments. Polymer paints are particularly suitable for the precise, immaculate finish demanded by Op art, Minimalist, and Photo-realist painters such as Bridget Riley, Morris Louis, Frank Stella, and Richard Estes.

Other mediums. *French pastels.* French pastels, with the sharpened lumps of pigment used by Ice Age artists, are the purest and most direct painting materials. Pastel pigments are mixed only with sufficient gum to bind them for drying into stick molds. Generally, they are used on raw strawboard or on coarse-grained tinted paper, although vellum, wood, and canvas have been also employed. These colours will not fade or darken, but, since they are not absorbed by the surface of the support, they lie as pigment powder and are easily smudged. Unfortunately, pastel colours lose their luminosity and tonality if fixed with a varnish and so are best preserved in deep mounts behind glass. Degas often overcame the fragile nature of true pastel painting by the unorthodox method of working on turpentine-soaked paper, which absorbed the powdery pigment.

Eighteenth-century portrait pastellists, such as Maurice-Quentin de La Tour, Jean-Baptiste Peronneau, Jean-Étienne Liotard, and Anton Raphael Mengs, blended the pigment with coiled paper stumps so the surface resembled that of a smooth oil painting. Later pastel painters, such as Degas, Toulouse-Lautrec, Mary Cassatt, Everett Shinn, Odilon Redon, and Arthur Dove, contrasted broad masses of granular colour, spread with the side of the stick, with broken contours and passages of loose cross-hatching and smudging. They often used the tinted ground as a halftone, and, according to the amount of manual pressure exerted on the chalk, they varied the degree of pigment opacity to extract a wide range of tints and shades from each pastel colour.

Oil pastels. Oil pastels are pigments ground in mastic with oil of turpentine, spermaceti, and poppy oil. They are used in a similar way to that of French pastels but are already fixed and harder, producing a permanent, waxy finish. Oil-pastel paintings are generally executed on white paper, card, or canvas. The colours can be blended if the surface of the support is dampened with turpentine or if they are overworked with turpentine. They are popular for small preparatory studies for paintings.

Glass paintings. Glass paintings are executed with oil and hard resin or with watercolour and gum on glass sheets. These have been a folk art tradition in Europe and North America and, from the 15th to the 18th century, were regarded as a fine art in northern Europe, where they have been more recently revived by such painters as Willi Dirx, Ida Kerkovius, Lily Hildebrandt, Klee, Oskar Schlemmer, and Heinrich Camperdonck. Colours are applied from the back in reverse order. Unpainted areas of glass are often coated with mercury, providing a mirror background to the coloured images; this creates the kind of illusionary, bizarre spatial relationship between the viewer and picture space sought by the modern artist Michelangelo Pistoletto with his use of photographic images fixed to a polished steel sheet. The colours seen through glass appear translucent, jewel-like, and, since they cannot be touched, even magical.

Folk art tradition

Ivory painting. Ivory painting was practiced in the 18th and 19th centuries in Europe and America for portrait miniatures. These were generally oval-shaped and designed as keepsakes, lockets, and mantle pictures. They were painted under a magnifying glass in fairly dry watercolour or tempera stippling, with sable- or marten-hair brushes on thin, semitranslucent ivory pieces. Corrections were made with a needle. The velvet quality of their colours was enhanced, on the thinner ivories, by the glow produced by a gold leaf or tinted backing.

Lacquer. Lacquer has been a traditional Chinese medium for more than 2,000 years. It combines painting with intaglio relief. Linen-covered wood panels are coated with chalk or clay, followed by many thin layers of black or red lacquer-tree resin. The surface is polished and a design engraved, which is then coloured and gilded or inset with mother-of-pearl. Layers of compressed paper or molded papier-mâché have also provided supports. In China and Japan, lacquer has been used principally for decorating shrine panels, screens, caskets, panniers (large baskets), and musical instruments.

Sand, or dry, painting. Sand, or dry, painting is a traditional magic art of the North American Indians; it is still practiced in healing ceremonies among the Navajos of New Mexico and Arizona. Ground sandstone, natural ochres, mineral earths, and powdered charcoal are sprinkled onto a pattern marked into an area covered with yellow-white sand. The patient sits in the centre of this vivid symbolic design of coloured figurative and geometrical shapes. Following the ritual, the painting is destroyed. These "floor" pictures influenced Jackson Pollock in his horizontally spread "action paintings."

Paper. From the end of the 18th century, profiles and full-length group portraits were cut in black paper, mounted on white card, and often highlighted in gold or white. A silhouette ("shade") might be first outlined from the sitter's cast shadow with the aid of a physionotrace.

Collage. Collage was the Dada and Synthetic Cubist technique of combining labels, tickets, newspaper cuttings, wallpaper scraps, and other "found" surfaces with painted textures simulating wood graining and marbling. Among the most lyrical and inventive works in this magpie medium are the so-called *Merz* collages by Kurt Schwitters. Frottage was Max Ernst's method of taking paper rubbings from surfaces, unrelated to one another in real life, and combining them to create fantasy landscapes. Cut paper shapes, hand coloured in gouache, were used by Matisse for his monumental last paintings; Piet Mondrian composed his famous "Victory Boogie Woogie" (1942–43) in coloured-paper cutouts.

Mechanical mediums. The use of mechanical mediums in painting has run parallel to similar developments in modern music and drama. In the field of cybernetics, painters have programmed computers to permutate drawings, photographs, diagrams, and symbols through sequences of progressive distortion; and light patterns are produced on television screens by deliberate magnetic interference and by sound-wave oscillations. Artists have also explored the expressive and aesthetic possibilities of linear holograms, in which all sides of an object can be shown by superimposed light images. Painters are among those who have extended the boundaries of filmmaking as an art form. Following the Surrealist film fantasies created by Berthold Bartosch, Jean Cocteau, Hans Richter, and Salvador Dalí, by Schlemmer's filmed ballets and Norman McLaren's hand-painted abstract animations, some painters and other graphic artists are now experimenting with video cassettes for television.

For some Conceptual artists the typewriter is the only equipment used when visual ideas are expressed in the form of instruction sheets alone. For example, typed proposals for defining the real space of an exhibition area with painted lines might invite the active participation of visitors (see below *Forms of painting: Modern forms*).

Mixed mediums. Some pictures are first painted in one medium and corrected or enriched with colour and texture in another. Examples of this kind of mixed mediums are the Renaissance tempera-oil technique, William Blake's relief etchings colour-printed in glue tempera and hand-finished in watercolour, and Degas's overpainted monotypes and his combinations of pastel, gouache, and oil. More recent examples are Richard Hamilton's photographs overpainted in oil colour, Dubuffet's patchwork assemblages of painted canvas and paper, and Klee's alchemy in mixing ingredients such as oil and distemper on chalk over jute and watercolour and wax on muslin stuck on wood.

FORMS OF PAINTING

Mural painting. Mural painting has its roots in the primeval instincts of people to decorate their surroundings and to use wall surfaces as a form for expressing ideas, emotions, and beliefs. In their universal manifestation in graffiti and in ancient murals, such as Ice Age cave paintings and protodynastic Egyptian frescoes, symbols and representational images have been spread freely and

indiscriminately across walls, ceilings, and floors. But, in more disciplined attempts to symbolize the importance and function of particular buildings through their interior decoration, murals have been designed for the restricted framework of specific surface areas. They therefore have to be painted in close relationship to the scale, style, and mood of the interior and with regard to such siting considerations as light sources, eye levels, the spectators' lines of sight and means of approach, and the emotive scale relationship between spectators and the painted images.

Early mural decorations

Early mural decorations for tombs, temples, sanctuaries, and catacombs were generally designed in horizontal divisions and vertical axes. These grid patterns were in harmony with the austere character of the interiors, and their geometrical plan enabled the artist to depict clearly the various episodes and symbols of a narrative subject. In these early traditions of mural design, in China, India, Mexico, Egypt, Crete, and Byzantium, no illusionary devices were used to deny the true flatness of the wall surface; images were silhouetted against a flatly painted ground framed by decorative dadoes (the decoration adorning the lower part of an interior wall) of stylized motifs in repeat patterns. By the early Renaissance, however, innovators such as Giotto, Masaccio, and Fra Angelico were placing figures within architectural and landscape settings, painted as if extensions to the real dimensions of the interior. The peak of technical skill and artistic expression was reached in the 15th and 16th centuries with the frescoes of Piero della Francesca, Michelangelo, and Raphael. The irregular shapes of wall areas and the distortions produced by convex surfaces were inventively exploited in the design. Intruding doors and windows, for example, were skillfully circumvented by sweeping pattern rhythms or were incorporated as features in the painting, and figures were foreshortened so as to appear to float across or to rise into cupolas (rounded vaults that form ceilings), lunettes (rounded spaces over doors or windows), and apses (domed projections of a church, usually at the east end or altar), the curving surfaces of which might be painted to simulate celestial skies. Existing structural wall features provided the divisions between narrative episodes. These were often supplemented by trompe l'oeil ("deceive the eye") columns, pilasters, arcading, balustrading, steps, and other architectural forms that also served to fuse the painted setting with the real interior.

With the increasing dependence upon tapestry hangings and stained glass as primary forms of interior decoration, mural painting suffered a decline in the Western world. Except for those given to Rubens, Tiepolo, Delacroix, and Puvis de Chavannes, there were relatively few important mural commissions in the period following the High Renaissance. In the 20th century, however, enlightened patronage has occasionally enabled leading modern artists to execute paintings for specific sites: Monet's "Water-Lilies" series for the Paris Orangerie, for example, and other murals in France by Vuillard, Matisse, Léger, Chagall, and Picasso; in Mexico and the United States by Orozco, Rivera, Tamayo, and David Siqueiros, and also in the U.S. by Matisse, Shahn, and Willem de Kooning; in Britain by Sir Stanley Spencer and Bawden; in Norway by Edvard Munch; in Holland by Karel Appel; and in Italy by Afro Basaldella.

Easel and panel painting. The easel, or studio, picture was a form developed during the Renaissance with the establishment of the painter as an individual artist. Its scale and portability enabled European artists to extend the range of themes, previously restricted to those suitable to mural decoration. Easel and panel forms include still life, portraiture, landscape, and genre subjects and permit the representation of ephemeral effects of light and atmosphere that the more intimate forms of Oriental art had already allowed the painters of scrolls, screens, and fans to express. Although easel paintings are occasionally commissioned for a special purpose, they are generally bought as independent art objects and used as decorative focal features or illusionary window views in private homes. They are also collected as financial investment, for social prestige, for the therapeutic escapism their subject may provide, or purely for the aesthetic pleasure they afford.

Panel paintings

Panel paintings, by strict definition, are small pictures designed for specific sacred or secular purposes or as part of a functional object. Although these wooden boards are sometimes categorized as a form of "decorative" rather than "fine" art, the best examples justify their place in museums alongside great easel paintings. Among the functions they originally served were as predellas (the facings to altar-step risers); devotional and ceremonial icons; portable, folding diptych and triptych altarpieces; shop and tavern signboards; mummy cases; and panel decorations of carriages, musical instruments, and cassoni. Many of them were painted by acknowledged masters, such as Fra Angelico, Paolo Uccello, and Antoine Watteau, as well as by anonymous folk artists.

Miniature painting. Miniature painting is a term applied both to Western portrait miniatures and to the Indian and Islāmic forms of manuscript painting discussed below. Portrait miniatures, or limnings, were originally painted in watercolour with body colour on vellum and card. They were often worn in jewelled, enamelled lockets. Sixteenth-century miniaturists, such as Hans Holbein, Jean Clouet, Nicholas Hilliard, and Isaac Oliver, painted them in the tradition of medieval illuminators. Their flat designs, richly textured and minutely detailed, often incorporated allegorical and gilded heraldic motifs. In 17th- and 18th-century Western portrait miniatures, the two-dimensional pattern of rich colours was developed by atmospheric tonal modelling into more naturalistic representations; these were sometimes in pastel and pencil or painted in oils on a metal base. Pantographs (reducing and enlarging copying instruments made on the lazy-tongs lever principle) might be used to transfer a drawing. Among the exponents of this naturalistic style were Francisco Goya, Fragonard, Samuel Cooper, John Hoskins, François Dumont, and the U.S. miniaturists Robert Field and Edward Green Malbone. The introduction of painted ivory miniatures was followed, in the 19th century, by a decline in aesthetic standards, although a classical simplicity was achieved by unsophisticated itinerant limners and by the German miniaturist Patricius Kittner. The painted miniature was eventually superseded by the small, hand-tinted photograph (see above *Mediums: Ivory painting*).

Manuscript illumination and related forms. Among the earliest surviving forms of manuscript painting are the papyrus rolls of the ancient Egyptian Book of the Dead, the scrolls of Classical Greece and Rome, Aztec pictorial maps, and Mayan and Chinese codices, or manuscript books. European illuminated manuscripts were painted in egg-white tempera on vellum and card. Their subjects included religious, historical, mythological, and allegorical narratives, medical treatises, psalters, and calendars depicting seasonal occupations. In contrast to the formalized imagery of Byzantine and early Gothic manuscript painters, Celtic illuminators developed a unique, abstract style of elaborate decoration, the written text being overwhelmed by intricate latticework borders, with full-page initial letters embraced by interlacing scrolls. The medieval Gothic style of illumination, in sinuous, linear patterns of flattened forms isolated against white or gilded grounds, had developed, by the end of the 15th century, into exquisitely detailed, jewel-like miniatures of shaded figures and spatial landscapes. These were often framed by gilded initial letters as vignettes or by margin borders in simulated half relief. With the advent of printing in the 15th century and a final, brilliant period of Flemish and Italian illumination, European manuscript painting survived only in official documents, maps, and in the form of hand-coloured, block-printed pages. Pennsylvanian-German birth and baptismal certificates in the U.S. and William Blake's hand-coloured engravings to the Bible and to his own poems were isolated revivals of those forms.

Abstract style of Celtic illuminators

Indian and Islāmic miniature painting, however, was practiced into the 19th century; and 11th-century Oriental albums of poem paintings in ink, on leaves of silk or paper, represent a tradition that was continued into modern times. The subjects of Near Eastern miniatures included religious and historical narrative, cosmic maps, and medical, palmistry, and astrological charts, as well as illustrations to poems, songs, and romantic epics. These

Tempera



"Wind from the Sea," egg tempera painting by Andrew Wyeth, 1947. In a private collection. 47 × 70 cm.



"Christ Discovered in the Temple," panel painting in egg tempera with gold leaf by Simone Martini, 1342. In the Walker Art Gallery, Liverpool. 49.5 × 35 cm.



"King Khusrau Seated upon His Throne," Persian miniature in glue tempera from the *Khamsa of Nizami*, 1524–25. In the Metropolitan Museum of Art, New York City. 32 × 23 cm.

A leaf from the *Calendar of a Book of Hours (April)*, manuscript illumination in egg tempera by Simon Bennick, early 16th century. In the Victoria and Albert Museum, London. 14 × 9.6 cm.

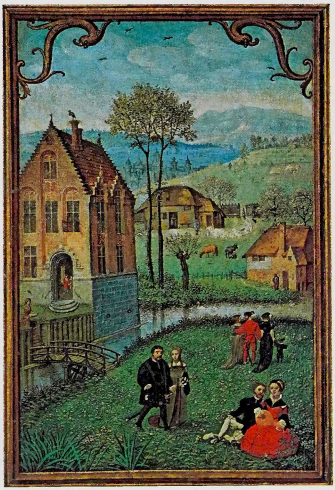
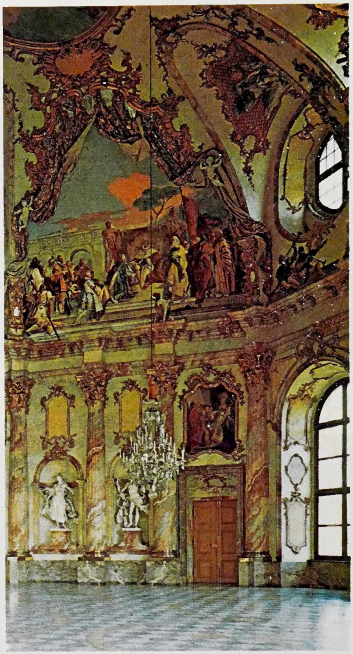


Plate 1: By courtesy of (top right) the Walker Art Gallery, Liverpool, (bottom left) the Metropolitan Museum of Art, New York, gift of Alexander Smith Cochran, 1913, (bottom right) the Victoria and Albert Museum, London; photograph, (bottom right) John Webb



Fresco of a court scene from the *Mahājanaka Jataka*, Cave I, Ajantā, India, 600–700.

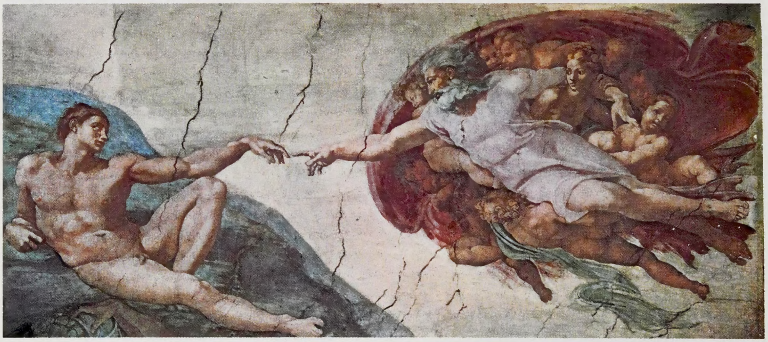


Frescoes by Giovanni Battista Tiepolo, 1750–52, decorating the Kaisersaal Residenz, Würzburg, West Germany, designed by Balthasar Neumann, 1719–44.



"Fowling in the Marshes," detail of a wall painting from the tomb of Nebamun, Thebes, Egypt, *c.* 1450 BC, XVIII dynasty. In the British Museum.

"The Creation of Adam," detail of the ceiling fresco in the Sistine Chapel, Vatican, by Michelangelo, 1508–12.



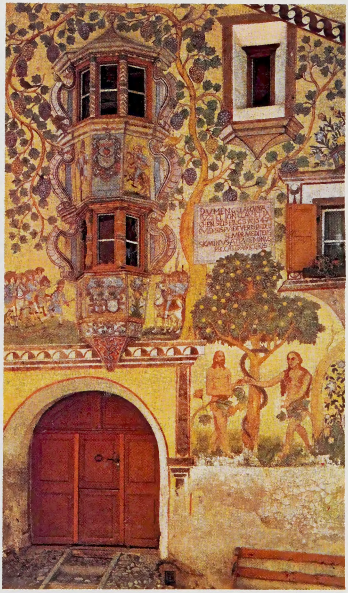
Wall painting



Frescoes by José Clemente Orozco, 1938–39, decorating the ceiling and drum of the dome of the Hospicio Cabañas, Guadalajara, Mexico.



Fresco cycle depicting scenes from the life of the Virgin and the life of Christ by Giotto in the Arena Chapel, Padua, Italy, *c.* 1305–09.



Folk art fresco on the Adam and Eve house, Ardez, Switzerland, 1647.

Plate 2: By courtesy of (centre left) the trustees of the British Museum; photographs, (top left) V. Paniabi—Shostal/EB Inc., (top right) A.F. Kersting, (bottom) SCALA—Art Resource/EB Inc.

Plate 3: By courtesy of (top right) Fondo Editorial de la Plastica Mexicana; photographs, (centre left) SCALA—Art Resource/EB Inc., (bottom right) Toni Schneiders

Oil and encaustic



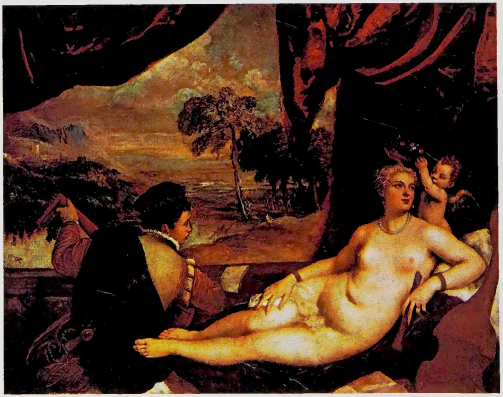
"Salisbury Cathedral," oil painting by John Constable, *c.* 1825. In the National Gallery, London. 53 × 77 cm.



"Smaragd Red and Germinating Yellow," oil painting by Hans Hofmann, 1959. In the Cleveland Museum of Art. 1.4 × 1 m.



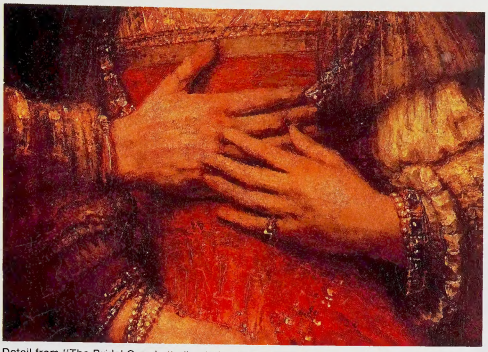
Detail from "Water Lilies," oil on canvas by Claude Monet, 1904. In the Musée de Jeu de Paume, Paris. Entire painting 90 × 92 cm.



"Venus and the Lute Player," oil painting by Titian, *c.* 1560. In the Metropolitan Museum of Art, New York City. 1.7 × 2.1 m.

Plate 4: By courtesy of (top left) the Contemporary Collection of the Cleveland Museum of Art, (top right) the trustees of the National Gallery, London, (bottom) the Metropolitan Museum of Art, New York, Munsey Fund, 1936; photograph, (centre right) SCALA / Art Resource, NY.

Plate 5: By courtesy of (top left) the Rijksmuseum, Amsterdam, (top right) the Metropolitan Museum of Art, New York, gift of Edward S. Harkness, 1917–18, (bottom left) the National Gallery of Art, Washington, D.C., Andrew W. Mellon Collection, 1937, (bottom right) the Fogg Art Museum, Harvard University, purchase, Alpheus Hyatt Fund

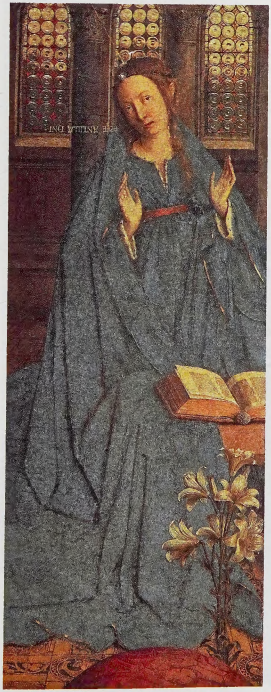


Detail from "The Bridal Couple," oil painting by Rembrandt, *c.* 1665. In the Rijksmuseum, Amsterdam. Entire painting 1.2 × 1.7 m.



Portrait of a boy, encaustic painting from al-Fayyūm, Egypt, probably 2nd century. In the Metropolitan Museum of Art, New York City. 33 × 18.5 cm.

Detail from "The Annunciation," oil painting by Jan van Eyck, *c.* 1434/36. In the National Gallery of Art, Washington, D.C. Entire painting 93 × 36.5 cm.



Detail from "Quos Ego" ("Neptune Calming the Tempest"), oil painting by Peter Paul Rubens, *c.* 1635. In the Fogg Art Museum, Harvard University. Entire painting 49 × 64 cm.





"The Burning of the Houses of Parliament," watercolour painting by J.M.W. Turner, 1834. In the British Museum. 23.4 × 32.4 cm.

Ink, gouache, and watercolour



"A Young Man Among Roses," watercolour miniature by Nicholas Hilliard, *c.* 1588. In the Victoria and Albert Museum, London. 13.4 × 7.4 cm.

"The Annunciation," anonymous Austrian or Swabian woodcut, hand-coloured with watercolour, 1450–70. In the National Gallery of Art, Washington, D.C. 27.3 × 19.2 cm.

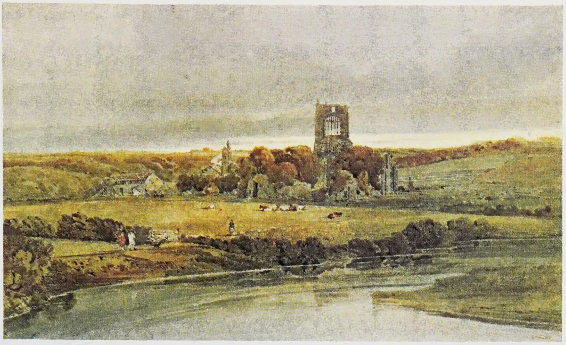




"The Monumental Turf," watercolour painting by Albrecht Dürer, 1503. In the Albertina, Vienna. 41 × 31.5 cm.

Plate 6: By courtesy of (top left) the Victoria and Albert Museum, London, (bottom left) the National Gallery of Art, Washington, D.C., Rosenwald Collection, (bottom right) the Albertina, Vienna; photograph, (top right) Hans Hinz, Basel

Plate 7: By courtesy of (top, centre right) the Victoria and Albert Museum, London, (bottom) the Museum of Fine Arts, Boston, Ross Collection; photographs, (top, centre right) John Webb, (centre left) Hollé Bildarchiv, Baden-Baden



"Kirkstall Abbey, Yorkshire—Evening," watercolour painting by Thomas Girtin, 1801. In the Victoria and Albert Museum, London. 30.4 × 51 cm.



"Landscape," detail of a Japanese handscroll, ink on paper by Sesshū (1420–1506). In the National Museum, Tokyo. Dimensions of entire scroll 147.3 × 35.6 cm.



French fan, gouache on paper, sticks of carved mother-of-pearl inlaid with gilt and set with pastes, *c.* 1750. In the Victoria and Albert Museum, London.

"The Return of Wen-chi to Ying-chuan," scene from the Chinese handscroll "The Life of Lady Ts'ai-chi," anonymous, ink and colour on silk, 12th century, Sung dynasty. In the Museum of Fine Arts, Boston. Detail of the scroll 25 × 55.8 cm.



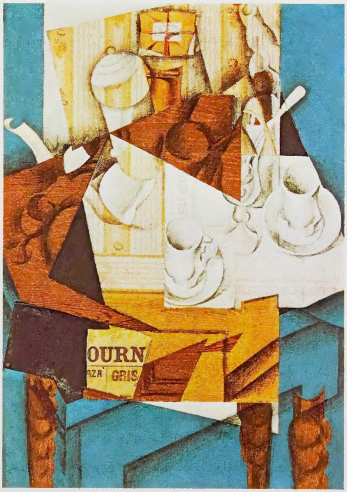


"Union I," epoxy paint and fluorescent alkyd painting by Frank Stella, 1966. In the Detroit Institute of Arts. 2.6 × 4.5 m.

Modern forms



"First Landing Jump," combine painting by Robert Rauschenberg, 1961. In the Museum of Modern Art, New York City. 2.3 × 1.8 m.



"Breakfast," pasted paper, crayon, and oil paint collage by Juan Gris, 1914. In the Museum of Modern Art, New York City. 81 × 60 cm.



"Grossinger's Bakery," oil painting by Richard Estes, 1972. In the Allan Stone Gallery, New York City. About 89 × 61 cm.

Plate 8. By courtesy of (top left) the Detroit Institute of Arts, gift of Friends of Modern Art, (top right) the Leo Castelli Gallery, New York, (bottom left) the Museum of Modern Art, New York, acquired through the Lillie P. Bliss Bequest, (bottom right) the Allan Stone Gallery, New York City

were generally painted in gouache on paper, with occasional gold- or silver-leaf embellishment. The linear design was first drawn with a brush in delicate contours and soft shading. Landscape and architectural detail was as well observed as in that of the principal figures.

The rapprochement established between text, painted borders, margin spaces, and illustration is characteristic of both Eastern and Western manuscript paintings. In Indian and Islāmic miniatures, for example, the panels of decorative script are integrated within the overall pattern as areas of textural enrichment; and, with the margin and inset frames, these panels serve also as concrete screens and prosceniums to the action depicted, the participants in the narrative episode making their exits and entrances across or behind them.

Scroll painting. Hand scrolls, traditional to China and Japan, are ink paintings on continuous lengths of paper or silk. They are unrolled at arm's length and viewed from right to left. These generally represent panoramic views of rivers, mountain and urban landscapes, and domestic interiors. They also illustrate romantic novels, Taoist and Buddhist themes, and historical and genre subjects. Narrative poetic commentaries were included as integral textures in the flowing design. The scrolls are remarkable for their vitality, the lyrical representation of atmospheric space, and for the rising and dipping viewpoints that anticipate the zooming motion-picture camera. The earliest surviving scrolls, such as Ku K'ai-chih's "Admonitions of the Instructress to the Court Ladies," date from the 4th century AD. Oriental hanging scrolls and Indian and Tibetan temple banners are forms similar to those of Western easel and panel paintings. Their subjects range from the seasons, domestic interiors, landscapes, and portraits to Vishnu epics, mandalas (symbolic diagrams of the universe), and temple icons. They are painted in ink or gouache on silk and paper and are usually mounted on embroidered or block-printed silk. The dramatic interplay of bold, flattened images against the open space of an unpainted or gilded ground influenced 19th-century Western Art Nouveau decoration.

Hanging scrolls and temple banners

Screen and fan painting. Folding screens and screen doors originated in China and Japan, probably during the 12th century, and continued as a traditional form into the 20th. They are in ink or gouache on plain or gilded paper and silk. Their vivid rendering of animals, birds, and flowers and their atmospheric landscapes brought nature indoors. In some screens each panel was designed as an individual painting, while in others a continuous pattern flowed freely across the divisions. Japanese screens were often painted in complementary Yin and Yang pairs. Large 12-panel Chinese coromandel lacquer screens were imported into Europe during the 17th and 18th centuries. French Rococo boudoir screens depicting *fêtes champêtres* (townspeople enjoying rural surroundings) and toile de Jouy (landscape or floral) pastoral themes were painted on silk or on wood panels in a flamboyantly scrolled, gilded framework. The designs of Art Nouveau screens were inspired by the Japanese tradition. Sidney Nolan's screens on Greek themes and the pastiches of Victorian paper-scrap screens by Pop art painters are recent Western revivals. Traditional to the Greek and Russian Orthodox churches is the iconostasis screen, which stands between the nave and sanctuary and displays icon panel paintings representing the Virgin, the saints, and narrative subjects.

Rigid fans are depicted in the paintings and reliefs of ancient Egypt, Assyria, Greece, and Rome, but the oldest surviving specimens are the round and folding fans from Japan and China. These were painted in India ink and colour on paper, card, and silk, the ground often sprinkled with gold dust or laid with gold or silver leaf. Spread freely across the mount, a calligraphic design depicted seasonal landscapes, genre scenes, and bird, flower, and animal motifs, with accompanying poems and commentaries. Leading Oriental painters produced much of their finest work in this form. In Europe, however, where fan painting had been rarely practiced until the 17th century, it was considered a minor art, and designs were often based on frescoes and easel paintings. The richest and most elegant of these were painted in France and Italy during the 18th century. Watercolour and gouache paintings and hand-coloured engraved designs were made on paper, card, kid, and gauze. Allegories and romantic pastoral landscapes were frequently designed as separate vignettes, linked by floral swags and border scrolls. Both sides of the mount might be painted. The guards and sticks of the spoke framework were in delicately carved wood or ivory, inlaid with gold leaf or mother-of-pearl. Round hand-screens of parchment, mounted on handles like lollipops, were popular in early 19th-century English society. Charles Conder was a notable fin de siècle ("end of the century," characterized by effete sophistication) fan painter, and, more recently, Oskar Kokoschka decorated a lively set of fans on an autobiographical narrative theme.

Round hand-screens

Panoramas. Panoramas were intended to simulate the sensation of scanning an extensive urban or country view or seascape. This form of painting was popular at the end of the 18th century. Notable examples are "The Battle of Agincourt" (1805), by R.K. Porter, and "Scheveningen," by Hendrik Willem Mesdag. Panoramas might be compared to Cinerama films and enjoyed as a stimulating optical entertainment, along with cyclorama drums (large pictorial representations encircling the spectator), trompe l'oeil diorama peep shows, and the show box, for which Thomas Gainsborough painted glass transparencies. More serious forms of panoramic painting are exemplified in Chinese Buddhist sanctuary frescoes, Oriental hand scrolls, Dürer's watercolour townscapes, Andrey Rublyov's 14th-century mural of Moscow, and Uccello's original sequence of three panels depicting the Battle of San Romano.

Modern forms. The concept of painting as a medium for creating illusions of space, volume, texture, light, and movement on a flat, stationary support has been challenged by many modern artists. Some recent forms, for example, have blurred the conventional distinctions between the mediums of sculpture and painting. Sculptors such as David Smith, Eduardo Paolozzi, and Philip Sutton have made multicoloured constructions; painters such as Jean Arp and Ben Nicholson have created abstract designs in painted wood relief, and Richard Smith has painted on three-dimensional canvas structures the surfaces of which curl and thrust toward the spectator. And, rather than deny the essential flatness of the painting support by using traditional methods of representing volume and texture, Robert Rauschenberg and Jim Dine have attached real objects and textures to the painted surface, and Frank Stella and Kenneth Noland have designed their irregularly shaped canvases to be seen as explicitly flat art objects. Rejecting earlier painting methods of reproducing effects of light with tonal contrasts and broken pigment colour, some artists have made use of neon tubes and mirrors. Instead of simulating sensations of movement by optical illusion, others have designed kinetic panels and boxes in which coloured shapes revolve under electric power. The traditional definition of painting as a visual, concrete art form has been questioned by recent aspects of Conceptual art, in which the painter's idea might be expressed only in the form of documented proposals for unrealized and often unrealizable projects. In "performance art" and "happenings," which employ techniques akin to those used in theatre, the artists themselves become a kind of medium.

IMAGERY AND SUBJECT MATTER

The imagery and subject matter of paintings in early cultures were generally prescribed by tribal, religious, or dynastic authorities. In some Eastern countries, traditional models survived into the 18th century and even later. With the Renaissance, however, images and themes in Western painting, reflecting the new spirit of Humanistic, objective curiosity and scientific research, came to be decided by the artist and his patron and, in more recent periods, by the artist alone.

Kinds of imagery. Within the various cultures the art of representing things by painted images has rarely shown a continuously developing pattern toward greater realism. More often, religious and philosophical precepts have determined the degree of naturalism permitted. Rules governing portrayals of the human figure have been particularly stringent in certain traditions of representational painting,

Attitudes toward the cosmic significance of man

reflecting different attitudes to the cosmic significance of man. For example, a belief in the inferiority of man in relation to an almighty deity is expressed in the faceless figures in early Jewish painting and in the dehumanized stylizations of Byzantine imagery; and his insignificance against the dynamic forces of nature is symbolized in Chinese landscape paintings by his puny scale within a monumental setting. An earlier view, which instead sought to glorify the spiritual, intellectual, and physical attributes of mankind, is typified in the noble figures of Greco-Roman art and in the renewed celebration of human physical beauty in the Renaissance and subsequent Neoclassical styles. The uniqueness of man among living things and the expression of his individual physical and emotional characteristics are exemplified in Japanese and northern European narrative and genre painting. Concomitant with the antipathy toward figurative representation in some cultures was a general distaste for the portrayal of all things of the exterior world, animals, landscape features, and other natural forms rarely appearing except as stylized images signifying spiritual forces of good and evil. The representational imagery of modern painting borrows freely from ancient and contemporary sources such as primitive and child art, Classical mythology, commercial advertising, press photography, and the allegories and fantasies of the motion picture and the comic strip. Nonrepresentational imagery is not restricted to modern painting but appears also in earlier forms such as Aurignacian (Paleolithic) decorative meanders, the scrollwork of Celtic illuminations, and the patterns of Islāmic Kūfic calligraphy (an angular variety of the Arabic alphabet). And the abstraction of natural forms into rudimentary symbols, characteristic of modern painting, is echoed in the "pin-men" conventions of Magdalenian caves, in Aztec pictograms, and Indian and Tibetan cosmic-diagram paintings.

Kinds of subject matter. *Devotional.* The range and interpretation of subjects in different forms of devotional painting express a particular attitude to the relationship between man and his deity. Early Christian and Buddhist murals, for example, portrayed an all-powerful, remote, and mysterious being, painted as a flat, formalized head or figure whose stern gaze dominated the interiors of temples, churches, and sanctuaries. Christian Last Judgments and Oriental hell paintings were intended to frighten the believer, while subjects such as the Virgin enthroned, the Assumption, and Buddha descending from Paradise sustained his faith with hopes for salvation and rewards of blissful immortality.

Narrative. When the autocratic ecclesiastical control over Western painting weakened under Renaissance Humanism, the religious narrative picture became a window onto a terrestial rather than a celestial world. Both emotional and physical relationships between the figures depicted were realistically expressed, and the spectator was able to identify himself with the lifelike representation of a worldly space inhabited by Christ, his disciples, and saints, wearing updated dress and moving naturally within contemporary settings. This kind of narrative interpretation persists in the modern religious paintings of Sir Stanley Spencer, where biblical environments are represented by the clipped hedgerows, the churchyards, and the front parlours of his neat, native English village of Cookham.

Allegorical narrative subjects

Allegorical narrative subjects might exalt the sensuous arts, as in the symbolic muses portrayed by Poussin and Luca Signorelli and the paradisiac gardens of 15th-century French illuminated manuscripts. But they might also carry warnings. In the 16th century, Pieter Bruegel the Elder, for example, combined overt and often grotesque symbols with subtle visual metaphors to point stern morals in such paintings as "The Triumph of Death" (alluding to the "wages of sin"), "The Land of Cockaigne" (attacking gluttony and sloth), and "Mad Meg" (ridiculing covetousness). Even Bruegel's apparently straightforward genre subjects, such as "The Peasant Dance" and the festival of "The Fight Between Carnival and Lent," conceal parables on man's follies and sins, while Hieronymus Bosch introduced abstruse, allegorical phantasmagoria into such traditional narratives as "The Temptation of St. Antony" and "The Prodigal Son" and made his "Garden of Delights" an expression of disgust rather than of joy. Botticelli's late paintings, probably produced under the influence of the 15th-century Italian monk and reformer Girolamo Savonarola, are other savagely pessimistic allegories: "The Story of Virginia Romana" and "The Tragedy of Lucretia," representing virtue upheld only by death, and "The Calumny of Apelles," in which envy, suspicion, deceit, guile, repentance, and truth are identified, like medieval mummers, by their costume, pose, and gesture. Rubens, however, found in allegorical symbolism a means of dramatizing mundane state commissions, such as "The Union of Scotland and Ireland" and "The Bounty of James I (Triumphing over Avarice)." Among famous 19th-century allegories are Delacroix's "Liberty Leading the People" and Pierre-Paul Prud'hon's "Crime Pursued by Vengeance and Justice."

Possibly the highest achievements in narrative illustrations to poetry and literature are found in Eastern miniatures and Oriental scrolls, such as the Persian paintings of Ferdowsi's 11th-century national epic poem, the *Shāh-nāmeh,* and the 12th-century Japanese scrolls of the *Genji monogatari* and the *Story of Ben Dainagon.* An example of modern literary painting is Sidney Nolan's narrative series portraying the Australian folklore hero Ned Kelly.

Ancient Greek and Roman mythologies have provided Western artists with rich sources of imagery and subject matter and with opportunities for painting the nude. Historical narrative painting includes Classical mythology and heroic legend, as well as the representation of contemporary events; examples include Benjamin West's "Death of Wolfe," Théodore Géricault's "Raft of the Medusa," and Goya's "The 3rd of May in Madrid."

Portraiture. The earliest surviving portraits of particular persons are probably the serene, idealized faces painted on the front and inside surfaces of dynastic Egyptian sarcophagi. The human individuality of the Roman mummy portraits of the 1st and 2nd century AD, however, suggests more authentic likenesses. Although portraits are among the highest achievements in painting, the subject poses special problems for the artist commissioned to paint a notable contemporary. The portraits of patrons by artists such as Raphael, Rubens, Hyacinthe Rigaud, Antoine-Jean Gros, Jacques-Louis David, and Sir Thomas Lawrence were required to express nobility, grace, and authority, just as the sultans and rajahs portrayed on frontispieces to Persian and Indian illuminated books and albums had understandably to be flattered as benevolent despots. Such concessions to the sitter's vanity and social position seem to have been disregarded, however, in the convincing likenesses by more objective realists such as Robert Campin, Dürer, Jan van Eyck, Velázquez, Goya, and Gustave Courbet. Probably the finest are the self-portraits and studies of ordinary people by Rembrandt and Van Gogh, where psychological insight, emotional empathy, and aesthetic values are fused. A more decorative approach to the subject is seen in the flattened portraits by Holbein, the Elizabethan and itinerant naïve U.S. limners, and the Far Eastern paintings of ancestors, poets, priests, and emperors. Like these paintings, the full-length portraits by Boucher, Gainsborough, Kees van Dongen, and Matisse display as much regard for the texture and form of their sitters' dress as for their facial features.

Photographs and painted portraits

Since the development of photography, however, portraiture has been rarely practiced for its own sake as a serious art form, except where artists such as Cézanne and Braque have used it as a subject for structural research or—like Amedeo Modigliani, Chaim Soutine, and Francis Bacon—for the expression of a personal vision beyond the scope of the camera.

Genre. Genre subjects are scenes from everyday life. Hunting expeditions and tribal rituals figure in prehistoric rock paintings. Domestic and agricultural occupations, with banquet scenes of feasting, dancing, and music, were traditional subjects for ancient Egyptian tomb murals. Far Eastern hand scrolls, albums, and screens brilliantly describe court ceremonies, the bustle of towns, and the hardships of the countryside. The depiction of earthly pursuits was forbidden under the strict iconography prescribed by the early Christian Church, but the later illu-

minated Books of Hours provide enchanting records of the festivals and occupations of northern European communities. In Renaissance painting, genre subjects were generally restricted to background features of portraits and historical narratives. Domestic scenes, however, not only provided Bruegel with subjects for moral allegories but, as with Rembrandt, were used to counterpoint the emotional intensity of a dramatic religious theme. The withdrawal of religious patronage in northern Europe directed painters toward secular subjects. The rich period of genre painting in 17th-century Holland is represented by the interiors, conversation pieces, and scenes of work and play by David Teniers the Younger, Frans Hals, Jan Steen, Gerard Terborch, Pieter de Hooch, Adriaen van Ostade, and, the finest, by Jan Vermeer. Pictures of rustic life had a special appeal for collectors in 18th-century France and England; these were the somewhat picturesque representations of peasant life painted by Jean-Baptiste Greuze, Boucher, George Morland, and Gainsborough. Jean-Baptiste-Siméon Chardin's paintings of servants and children, however, exhibit a timeless dignity and grandeur. The harsher realities of working life were depicted by Jean-François Millet, Daumier, Courbet, Van Gogh, and Degas; the robust gaiety of cafés and music halls was captured by Toulouse-Lautrec, John Sloan, Everett Shinn, and Walter Richard Sickert; and intimate domestic scenes were recorded by Bonnard and Vuillard. Modern genre movements have included the American Scene painters, the Ashcan and Kitchen Sink schools (represented by such painters as George Wesley Bellows, Jack Smith, and Derrick Greaves), the Camden Town and Euston Road groups (Frederick Spencer Gore, Sir William Coldstream, and Victor Pasmore), and the Social Realists in England and in the United States (Robert Henri, Stuart Davis, and Maurice Prendergast).

Landscape. Idealized landscapes were common subjects for fresco decoration in Roman villas. Landscape painting (as exemplified by a Chinese landscape scroll by Ku K'aichih dating from the 4th century) was an established tradition in the Far East, where themes such as the seasons and the elements held a spiritual significance. In Europe, imaginary landscapes decorated 15th-century Books of Hours. The first naturalistic landscapes were painted by Dürer and Bruegel. Landscapes appeared in most Renaissance paintings, however, only as settings to portraits and figure compositions. It was not until the 17th-century Dutch and Flemish schools—of Rembrandt, Jacob van Ruisdael, Meindert Hobbema, Aelbert Cuyp, Rubens, and Hercules Seghers—that they were accepted in the West as independent subjects. The most significant developments in 19th-century painting, however, were made through the landscapes of the Impressionists and the Neo-Impressionists and Postimpressionists. Styles in landscape painting range from the tranquil, classically idealized world of Poussin and Claude, the precise, canal topography of Francesco Guardi and Canaletto, and the structural analyses of Cézanne to the poetic romanticism of Samuel Palmer and the later Constables and Turners and the exultant pantheism of Rubens and Van Gogh. Modern landscapes vary in approach from the Expressionism of Oskar Kokoschka's cities and rivers, Maurice de Vlaminck's wintry countrysides, and John Marin's crystalline seascapes to the metaphysical country of Ernst, Dalí, and René Magritte and the semi-abstract coastlines of Nicolas de Staël, Maria Helena Vieira da Silva, and Richard Diebenkorn.

Still life. The earliest European still-life painting is usually attributed to Jacopo de' Barbari (*i.e.,* "Dead Bird," 1504). In Western paintings, still life often appears as a minor feature of the design; but until the 17th century it was not generally painted for its own sake, although it was already traditional to Far Eastern art. The subject is particularly associated with northern European painting, and the choice of objects very often has a religious or literary significance: wine, water, and bread symbolizing the Passion; skulls, hourglasses, and candles, the transience of life; and selected flowers and fruits, the seasons. Flower painting, especially, held a spiritual and emotional meaning for Japanese artists and for 19th-century European painters, such as Odilon Redon, Paul Gauguin, and Van Gogh. Still life has been expressed in many different ways: Giuseppe Arcimboldo's witty arrangements of fruit, flowers, and vegetables made into fantastic allegorical heads and figures; the sensuous representation of food by Frans Snyders, Goya, and William Merritt Chase; the trompe l'oeil illusionism of François Desportes and William Harnett; the formal decoration of folk artists or primitives such as Henri Rousseau and Séraphine and of modern painters such as Matisse, Dufy, and Pat Caulfield; the semi-abstract designs of Picasso, Gris, and William Scott; and, probably at its highest level of expression, the majestic still lifes of Chardin, Cézanne, and Giorgio Morandi.

Subject matter of still lifes

Other subjects. Since ancient times, animals and birds have provided the primary subject matter of a painting or have been included in a design for their symbolic importance. In the paintings of prehistoric caves and dynastic Egyptian tombs, for example, animals are portrayed with a higher degree of naturalism than human figures. Their texture, movement, and structure have provided some artists with a primary source of inspiration: the classical, anatomical grace of a George Stubbs racehorse and a more romantic interpretation in the ferocious energy of a Rubens and Géricault stallion; the vivid expression of rhythmically co-ordinated movements of deer by Tawaraya Sotatsu and Antonio Pisanello; the weight and volume of George Morland's pigs and Paul Potter's cows; the humanized creatures of Gothic bestiaries and of Edward Hicks's "Peaceable Kingdom"; and, finally, Dürer's "The Hare," which is possibly as famous as Leonardo's "Mona Lisa."

Increasing interest is shown in notable painters' versions of other artists' works. These are not academic copies (such as the study made by Matisse, when a student, of Chardin's "La Raie"), but creative transcriptions. Examples that can be appreciated as original paintings are those by Miró of Sorgh's "Lute Player"; by Watteau of Rubens' "Apotheosis of James I"; by Degas of Bellini's "Jealous Husband"; by Caulfield of Delacroix's "Greece Expiring on the Ruins of Missolonghi"; by Larry Rivers of Jean-Auguste-Dominique Ingres's "Mlle Rivière"; and by Picasso of Manet's "Déjeuner sur l'herbe," Velázquez' "Las Meñinas," and Delacroix's "Woman of Algiers" (which produced Roy Lichtenstein's "Femmes d'Alger, After Picasso, After Delacroix"). Picasso has also painted free versions of works by El Greco, Lucas Cranach, Poussin, and Courbet, as Rubens had of Mantegna and Titian, Rembrandt of Persian and Indian miniatures, Cézanne of Rubens and El Greco, and Van Gogh of Millet, Gustave Doré, and Delacroix.

The subject of an abstract painting

In an abstract painting, ideas, emotions, and visual sensations are communicated solely through lines, shapes, colours, and textures that have no representational significance. The subject of an abstract painting may be therefore a proposition about the creative painting process itself or exclusively about the formal elements of painting, demonstrating the behaviour of juxtaposed colours and shapes and the movements and tensions between them, their optical metamorphosis and spatial ambiguities. Many abstracts, however, are more than visual formal exercises and produce physical and emotional reactions in the spectator to illusions of shapes and colours that appear to rise and fall, recede and advance, balance and float, disintegrate and re-form; or of moods created of joy, sadness, peace, or foreboding; or of effects produced by light or by flickering or throbbing movement. Some abstracts evoke the atmosphere of a particular time, place, or event; and then their titles may be significant: "Pancho Villa, Dead and Alive" (Robert Motherwell); "Late Morning" (Bridget Riley); "Broadway Boogie Woogie" (Mondrian); "Gold of Venice" (Lucio Fontana); "Capricious Forms" (Kandinsky).

SYMBOLISM

Most early cultures developed iconographic systems that included prescriptions for the site, design, function, form, medium, subject matter, and imagery of their painting. The siting of early Byzantine murals, for instance, echoed the symbolic, architectural planning of the basilica. Thus, a stylized, linear image of Christ, surrounded by heavenly

hosts, occupied the central dome; the Virgin was represented in the apse; and stiff figures of apostles, prophets, martyrs, and patriarchs occupied the aisle walls. The format of early devotional paintings was also prescribed, Christian and Buddhist deities being placed in the focal centre of the design, above the eye level of the audience and larger than surrounding figures. And, in the conventional arrangement of a Christian subject such as the Holy Trinity, a central, bearded, patriarchal God, flanked by archangels, presented Christ on the cross; between them was a dove, representing the Holy Spirit. In a rendering of the risen Christ, the Son faced the audience, with the Virgin Mother on the left and St. John on the right of the design. In the Far East a traditional format depicted Buddha on a lotus throne or in a high chariot drawn by oxen across clouds, surrounded by figures representing the planets. Deities generally appear against undefined grounds of white (signifying eternity or nothingness), blue (the celestial vaults), or gold (representing heavenly light by radiating lines or the spiritual aura by a nimbus). The elaborate surface preparation of supports and the painstaking execution with the finest materials symbolized the intention that paintings dedicated to a deity should last forever. The imagery, subject matter, and form might also have a mystical function: the realistic rendering of animals in contrast to the perfunctory human representations in Ice Age rock paintings, thought to signify a wishful guarantee for success in hunting; the earthly pleasures depicted on ancient Egyptian tomb murals intended to secure their continuance for the deceased; and the North American Indian sand paintings designed for magic healing ceremonies and the Tantric (relating to Tantrism, a school of Mahāyāna Buddhism) mandalas used for meditation and enlightenment.

Symbolism in Eastern painting—intended to deepen the experience of a picture's mood and spirituality—is more generalized and poetic than in Western art. Both the execution and the subject matter of Buddhist Chinese and Japanese painting have a religious or metaphysical significance: the artist's intuitive, calligraphic brush movements symbolizing his mystical empathy with nature and his cyclic landscape and flower subjects expressing his belief in the spiritual harmony of natural forms and forces. Much of Indian symbolism is visually emotive, images such as snakes, plantain leaves, twining creepers, and rippling water being overtly sexual. And, although symbolic attributes and colour codes identify Indian mythological characters (for example, the four arms of the terrible goddess Kālī and the blue skin of the divine lover Krishna), the formal character and colour scheme of settings generally reflect the narrative's emotional mood (for example, vibrant, dark-blue, cloudy skies and embracing, purple-black glades evoking amorous anticipation and red grounds expressing the passions of love or war).

Western symbolic systems

Western symbolic systems, however, are more intellectually directed, their imagery having precise literary meanings and their colour codes intended primarily for narrative or devotional identification. The iconographic programs of the early Christian churches, for example, laid down complex formulas for the viewpoints, gestures, facial expressions, and positions of arms, hands, and feet for religious figures. An elaborate Ethiopian Christian iconographic system was followed until very recently, and elsewhere traditional methods survive of identifying archangels and saints by their attributes and by the symbols of martyrdom that they display: distinguishing white-bearded St. Peter from black-bearded St. Paul, for example, and portraying St. Catherine with a wheel and St. Bartholomew with a knife and skin. Christian iconography adopted and elaborated Greco-Roman and Jewish symbolic imagery: the pagan signs of the vine and the fish, for example, and the image of Christ as the Good Shepherd based on the Greek Hermes Kriophoros. Medieval and Renaissance writings define an immense vocabulary of symbolic images, such as the crescent, sea urchin, and owl signifying heresy, the toad and jug representing the devil, and the egg and bagpipes as erotic symbols (all of which appear in Hieronymus Bosch's 15th-century narrative moralities). Angels and devils, hellfire and golden paradise, heavenly skies and birds in flight representing spirituality and rebirth are examples of the similarity of symbolic meaning for many religious, mythological, and allegorical traditions. The significance of images common to several cultures, however, may also be very different: the dragon representing avarice in European medieval allegory symbolizes friendliness in Japanese Zen painting; and the snake, symbol of temptation and eroticism in the West, signifies, by its skin shedding, the renewal of life in Far Eastern iconography.

BIBLIOGRAPHY

General reference: Both KIMBERLEY REYNOLDS with RICHARD SEDDON, *Illustrated Dictionary of Art Terms* (1981, reissued 1984); and RALPH MAYER, *A Dictionary of Art Terms and Techniques* (1969, reissued 1981), include modern art references. Two works by HAROLD OSBORNE (ed.), *The Oxford Companion to Art* (1970), and *The Oxford Companion to Twentieth-Century Art* (1981), provide extensive bibliographies. Also see RENÉ HUYGHE (general ed.), *Larousse Encyclopedia of Prehistoric and Ancient Art*, rev. ed. (1966, reissued 1981); *Larousse Encyclopedia of Byzantine and Medieval Art*, rev. ed. (1966, reissued 1981); *Larousse Encyclopedia of Renaissance and Baroque Art* (1964, reissued 1981); and *Larousse Encyclopedia of Modern Art* (1965, reissued 1981).

Design: Good surveys of the subject include FREDERICK MALINS, *Understanding Paintings: The Elements of Composition* (1981); and JOHANNES ITTEN, *Design and Form: The Basic Course at the Bauhaus*, rev. ed. (1975; originally published in German, 1963). Works on colour include JOSEF ALBERS, *The Interaction of Color* (1963, reissued with rev. plate section, 1975); JOHANNES ITTEN, *The Art of Color* (1961, reissued 1973; originally published in German, 1961); FABER BIRREN, *Creative Color* (1961), and (ed.), *A Grammar of Color: A Basic Treatise on the Color System by Albert H. Munsell* (1969); ROBERT L. HERBERT, *Neo-Impressionism* (1968); WILLIAM INNES HOMER, *Seurat and the Science of Painting* (1964, reprinted 1978); and BARBARA ROSE, "The Primacy of Color," *Art International*, 8:22–26 (1964). The influence of photography on painting is examined in AARON SCHARF, *Art and Photography* (1968, reissued 1974); and KAREN TSUJIMOTO, *Images of America: Precisionist Painting and Modern Photography* (1982).

Mediums: Standard works on most painting materials, supports, surfaces, and techniques include RALPH MAYER, *The Artist's Handbook of Materials and Techniques*, 4th rev. ed. (1982), with extensive bibliography; HILAIRE HILER, *The Painter's Pocket Book of Methods and Materials*, 3rd ed. rev. by COLIN HAYES (1970); KURT HERBERTS, *The Complete Handbook of Artist's Techniques* (1958; trans. from the German); MARIA BAZZI, *The Artist's Methods and Materials* (1960; originally published in Italian, 1956), with a bibliography of important treatises on mediums and techniques; FREDERIC TAUBES, *A Guide to Traditional and Modern Painting Methods* (1963); and MAX DOERNER, *The Materials of the Artist and Their Use in Painting*, rev. ed. (1949, reprinted 1969; originally published in German, 4th ed., 1933). For discussion of tempera, see the appropriate sections in DANIEL V. THOMPSON, *The Materials of Medieval Painting* (1936; reprinted as *Materials and Techniques of Medieval Painting*, 1956). Watercolour painting is treated in WALTER KOSCHATZKY, *Watercolor: History and Technique* (1970; originally published in German, 1969). For treatment of ink painting, see FEI CH'ENG WU, *Brush Drawing in the Chinese Manner* (1957); and OSVALD SIRÉN, *Chinese Painting: Leading Masters and Principles*, 7 vol. (1973). FRED GETTINGS, *Polymer Painting Manual* (1971), is a thorough and well-illustrated guide to acrylic painting. Works on other mediums include JEAN GUICHARD-MEILI, *Matisse Paper Cutouts* (1984; originally published in French, 1983); and HARRIET JANIS and RUDI BLESH, *Collage*, rev. ed. (1967).

Forms: WILLIAM G. ARCHER, *Indian Miniatures* (1960), and *Indian Paintings from the Punjab Hills: A Survey and History of Pahari Miniature Painting* (1973); MARK ZEBROWSKI; *Deccani Painting* (1983); JANET WOODBURY ADAMS, *Decorative Folding Screens: In the West from 1600 to the Present Day* (U.S. title: *Decorative Folding Screens: 400 Years in the Western World*, 1982); ELISE GRILLI, *The Art of the Japanese Screen* (1971); ROSELEE GOLDBERG, *Performance: Live Art, 1909 to the Present* (1979); and ADRIAN HENRI, *Environments and Happenings* (U.S. title: *Total Art: Environments, Happenings, and Performance*, 1974).

Imagery: E.H. GOMBRICH, *Art and Illusion*, 4th ed. (1972); GYORGY KEPES (ed.), *Sign, Image, Symbol* (1966); LEON M. ZOLBROD, *Haiku Painting* (1983); LUCY R. LIPPARD *et al.*, *Pop Art* (1960, rev. ed. 1970); and J.H. MATTHEWS, *Eight Painters: The Surrealist Context* (1982).

Subject matter: HOWARD DANIEL, *Encyclopedia of Themes and Subjects in Painting* (1971), a concise survey of Western

mythological and religious subjects; DAVID ROSAND, *Painting in Cinquecento Venice* (1982); SVETLANA ALPERS, *The Art of Describing: Dutch Art in the Seventeenth Century* (1983); NORMAN BRYSON, *Word and Image: French Painting of the Ancient Régime* (1982); MICHAEL LEVEY, *The Painter Depicted: Painters as a Subject in Painting* (1982); JOHN POPE-HENNESSY, *The Portrait in the Renaissance* (1966, reissued 1979), includes interpretive discussions of the works and extracts from the artists' letters; KENNETH CLARK, *Landscape into Art* (1949, reissued 1975); A. RICHARD TURNER, *The Vision of Landscape in Renaissance Italy* (1966, reprinted 1974); JOSEPH S. CZESTOCHOWSKI, *The American Landscape Tradition* (1982); ROGER BOULET, *The Canadian Earth: Landscape Paintings by the Group of Seven* (1982); MICHAEL JACOBS, *Nude Painting* (1979); and KENNETH CLARK, *The Nude: A Study of Ideal Art* (1976).

Symbolism: ERWIN PANOFSKY, *Studies in Iconology* (1939, reissued 1972), and *Meaning in the Visual Arts* (1955, reprinted 1982); F.D.K. BOSCH, *The Golden Germ* (1960); CARL G. JUNG *et al., Man and His Symbols* (posthumous ed. 1964, reprinted 1979), with excellent illustrations; RUDOLF WITTKOWER, *Allegory and the Migration of Symbols* (1977), includes Eastern imagery; PAUL FRANKL, *The Gothic: Literary Sources and Interpretations Through Eight Centuries* (1960); GEORGE FERGUSON, *Signs and Symbols in Christian Art* (1954, reprinted 1973); JOAN EVANS, *Monastic Iconography in France: From the Renaissance to the Revolution* (1970); and JITENDRA NATH BANERJEA, *The Development of Hindu Iconography*, 2nd rev. ed. (1956, reprinted 1974).

Writings: IRMA A. RICHTER (ed.), *Selections from the Notebooks of Leonardo da Vinci* (1952, reprinted 1977), illustrated; GIORGIO VASARI, *Vasari on Technique*, ed. by G. BALDWIN BROWN, trans. by LOUISA A. MACLEHOSE (1907, reprinted 1961); *The Mustard Seed Garden Manual of Painting, 1679–1701*, included in MAI-MAI SZE, *The Tao of Painting*, 2nd ed., 2 vol. (1963); ELIZABETH G. HOLT (ed.), *A Documentary History of Art*, 2nd ed., 3 vol. (1957–65); JOHN REWALD (ed.), *Cézanne's Letters*, 5th ed. (1982); VINCENT VAN GOGH, *The Complete Letters of Vincent van Gogh*, 2nd ed., 3 vol. (1978); FERNAND LÉGER, *Functions of Painting*, ed. by EDWARD F. FRY (1973; originally published in French, 1965); ROBERT DELAUNAY and SONIA DELAUNAY, *The New Art of Color*, ed. by ARTHUR A. COHEN (1978); ROBERT MOTHERWELL (ed.), *The Dada Painters and Poets: An Anthology*, 2nd ed. (1981); MARCEL JEAN (ed.), *The Autobiography of Surrealism* (1980); WASSILY KANDINSKY, *Kandinsky, Complete Writings on Art*, 2 vol., ed. by KENNETH C. LINDSAY and PETER VERGO (1982), with the original illustrations; PAUL KLEE, *Pedagogical Sketchbook* (1953, reprinted 1977; originally published in German, 1925), *On Modern Art* (1948, reprinted 1966; originally published in German, 1945), *The Thinking Eye*, ed. by JÜRG SPILLER (1961, reprinted 1969; originally published in German, 1956), and *The Diaries of Paul Klee, 1898–1918*, ed. by FELIX KLEE (1964, reissued 1968; originally published in German, 1957); MARCEL DUCHAMP, *The Essential Writings of Marcel Duchamp: Salt Seller*, ed. by MICHEL SANOUILLET and ELMER PETERSON (1975; originally published in French, 1958), and *Marcel Duchamp, Notes*, ed. by PAUL MATISSE (1983); HENRI MATISSE, *Notes of a Painter*, included in ALFRED H. BARR, *Matisse: His Art and His Public* (1951, reprinted 1966); EDWARD F. FRY, *Cubism* (1966, reprinted 1978; trans. from the French and German); PIERRE DAIX, *Cubists and Cubism* (1982; originally published in French, 1982); PABLO PICASSO, *Picasso on Art: A Selection of Views*, comp. by DORE ASHTON (1972); KATHARINE KUH, *The Artist's Voice: Talks with Seventeen Artists* (1962); and *The New York School*, foreword by MAURICE TUCHMAN (1970), with an extensive bibliography.

(P.D.O.)

The History of Western Painting

Painting, the execution of forms and shapes on a surface by means of pigment (but see also DRAWING for discussion of depictions in chalks, inks, pastels, and crayons), has been continuously practiced by humans for some 20,000 years. Together with other activities that may have been ritualistic in origin but have come to be designated as artistic (such as music or dance), painting was one of the earliest ways in which man sought to express his own personality and his emerging understanding of an existence beyond the material world. Unlike music and dance, however, examples of early forms of painting have survived to the present day. The modern eye can derive aesthetic as well as antiquarian satisfaction from the 15,000-year-old cave murals of Lascaux—some examples testify to the considerable powers of draftsmanship of these early artists. And painting, like other arts, exhibits universal qualities that make it easy for viewers of all nations and civilizations to understand and appreciate.

The major extant examples of early painting anywhere in the world are found in western Europe and the Soviet Union. But some 5,000 years ago, the areas in which important paintings were executed shifted to the eastern Mediterranean Sea and neighbouring regions. For the purposes of this article, therefore, Western painting is to be taken as signifying painting not only in Europe but also in regions outside Europe that share a European cultural tradition—the Middle East and Mediterranean Basin and, later, the countries of the New World.

Western painting is in general distinguished by its concentration on the representation of the human figure, whether in the heroic context of antiquity or the religious context of the early Christian and medieval world. The Renaissance extended this tradition through a close examination of the natural world and an investigation of balance, harmony, and perspective in the visible world, linking painting to the developing sciences of anatomy and optics. The first real break from figurative painting came with the growth of landscape painting in the 17th and 18th centuries. The landscape and figurative traditions developed together in the 19th century in an atmosphere that was increasingly concerned with "painterly" qualities of the interaction of light and colour and the expressive qualities of paint handling. In the 20th century these interests contributed to the development of a third major tradition in Western painting, abstract painting, which sought to uncover and express the true nature of paint and painting through action and form.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 628. (Ed.)

This article is divided into the following sections:

European Stone Age 337
Aegean and eastern Mediterranean Metal Age 337
  Early Bronze Age (3000–2000 BC) 338
  Middle Bronze Age (2000–1600 BC) 338
  Late Bronze Age (1600–1100 BC) 338
Ancient Greek 339
  Dark Ages (1200–900 BC) 339
  Geometric period (*c.* 900–700 BC) 339
  Orientalizing period (*c.* 700–625 BC) 340
  Archaic period (*c.* 625–500 BC) 340
  Classical period (*c.* 500–323 BC) 341
  Hellenistic period (*c.* 323–1st century BC) 342
Western Mediterranean 342
  Etruscan 342
  Roman 343
    Pagan
    Early Christian
    Book illustration in antiquity
Eastern Christian 346
  Early Byzantine period (330–717) 346
  Iconoclastic Age (717–843) 347
  Middle Byzantine period (843–1204) 347
  Late Byzantine period (1204–1453) 347
  Post-Byzantine Russia 349
  Regional variations in eastern Christian painting 349
Western Dark Ages and medieval Christendom 350
  Dark Ages 350
    Rome and Italy, *c.* 600–850
    England and Ireland, *c.* 650–850
    Merovingian Gaul
  Early Middle Ages 351
    Carolingian Empire
    The 10th century
    Late Anglo-Saxon England
    France
    Ottonian Germany
  Romanesque 353
    Italy
    France
    England
    Spain
    The Meuse Valley
    Germany and Austria
  Late 12th century 354
  Gothic 355
    Early Gothic
    High Gothic
    Italian Gothic
    International Gothic
    Late Gothic
Renaissance 358
  Early Renaissance in Italy 359
  High Renaissance in Italy 362
  Italian Mannerism and late Renaissance 365
  Renaissance outside Italy 366
    France
    Spain
    Germany
    Low Countries
    Bohemia
Baroque 368
  Early and High Baroque in Italy 368
  Late Baroque and Rococo 369
    Spain and Portugal
    Low Countries
    France
    Britain
    Colonial Americas
    Central Europe
    Poland
    Russia
    Scandinavia
Neoclassical and Romantic 375
  Neoclassicism 375
    Britain
    France
    Germany and Austria
    Italy
    Other countries
  Romanticism 376
    Britain
    Germany
    France
    United States
    Russia
Modern 380
  Origins in the 19th century 381
    Impressionism
    Symbolism
    The end of the 19th-century tradition
  The 20th century 384
    Cubism and its consequences
    Fantasy and the irrational
    After 1945
    New forms
Bibliography 389

European Stone Age

Upper Paleolithic. During the Upper Paleolithic Period, just before the final retreat of the glaciers at the end of the last Ice Age (15,000–10,000 BC), much of Europe was peopled by small bands of nomadic hunters preying on the migratory herds of reindeer, cattle, bison, horses, mammoth, and other animals whose bodies provided them with food, clothing, and the raw materials for tools and weapons. These primitive hunters decorated the walls of their caves with large paintings of the animals that were so important for their physical well-being. Most surviving examples of such murals have been found in France and Spain (see PREHISTORIC PEOPLES AND CULTURES), but similar figures from caves in the Ural Mountains in the Soviet Union may indicate that the practice was more widespread than has been supposed.

Ever since the first examples of these paintings came to light in the late 19th century, they have excited admiration for their virtuosity and liveliness. The simplest figures are mere outline drawings, but the majority combine this technique with sophisticated shading and colour washes that modulate the surface and suggest the differing textures of pelts, horn, and bone. Volume is indicated by carefully controlled changes in the thickness of brushstrokes, and the astonishingly advanced draftsmanship conveys a considerable sense of movement and life. Most of the animals were originally depicted as individual figures without narrative import, and what appear to the modern observer to be sophisticated groupings of figures are, in reality, the end result of a long additive process.

The lack of a clear narrative element in these paintings has caused problems in their interpretation. Man is seldom portrayed, and depictions of human figures unambiguously interacting with the numerous animal figures are rare. One of the few exceptions to this rule is a scene at Lascaux in southern France depicting a bison butting a falling male figure. The "Sorcerer" at Les Trois Frères, also in southern France, is more characteristic. Although he is draped in the skin of an animal and seems to be engaged in stalking or a ritual dance, his complete isolation from any other figure leaves his exact significance unclear. It is also interesting that, in contrast to the obvious care taken in the detailed portrayal of animals, the few human figures are usually executed in a perfunctory and schematized fashion. Sometimes the only hint of man is provided by depictions of darts wounding or killing a few of the animal figures. These projectiles have been interpreted as exercises in sympathetic magic designed to induce success in a future hunt. Conversely, they might just as easily commemorate past kills. But certain features suggest that such simple explanations do not tell the whole story: first, such portrayals are rare (in inverse proportion to the amount of scholarly discussion they have engendered) and, second, the beasts that are shown as wounded—indeed the vast majority of the species depicted on the cave walls—were not significant items in the diet of the cave artists. Contemporary habitation deposits indicate that most of the meat consumed came from reindeer, and reindeer appear almost as infrequently as man himself among the surviving paintings. One fact is clear: individual initiative seems paramount, both in the execution of the animal figures and in the recording of the activities of the isolated humans. Any hint of social interaction is absent, and it has been assumed that society as such existed at a relatively low level. Nature provided the impetus for change, and in the art of the following period man finally emerged as part of a community.

Mesolithic. At the end of the Ice Age the great herds that had provided sustenance for the Paleolithic hunters disappeared from France and Spain. Forests cloaked the landscape and harboured much smaller groups of deer and related species. These were fleet and elusive and, in consequence, much more difficult to hunt and kill. Thus, although the climate was warmer than before, it was much harder to live by hunting alone. Man had to modify his hunting techniques and forage for the seeds and fruits that the forests provided, or the fish and shellfish that he could find in rivers or on the coasts. Cooperation was essential, and the new situation is clearly reflected in the art of the period.

In the southern and eastern parts of what is now Spain, small bands of such hunter-gatherers left a record of their activities in the rock shelters where they camped periodically. In some ways the new paintings resemble the old. Although a simple silhouette technique for the most part replaced the outline and shading techniques of the Paleolithic style, facility of brushwork and accuracy of observation continued to imbue the new creations with a vivacity and sense of movement similar to those of their predecessors. There are obvious conceptual differences between the two artistic complexes, however. The new paintings constitute the first real compositions having a clear narrative meaning, and man finally emerges as the chief actor in the dramas played out on the rock walls. At Remigia three hunters are depicted stalking a leaping ibex, while at Los Caballos a line of archers fires arrows into a small herd of panic-stricken deer, presumably driven into the ambush by beaters. Scenes of battle or groups of dancers also occur, while social status is implied in a carefully executed archer found at Santolea: he is dressed in painstakingly portrayed finery and is flanked by two other figures. This emphasis on man is new, but even more significant is the element of cooperation as part of a group whose social cohesion in warfare, hunting, or ritual was probably necessary if the group was to survive and prosper.

Social status implied

Neolithic. The subsequent Neolithic Period saw the introduction from western Asia of farming and the raising of domesticated animals. The new way of life appeared in the Balkans sometime before 6000 BC and rapidly spread across Europe. For the first time man was able to live a relatively settled village life and accumulate a wide range of household goods. So far as large-scale painting is concerned, however, this period is something of a disappointment. Thus far, there is no evidence that the farming communities decorated their house walls with painted designs, in this at least failing to imitate their Asiatic mentors whose walls, as in the shrines at Catalhüyük in Turkey, were often embellished with ambitious decorative schemes. In different places and at different times the European farmers did indeed indulge their aesthetic drive by producing highly decorated painted pottery whose patterns reflect contemporary basketwork or textiles. Few of these styles include human or animal figures and, despite their undoubted charm, these vases are the products of craft traditions that have little to do with large-scale art.

Aegean and eastern Mediterranean Metal Age

In Greece and the Aegean, influence from the adjacent areas of western Asia helped promote the rise of small towns by about 3000 BC. The cultural development is usually divided into three separate strands: Minoan on Crete, Cycladic on the islands of the central Aegean, and Helladic on the Greek mainland. A fourth area, Cyprus, is often included in this development, though its culture was closer to those of Syria and Asia Minor and it was only during the 13th century BC that Greek invaders brought Cyprus fully into the Aegean orbit. (P.J.C.)

The Metal Age in Europe started in the early 3rd millennium BC, when the peoples around the Aegean Sea began to work copper, under the influence of the neighbouring peoples of western Asia. By 2500 BC coppersmiths were also active across the Alps. Bronze began to be used in Europe at the beginning of the 2nd millennium BC, and iron was used in Greece by the 11th century BC and north of the Alps by the 8th century BC. Bronze was always a luxury item because the sources of its constituent metals, tin and copper, occurred in scattered deposits, often far from the producing centres. Its use, therefore, encouraged trade. But iron, when it came into use, was cheaper and easier to work; moreover, the ore lodes were often close at hand. Its use, especially for agricultural implements, allowed more intensive exploitation of the countryside, especially those areas where heavy soils had precluded farming with more primitive tools. The end of this period is usually placed at the point where written records supplement the archaeo-

logical record. In Greece and Italy this happened during the 8th century BC or a little later, whereas in northwestern Europe the Celtic and Germanic peoples had to wait for the Roman conquests of the 1st century BC before emerging into history. Beyond the imperial frontiers old patterns continued longer. Throughout this long period it was the Mediterranean, with its flourishing towns and cities, that produced major works of painting.

(O.K.-J./P.J.C.)

EARLY BRONZE AGE (3000–2000 BC)

Early Minoan. In Crete the Early Minoan peoples lived in small towns and villages with a basically agricultural economy. Although traces of their houses have been preserved, it is clear that they did not paint their walls with decorative designs. The fine plaster introduced at this time, however, did provide the basis for later developments. Their pottery was at first plain or decorated with simple, arresting patterns of straight lines. (Pottery is an important source for modern knowledge of painting in the last three millennia BC because, although fired clay objects—even when decorated—may be broken, they are not easily pulverized, so many fragments have survived.) In the following phase (2500–2200 BC) a similar style flourished, though other vases with a mottled surface imitating variegated stones were produced. During the third phase (2200–2000 BC) most fine vases were decorated with designs in white or cream paint on a dark ground. Elegant running spirals and other curvilinear motifs, as well as the occasional use of other colours, revolutionized the style and paved the way for the greater advances of the Middle Minoan period.

Running spirals and other curvilinear motifs

Early Cycladic. In the islands there was little interest in painted designs. Most decoration consisted of incised or impressed geometric schemes, though there were some vases with similar designs in paint. The typical pottery of the second and third phases (2500–2000 BC) was decorated in semilustrous paint, either as an allover wash or in angular patterns.

Early Helladic. On the Greek mainland there was a similar lack of interest in painted decoration on pots. Although monumental buildings have been found in the Peloponnese dating to the Early Helladic II period (2500–2200 BC), none of these had decorated walls. New settlers arrived about 2200 BC and destroyed the old centres of power. Their houses were primitive affairs and only a few of their finer vases bore painted designs, these being of straight lines or other simple patterns.

MIDDLE BRONZE AGE (2000–1600 BC)

Middle Minoan. The Middle Minoan period saw the evolution of a monarchical society based on palaces situated in the most fertile districts of Crete. There were undoubtedly frescoes in these large buildings before 1600 BC, but little survived the disastrous earthquake of about 1700 BC, and once again it is the pottery that gives the best idea of contemporary aesthetics. The decorative style is basically a development of the previous period's. Curvilinear patterns in white, yellow, and red swirl around the surfaces of these bulbous vases. The latest Middle Minoan style is similar, but its static formality seems better suited to wall decoration, and it is likely that monumental frescoes from the old palaces influenced the vase painter. The combination of modeled flowers and animals with painted motifs on the vases certainly reflects similar developments in wall painting, where stucco reliefs were combined with simple painted backgrounds.

Middle Cycladic and Middle Helladic. On the mainland and in the islands, native styles of plain or simply painted pottery continued to be executed, but Cretan influence was felt toward the end of the period in both areas, and they began to be drawn into the wider cultural orbit characteristic of the following period.

LATE BRONZE AGE (1600–1100 BC)

Wall paintings. The three separate areas of the Aegean were brought into intimate contact during the Late Bronze Age; indeed the whole eastern Mediterranean saw intense cross connections and cultural diffusion. Great palaces arose on the Greek mainland and Crete and even on some of the lesser islands. Although there were probably differences in the natures of the societies that built them (resulting in fortified structures on the Greek mainland and unfortified ones on Crete), the palaces and great houses were decorated with complex frescoes whose style was based on Cretan models. Many of the figured scenes are merely decorative and depict landscapes with birds and animals or figures gathering flowers. Others show ceremonies connected with a cult or the court (Plate 1, "The Toreador Fresco") and were probably useful in bolstering the power of the royal or priestly classes. The style is a combination of dark outline drawing, to delimit the object shown, and solid painted areas within it. On some birds and animals the feathers or pelts are imitated by slightly more impressionistic brushstrokes. Most of these frescoes are in fragmentary condition, but a better idea of what they must once have looked like can be gained from the house walls at Akrotíri on Thera (one of the Cyclades of the southern Aegean). Thera was destroyed by volcanic eruption during the 15th century BC and is often referred to as the Greek Pompeii. The wall paintings there were heavily influenced by those of Crete, both as to style and subject matter, though the popularity of outline figures on a pale background stemmed from the local pottery tradition. One of the most exciting discoveries is a long frieze depicting a fleet of gaily decorated ships sailing against a backdrop of hilly islands with towns, shepherds, and hunters scattered along the shores or set upon the forested peaks among gushing streams. Another painting shows a group of women at a religious festival and—in the first known instance at this period—ordinary people: two boys boxing and a fisherman proudly displaying his catch (Plate 1). These paintings decorated well-to-do houses. In the great palaces of Crete and on the Greek mainland many of the scenes are rather more formal. At Knossos on Crete there are long lines of offering bearers in the vestibule leading to the state rooms. The throne in one ritual chamber is flanked by fresco paintings of griffins whose presence must have had a protective value. Griffins also flank the throne at Pylos in Greece, and the same site has produced fragments of another fresco showing battle scenes. Mycenae (also on the Greek mainland) possesses a small sanctuary whose walls are decorated with ritual episodes, and religious ceremonies do indeed appear to have been an important part of the wall painters' repertoire. There are, however, none of the historical or annalistic scenes so characteristic of the palaces and temples of western Asia and Egypt. In particular there are no depictions of investitures or battles with accompanying inscriptions; in short, Aegean paintings are far less bombastic than their Middle Eastern equivalents. This is not to say that the visitor would have been less impressed by the ruler's power in these first great European civilizations, merely that the iconographic emphases were different. (P.J.C.)

Differences between Asia and Europe

Vase paintings. *Late Minoan.* The light-on-dark style of pottery was by now replaced by dark-on-light ornamentation. At first (roughly 1600–1500 BC), curvilinear patterns and simple designs of vegetation predominated. Between 1500 and about 1450 BC, however, there flourished the Marine style, possibly the most successful of all Minoan pottery styles (Figure 1). Nearly every form of marine life is accurately reproduced in a riotous allover arrangement: octopuses, argonauts, dolphins, and fish, against a background of rocks and waves. In the 70 or 80 years after 1450 BC, the spontaneity of the early Marine style degenerated into a rigid formality. Subsequently, Late Minoan pottery became little more than a provincial version of Mycenaean ware.

Late Mycenaean. For about two and a half centuries after around 1600 BC, Mycenaean pottery painting echoed Minoan. After the eclipse of Knossos, however, Minoan influence declined, and Mycenaean potters fell back on their own resources. Minoan plant and marine motifs became simpler until virtually unrecognizable as representations of anything in real life. A figure style also developed. Adapted at first from frescoes and later from textiles, this style is seldom successful, however. Unlike the classical Greeks who came later, the Mycenaean potters were not



Figure 1: The "Octopus Vase," Marine style vessel from Palaikastron, Crete, c. 1500 BC. In the Archaeological Museum, Iráklion. Height 27.9 cm.
From *Crete and Mycenae* published by Thames & Hudson, London, and Harry N. Abrams, New York; photograph, Hirmer Fotoarchiv, München

able to adapt their fresco style so as to form a convincing figure style for vases.

Late Cypriot. The Cypriot pottery of the Late Bronze Age is of three main kinds: (1) a handmade ware with a glossy brown surface called base-ring ware, vases and statuettes of humans and animals being the most common examples of this type, (2) white-slip ware, in which handmade vases of a leathery appearance are decorated with patterns in black on a white slip (slip is liquid clay covering the pottery body), and (3) local imitations, made on the wheel, of imported Mycenaean pottery, which was evidently popular. (R.A.Hi./P.J.C.)

Ancient Greek

At the root of Greek art was the desire to explore man and the nature of his experience. Even divine subjects were cast in terms of human behaviour, and both gods and epic heroes could at times stand as representations of and models for contemporary political achievement. The seemingly naturalistic outward forms characteristic of Greek art have continued to fascinate Western artists to the present day, and the history of Western painting is full of classical revivals that have aimed at recapturing the spirit of the Greek original. Art, however, is deeply rooted in the society that creates it, and these classical revivals usually say more about those who are attempting the revival than they do about the Greek art that served as the model. Attempts to re-create the spirit and form of antique art do serve, however, as a reminder that a mere description of form does not reveal the whole truth about the art of an ancient culture. This section defines, therefore, the reasons for certain developments as well as the technical advances themselves.

A major stumbling block has been the difficulty in defining the ancient Greek attitude to art. Certainly it is clear that there was no concept of "art for art's sake" before the Hellenistic period (roughly the last three and a quarter centuries BC). Great works of art were functional: they served as gifts to the gods, monuments to the dead, or commemorations of events in the life of a city. The Greek language itself made no distinction between art and craft: both were called *technē*: a great work of art was simply an exceptional piece of workmanship (*aristourgēma*). This lack of linguistic variety should not be made too much of, however, for the actions of the artists indicate that they were exceptionally proud of their work. For the first time in the history of art, painters signed their works, and both painters and sculptors explored new means of expression. The greatest sculptors sometimes wrote books detailing their philosophy of art, and there was obviously a body of philosophical thought behind the more important advances in the painter's technique during the 5th and 4th centuries BC. By the late 5th century BC this became a basis for discussion by the philosophers themselves, indicating that, by then at least, a theory of art coexisted with the corpus of workshop techniques that might reasonably be called the practice of art. (P.J.C.)

Paintings on wall plaster, wood, and marble panels are easily eradicated, and most ancient paintings were destroyed long ago. Many fine examples, some of the highest quality, have survived, however. These are the funerary paintings on stelae (decorated stone slabs) or burial chamber walls in northern Greece and Macedonia, whose rich kings and nobles could afford the best talents from the southern cities. Contemporary vase paintings—so long as vase painting continued—often depict the same subjects and sometimes faintly reflect the style and composition of monumental frescoes, but they were in no sense accurate or even deliberate copies. The paintings on vases, now the main evidence for the development of Greek draftsmanship, were hardly mentioned by ancient writers and, although in great demand, were evidently not considered important works of art. (B.As./P.J.C.)

DARK AGES (1200–900 BC)

During the 13th century BC the great palatial centres of the Aegean world came to a violent end. Both internal dissension and foreign invasion seem to have played a part in this development, and, if the exact course of events is still obscure, the end result is quite clear: Greece was severely depopulated and impoverished. The small, scattered settlements that took the place of the great Mycenaean and Minoan kingdoms were not able to support the luxury arts that had flourished in the Bronze Age palaces. No wall paintings are known from this period, and the sophisticated Bronze Age aesthetics was lost. Before the end of the 11th century BC Greece began a steady recovery, and a secure basis was laid for all future developments. At Athens, a city that had won a position of importance in Greece only at the end of the Bronze Age, the potters invented a new painted style, which has been called the Protogeometric. Old Bronze Age shapes persisted, but they became tauter and better proportioned. In addition, the old patterns were executed with a new finesse, aided by improved equipment—a multiple brush and compasses. Using these, the painter decorated selected zones of the vase with distinctive concentric circles and semicircles, simple zigzags, and wavy lines. The vases were well potted and restrained and successful in their decoration. The simple precision of their patterns is a quality that remained dominant in Greek vase painting as well as in the other arts. Other Greek cities besides Athens adopted the Protogeometric style as well.

Protogeometric patterns

GEOMETRIC PERIOD (C. 900–700 BC)

The Geometric style arose in Athens about 900 BC. It built upon the foundations of the previous period, though the area covered by painted patterns expanded and new motifs were incorporated into the painters' repertoire. The meander, swastika, and crenellation (battlement) patterns were prominent and, together with the older concentric circles, were used by the painters to push back the large areas of solid black characteristic of Protogeometric vases and to create a pleasing halftone decorative effect. A few human and animal figures were introduced into this otherwise severely geometric scheme, but it was not until about 760 BC that a renewed interest in figures became paramount. The major achievement in this development was that of the Dipylon Master, who specialized in monumental vases used as markers over the graves of rich Athenians (Figure 2). These vases incorporated scenes with animal and human figures: funerals, battles, and processions as well as files of deer or goats. The figures were not conceived in realistic terms; rather, they were formalized into geometric shapes whose schematic appearance did the least possible damage to the overall decorative pattern. That this was deliberate is indicated by the fact that newly introduced types, such as sows and piglets, are more naturalistic at

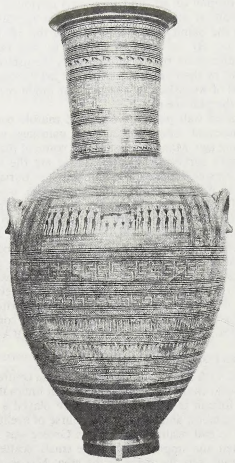


Figure 2: Attic Geometric amphora with *prothesis* (scene depicting the dead lying in state with mourners), 8th century BC. In the National Archaeological Museum, Athens. Height 1.55 m.
Alison Frantz

the time of their first appearance than in their subsequent development, when the artists learned how to cast them in a more formalized mold. Nevertheless, the introduction of schemes involving figures marked the beginning of the end for the Geometric style, for later painters became more and more fascinated with this aspect of decoration, and the older pattern work languished. By the end of the 8th century BC the figures had become much more naturalistic and were joined by floral patterns introduced from western Asia, leading to the rise of new styles in which men and gods occupied the most important positions.

The reasons for the introduction of figures, even the exact significance of such decoration, are problematic. On the simplest level, the subject matter is a factor: battles and funerals can be related to the lives of the aristocratic patrons whose graves were marked out by the Dipylon

Anderson—Alinari from Art Resource/EB Inc.



Figure 3: The "Chigi Vase" with battle, procession, and hunting scenes, Proto-Corinthian, *c.* 640 BC. In the National Museum of Villa Giulia, Rome. Height 26.2 cm.

vases. Some scholars believe, however, that the figured scenes include episodes from the heroic past or that the whole of the new iconography was cast in a heroic mold, indicating a basic identification between the aristocrats of the 8th century BC and their epic forebears. Athens in the Geometric period remained the centre from which the vase-painting studios of other cities took their inspiration.

ORIENTALIZING PERIOD (C. 700–625 BC)

Important changes in vase painting

About 700 BC important changes took place in vase painting. Floral motifs, animals, and monsters borrowed from the art of Syria and Phoenicia delivered the coup de grace to an already debased Geometric style. In Athens the new style is called Proto-Attic and includes, for the first time, scenes referring unambiguously to Greece's heroic past. The exploits of Heracles, Perseus, and other heroes were painted, often on large vases used as burial containers. The bodies of men and animals were depicted in silhouette, though their heads were drawn in outline; women were drawn completely in outline. The brushwork is bold, even sloppy on occasion, and the general effect is monumental and very impressive.

At Corinth, painting followed a different course during the 7th century BC. Corinthian painters also borrowed Oriental motifs, but their predilection for small vases, whose surfaces were divided into horizontal registers and covered with numerous tiny and beautifully drawn figures, created a miniaturist style called Proto-Corinthian (Figure 3). By the end of the century human or mythological figures were rare, and the backgrounds of the animal and narrative scenes were filled with incised floral rosettes. Corinthians introduced the black-figure technique, which, although seeming to owe something to Asian influence, is essentially native to Greece. In black-figure technique figures were painted on the naturally pale clay surface of the vase in a lustrous black pigment and then incised to indicate details of anatomy and drapery. Added colours enhanced the liveliness of these scenes. The high quality of these Proto-Corinthian vases led to a flourishing export trade, and in the later 7th century BC they were exported throughout the Mediterranean.

ARCHAIC PERIOD (C. 625–500 BC)

Corinth remained the leading exporter of Greek vases until about 550 BC, though mass production quickly led to a drop in quality. These later vases were decorated with unambitious and stereotyped groups of animal or human figures; there was little or no interest in narrative. By the late 7th century BC Athenian artists had adopted many of the stylistic features of Corinthian pots, as well as the black-figure technique. Files of animals became popular at Athens, but the artists always maintained an interest in the narrative scenes that had been so popular in the Proto-Attic style. The finest example of the marriage of Corinthian discipline and Attic invention is the François vase (in the Archaeological Museum in Florence), produced about 570 BC and exported to Etruria in Italy. Its surface is divided into horizontal friezes containing hundreds of carefully drawn, tiny figures showing episodes from Greek myth. The professionalism of the Attic masters, so clearly displayed on this and other contemporary vases, contrasted with the laziness of the Corinthian painters, and it is hardly surprising that the Attic products soon captured the foreign markets.

The first generation of Athenian painters after 500 BC concentrated on large-scale narrative scenes. One, Exekias, was fond of heroes. His superb draftsmanship and sense of the monumental was emphasized by exceedingly detailed use of incision to indicate the patterns on drapery, weapons, and anatomy (Figure 4). The Amasis Painter, on the other hand, preferred the wild cavortings of the wine god, Dionysus, and his band of drunken followers.

Many old conventions retained

In general, many old conventions were retained. Men were still painted in black on the red ground of the vase; women had white skins. But some of the work of the Amasis Painter and his contemporaries used an outline technique for women and certain other figures, and it must soon have become obvious that the brush allowed greater freedom than the graver. By about 530 BC several



Figure 4: Amphora with Ajax and Achilles playing a board game, painted by Exekias, *c.* 550–540 BC. In the Vatican Museum. Height 61 cm.
Hirmer Fotoarchiv, Munchen

painters took the momentous decision to dispense with the old black-figure technique entirely and show all their figures in outline, the details being indicated only with the brush. The background of the vase was now painted solid black and the figures stood out dramatically against this sombre field. This is called the red-figure technique, and, in the hands of artists such as Euthymides and Euphronius, the style rapidly gained ground. It had several advantages over black-figure. Incising the older decoration was painfully laborious, and it was almost impossible to vary the thickness or intensity of the incised lines. The painted line, however, could be made thicker or thinner depending on the amount of pressure applied and the amount of paint on the brush; it could also be made lighter by diluting the glaze. Red-figure artists took advantage of all these tricks and found that it was possible to depict complicated groups of overlapping figures or incidents involving violent action. Cities other than Athens and Corinth had studios producing black-figure vases; of these the most distinguished were in Sparta and eastern Greece. By the end of the Archaic period, however, only Athens was producing and exporting finely decorated pottery in any quantity.

It has always been assumed that vase painting in the Orientalizing and Archaic periods mirrored developments in monumental art, and to a certain extent this seems to be true. Not many paintings on monuments survive, but a sufficient number exist to give a general idea of their form and technique. Temple models of the late Geometric and Orientalizing periods are decorated in a way that suggests that temples had paintings on their walls; fragments of such paintings have actually been found at the temple of Poseidon at the Isthmus of Corinth. The earliest reasonably well-preserved temple decoration, however, comes from the temple of Apollo at Thermon, in central Greece, and dates from the later 7th century BC. The temple roof was decorated with a series of square terra-cotta frieze plaques, called metopes, bearing mythological scenes. Although there are several similarities to contemporary vases, there are also important differences: black-figure incision is confined to relatively minor details of drapery, and the figures themselves are drawn in outline, the women then being overpainted in white. Among vases of this period, only the brightly painted drinking cups from the island of Chios seem at all similar in technique. Other terra-cotta plaques painted in a similar, though more developed, style have been found in Italy at Caere (where they decorate the interior walls of a temple) and on the Acropolis, at Athens, indicating that there was probably a continuous tradition in this technique.

More important, because more numerous, are the many paintings on stucco. These are found in Italy and Asia Minor, as well as in Greece. They were painted by Greeks or artists working under intense Greek influence. At Pitsa, near Corinth, votive plaques covered in stucco and then painted have been found. There was a flourishing school of Greek painters who decorated tombs in the colonies of southern Italy. In Asia Minor, two tombs dating from the Late Archaic period have been found near Elmalı, in ancient Lycia (what is now southwestern Turkey). Although depicting scenes from the life of a Lycian prince, they were certainly painted by Greeks. With the exception of the plaque from Pitsa, a minor work, all these paintings come from provincial areas of the Greek world and probably do not represent the very finest of paintings then in existence, but many are highly competent pieces of work and they do give some idea of the state of monumental painting at the time. As on the vases, the greatest emphasis was on finely controlled line. Colours were applied in flat, undifferentiated masses, and there was no attempt at shading, perspective, or illusionistic treatment. At Karaburun, near Elmalı, variety was introduced by the use of finely detailed motifs on the clothing of the prince, an effect closer to the work of Exekias than to the practices of the early red-figure vase painters.

Stucco paintings more numerous

CLASSICAL PERIOD (C. 500–323 BC)

Early Classical (c. 500–450 BC). The Early Classical period is deemed to have begun after Athens' double defeat of the Persian invaders in 490 and 479 BC, but a new feeling of self-confidence was already in the air about 500 BC, possibly as a result of the firm establishment of democracy in Athens 10 years earlier. By now the Archaic colour and pattern were gone from vase painting, to be replaced by sobriety and dignity. The artist's ability to render anatomy in line had reached the point where he could accurately indicate the roundness of a figure without shading. The artist was still bound, however, to a strict profile view of heads, with few frontal, and even fewer three-quarter, views of the features. The vase painters of the first quarter of the 5th century BC included some of the finest Athens was ever to produce. One, the Cleophrades Painter, has often been called the "painter of power" since his intense, majestic subjects are rich in psychological insight. Although not all his vases concern scenes of violence, perhaps the vase that captures his spirit best is the kalpis, or wine jar, depicting the sack of Troy. It has been suggested that the extreme cruelty and tragedy present in this scene may well reflect Greek shock at the brutal sack of Miletus by Persian troops in 494 BC.

Another artist of this period was the Berlin Painter. His finest vases are almost completely covered in black glaze. Isolated or small groups of overlapping figures of extreme delicacy are posed on each side of the vases. The brushwork is exceptionally fine, and in these vases there is a sombre mood of introspection that also characterizes many contemporary sculptures. The work of this fine artist, though, is a relatively isolated phenomenon, except in funerary art where inaction and otherworldliness are appropriate. Most vase painters preferred a more narrative approach, and these narratives often reflected contemporary political developments. In 510 BC the tyranny (a tyrant at that time was a ruler, not necessarily brutal, who ruled unconstitutionally) of the Peisistratids had been overthrown in Athens, and the new democratic rulers, seeking among the heroes of the past a suitable patron, chose Theseus, an ancient king of Athens who had been credited with the union of the whole of Attica under the rule of its chief city. The new democracy fought off attempts to reinstate the tyrants, as well as defeating the two Persian invasions. It is therefore hardly surprising that the vase painters responded to the general enthusiasm and civic pride by adopting Theseus as a frequent subject. This development was reflected in monumental painting. About 460 BC the Painted Stoa at Athens was decorated with a series of paintings representing famous battles, including both legendary and historical events involving Athenians. Thus, probably for the first time in Greek history, painters placed their talents at the service of the state—moreover, a state that used them to decorate purely secular buildings. Panaenos, the brother or nephew of the sculptor Phidias, executed a picture of the Battle of Marathon for the Painted Stoa and, sometime later, included a painting of

Exceptionally fine brushwork

Greece and Salamis personified on the throne for the cult statue of Zeus at Olympia. This brought the depiction of political achievement into the very temples of the gods.

None of the Early Classical architectural paintings has survived, but a reasonable idea of what they might have looked like may be gleaned from the work of various vase painters who seem to have been working under the influence of the monumental artists. The great wall painter Polygnotus is said to have depicted figures at different depths in his compositional field, and similar compositions occur in the work of the Niobid Painter (Figure 5), although the lack of scope for such compositions on vases generally makes this something of an isolated example. Micon was another celebrated wall painter; both he and Polygnotus worked in Athens and Delphi. Ancient descriptions of their work dwell on features and moods that are easy to envisage in the light of extant contemporary vase painting and the Olympia sculptures, to which they seem to have been similar in spirit. The effect of wall paintings on white plaster may also be imagined by examining various white-ground vases intended for the tomb, where there is a concentration on calligraphic line and colour applied in flat areas without any use of shading. In other words, for all its achievements, Greek painting was still closer to drawing than anything that might today be regarded as exhibiting true painterly qualities.

By courtesy of Cliche Musees Nationaux



Figure 5: Calyx krater showing unidentified subjects with Heracles and Athena, by the Niobid Painter, c. 455–450 BC. In the Louvre, Paris. Height 54 cm.

Decline of Greek vase painting

High Classical (c. 450–400 BC). Because Greek vase painting consists essentially of the delineation of form by line, it could not follow monumental wall or panel painting once the latter began to depart significantly from their common traditions. This happened during the second half of the 5th century BC, and vase painting, while surviving for a time by looking to sculpture as a source of inspiration, went into a swift decline from about 400 BC.

There were certainly revolutionary changes in monumental painting technique. The Athenian painter Apollodorus introduced *skiagraphia* (literally "shadow painting"), or shading technique. In its simplest form this consists of hatched areas that give the illusion of both shadow and volume. A few of the white-ground vases exhibit this technique in a discreet fashion, but its true potential comes out in the great cycle of wall paintings that decorate the small royal tomb at Vergina, in Macedonia. The paintings (Plate 2), executed in the 4th century BC, represent the abduction of Persephone by Hades. The figures are defined less by an outline technique than by complicated patterns of shading and contour lines.

Another technique that also may have been included within the concept of *skiagraphia* by the ancient Greeks can be found in the treatment of Persephone's drapery: the reddish pink mantle is overlaid with slabs of darker red to create realistic patterns of light and shade, and then still darker lines are used to indicate the folds. This tomb is of the utmost importance for understanding the development of Greek painting, since it contains the earliest first-rate monumental wall painting to have survived. Therefore, it would be premature to generalize about the state of painting at that time solely from either vase painting or later Roman works, which, it has been argued, were based on Greek originals.

High point of ancient Greek painting

Late Classical (c. 400–323 BC). All authorities agree that the Late Classical period was the high point of ancient Greek painting. Within its short span many famous artists were at work, of whom Zeuxis, Apelles, and Parrhasius were the most renowned. Technique advanced considerably during this period. Zeuxis built on the discoveries of Apollodorus, and his pupil Apelles, who lived in the later 4th century BC, worked along the same lines but achieved even greater fame. They appear to have added the concepts of highlighting and subtle gradations of colour. Late Classical monuments such as the Great Tomb at Leukadia, in Macedonia, suggest that one of the means at their disposal was the juxtaposition of lines of different colours to create optical fusion—in other words, a true painterly style in the modern sense of the term. Parrhasius, in contrast, was a conservative and insisted on the priority of something called linear style, which is assumed to be closer to drawing than painting. His influence has been detected in the figure of Hermes at Leukadia and in the Lion Hunt and Dionysus mosaics at Pella, also in Macedonia.

In Athens, red-figure vase painting was in decline, and the majority of vases were painted with showy scenes, using much added colour and gilding. Occasionally there is a glimpse of brilliant line drawing, but the technique barely survived the century.

HELLENISTIC PERIOD (C. 323–1ST CENTURY BC)

The Hellenistic period began with the incorporation of the Persian Empire into the Greek world, specifically with the death of Alexander the Great (323 BC). In art history terms, however, a new relationship of painter and patron had begun slightly earlier. Apelles executed works depicting the tyrant of Sicyon and was later court painter to Alexander the Great. His career, in fact, spans the division between the two periods. The major monument for the new period is the Great Tomb at Vergina, the exact date of which should lie between the death of Philip II of Macedon, in 336 BC, and the death of his son Philip III, in 317 BC. The facade of the tomb is decorated with a large wall painting depicting a royal lion hunt. The background was left white, landscape being indicated by a single tree and the ground line. The figures themselves were painted in the fashion Apelles is assumed to have introduced, and there are sophisticated examples of optical fusion and light and shadow.

Change in patronage and subject matter

Very similar in style is the famous Alexander mosaic from Pompeii, almost certainly a copy of an original painting executed about the same time as that at Vergina. Apart from the interesting developments in technique discernible during the 4th century BC, an important change in patronage and choice of subject matter occurred. The great patrons were kings and tyrants, and many paintings exalted their claims to rule. After the 4th century BC there were few advances until the Roman period. One Demetrius of Alexandria is said to have specialized in "topographic" paintings, but the exact meaning of this word remains unclear. All other surviving Hellenistic works are of low quality. (Jo.Bo./P.J.C.)

Western Mediterranean

In the Metal Age, western Mediterranean cultures were similar at many points. The area occupied by them extended from Illyria (the northwestern part of the Balkan Peninsula) in the east to the Atlantic shores of the Iberian Peninsula in the west and from the shores along the Gulf of Lion and the Ligurian Sea (*i.e.*, the coasts of what are now southern France and northwestern Italy) and the top of the Adriatic in the north to a line stretching from Sicily to Gibraltar in the south. Of the earliest painting in classical antiquity, however, little remains except the frescoes on the tombs of the Etruscans. (R.Bl./P.J.C.)

ETRUSCAN

During the 8th and 7th centuries BC the Greeks founded many colonies in southern Italy, partly in order to ex-

pand their trade with the native peoples of Etruria, who controlled rich mineral deposits. In the Archaic period (6th century BC) these native settlements, scattered across the landscape of present-day Tuscany and Lazio in the area north of Rome, evolved into flourishing city-states whose culture was heavily dependent on influences from Greek art. More in the way of Etruscan painting has survived than in the case of Greek painting. The Etruscans buried their dead in large chamber tombs cut into bedrock; in many of these, especially in central Italy at Tarquinii (modern Tarquinia), Clusium (modern Chiusi), and Caere (modern Cerveteri), the walls of the tomb chambers were covered with plaster, and lively scenes were painted on them. Although some of these frescoes show scenes from Greek mythology, the overwhelming majority depict events in the lives of the Etruscans themselves. Funeral games were very popular subjects; perhaps the best-known depictions are those on the Tomb of the Augurs at Tarquinii, with its scenes of wrestlers, dancers, musicians, and a banquet. These paintings date from the late 6th century BC and, although the style of painting changed somewhat in later periods, the types of scene represented remained standard. The Archaic period saw the gradual evolution of an Etruscan style of wall painting whose inspiration is probably to be found in the Ionian colonies of southern Italy. By the early 5th century BC, however, the Athenian style began to predominate, and it ushered in the Classical period as well. There are many classical tombs at Clusium, including the Tomb of the Monkey.

Clusium's cultural lead

This inland city seems to have taken a cultural lead during the 5th century BC; certainly it contains competently executed works that made use of the new stylistic discoveries of mainland Greece—shading, hatching, and simple dimensional effects. There are few surviving later classical monuments in Etruria, and they seem to add little to the style established during the 5th century BC. It was only with the advent of Roman political and cultural influence during the Hellenistic period that an Etruscan renaissance in painting took place. The earliest examples of the new style are the Orcus tomb at Tarquinii and the Golinia tombs at Orvieto (south of Clusium), where there is some use of chiaroscuro effects as well as simpler means of shading. Tombs in Vulci and Tarquinii of the 1st century BC carry the development of these techniques even further. In the François Tomb at Vulci there is a celebrated fresco known as the "Sacrifice of the Trojan Prisoners." It is next to a historical scene showing wars between Etruscan and Roman princes during the Archaic period. This renewed interest in mythological or legendary equivalents of actual historical events is yet another hint that the Greek Hellenistic allegorical tradition was beginning to take hold. The same sacrificial scene and others depicting the deaths of the Theban princes during the war of the Seven Against Thebes were extremely popular on the ash urns that were used as burial jars by the late Etruscans, and it may very well have been through them that a taste for myth allegory was imparted to the Romans at this time.

ROMAN

Etruscan influence

Pagan. During the Archaic period Rome was ruled by Etruscan and Etruscanized kings. The city's temples were built and decorated in the Etruscan manner and most features of Etruscan culture were present. Although the Romans did not build painted tombs for their dead, they may have employed Etruscan artists to decorate the painted walls of the temples. When the republic was founded at the end of the 6th century BC, much of this Etruscan influence survived, especially the tendency to use painting for political purposes. Accounts of temple decoration during the 4th and 3rd centuries BC mention depictions of triumphal processions. The probable style of these is visible in the contemporary tombs of Tuscany. It was to Greek artists that the Romans turned when, in the 3rd century BC, they first came into contact with the flourishing Greek cities of southern Italy and the eastern Mediterranean. Contact was usually in the form of war, and soon Greek works of art were being brought to Rome as booty in order to decorate the temples set up as memorials to victorious campaigns. Greek artists followed the works of art as it became increasingly clear that Rome offered the best and most consistent source of patronage. In 168 BC Lucius Aemilius Paullus, the victor over the Macedonian king Perseus at the Battle of Pydna, employed Metrodorus, an Athenian painter, to execute panels depicting events in his victorious campaign. It is significant, perhaps, that Metrodorus was a philosopher as well as a painter and that he was also employed by Paullus in educating his children. Tradition states that Demetrius, an Alexandrian "place painter" (*topographos*), was working in Rome by 164 BC. The exact meaning of his title is problematic, but it could mean that he painted landscapes, later to become a favourite motif in the decoration of Roman houses. Some Alexandrian tombs of the 2nd century BC do indeed represent gardens and groves as seen through colonnades or windows in the wall of the tomb chamber. A late Hellenistic version of the type is preserved on floor mosaics from Pergamum in Asia Minor and from Italy itself. Paintings on Delos show floral motifs in an illusionistic style against a dark ground. This was later to become the Roman garden scene, usually set against a cool, dark background, that is found so often in the colonnades of Pompeian courtyards. (P.J.C.)

Wall paintings of the Roman period, for instance those from Pompeii, vary so much in their treatment of any one subject that it is hazardous to conjecture which version is likely to be closest to any earlier Greek painting, even supposing there was definite copying. With the exception of the Alexander mosaic already mentioned (see above *Hellenistic period*)—evidently a direct copy of a painting of the 4th century BC—there is nothing in painting to correspond to the straightforward copying of Greek statues that was apparently so abundant under the Roman Empire. (B.As./P.J.C.)

Virtually the only example of painting in Rome and Latium to have survived from before the 1st century BC is a fragment of a historical tomb painting with scenes from the Samnite Wars, found in a family tomb on the Esquiline and probably dating from the 3rd century BC (Palazzo dei Conservatori, Capitoline Museums, Rome). In addition to Metrodorus and Demetrius, ancient writers mention the names of three painters, each of whom worked in a temple: Fabius Pictor, in the Temple of Salus in Rome at the end of the 4th century; Pacuvius, a dramatist and native of Brundisium, in the Temple of Hercules in the Forum Boarium in Rome during the first half of the 2nd century; and Lycon, an Asiatic Greek, in the Temple of Juno at Ardea in the late 3rd or early 2nd century. Nothing is known about the work of these artists.

At Pompeii during the 2nd century BC the interior walls of private houses were decorated in a so-called Incrustation, or First, style; that is, the imitation in painted stucco of veneers, or *crustae* ("slabs"), of coloured marbles. But in the second half of the 1st century BC, there suddenly appeared in Rome and in the Campanian cities (the most famous of which is probably Pompeii) a brilliant series of domestic mural paintings of the so-called Second style, the aim of which was to deny the walls as solid surfaces confining the room space. This was sometimes done by covering the whole area of the walls with elaborate landscapes, in which depth, atmosphere, and light are rendered in a highly pictorial, illusionistic manner. Such are the *Odyssey* paintings (Plate 1) found in a Roman house on the Esquiline (now in the Biblioteca Apostolica Vaticana, Vatican City), which consist of a continuous flow of episodes that unfold, filmlike, beyond a colonnade of pilasters, with vertical, bird's-eye-view perspective and human figures strictly subordinated to their settings. Other wall paintings, such as those in a room from Livia's Villa at Prima Porta (transferred to the Museo Nazionale Romano), represent a great park or garden filled with trees, shrubs, flowers, and birds, with no pilasters in the foreground to interrupt the prospect and no human figures to distract attention.

The possibility of Hellenistic models for this type of painting has already been mentioned, though the surviving Hellenistic precursors were no preparation for the important Roman developments. Most examples of the type, which survived into the Fourth style, have been found on the back walls of the colonnades running around

real gardens. The cool painted scenes would have given the illusion that an idler in this part of the house was surrounded by shrubs or groves of trees. Another type of landscape combined sacred and idyllic features and was often placed as though behind elaborate stage buildings. These monotone compositions held sacred columns or rustic shrines and were closely related to other illusionistic scenes peopled with little figures whose antics, the written sources make clear, were a source of endless amusement for the householder and his guests.

Frieze in the Villa of the Mysteries

A celebrated frieze of life-size figures, depicting Dionysiac initiation rites and the prenuptial ordeals of a bride, in the so-called triclinium of the Villa of the Mysteries (or Villa Item) outside the Herculaneum gate of Pompeii, also belongs to the Second style (Plate 2). There the walls are denied by the device of substituting for them a narrow stage on which the figures carry out the ritual before a drop scene of continuous painted panels. But the most common Second-style paintings are known as Architectural and show a threefold horizontal division of the wall into dado, central area, and cornice, combined with a triple vertical scheme of design that consists of a large central panel (in the main, intermediate horizontal area), framed by flanking columns and a pediment, and two smaller panels on either side. The central panel and often the lateral panels as well are views seen through windows that break through the walls and link the spectator with the world outside, as in the house of Augustus on the Palatine in Rome.

In the Third style, which covers most of the Augustan period, the central panel picture on a wall is no longer thought of as a scene through a window but as a real picture hung on or inserted into a screen or woven into a tapestry, which partially conceals an architectural vista behind it. The columns, entablatures, and so on are completely unreal and so complicated that this Third style is sometimes dubbed Ornate.

The Fourth style, which runs from the close of the Augustan Age to the destruction of Pompeii and its fellow Campanian cities in AD 79, is less homogeneous than its predecessors and exhibits three main variants: first, an architectural design soberer and more realistic but still with a central screen or tapestry partly covering a retreating vista; second, an architectural layout that imitates a *scaena* ("stage background"); and third, a method (sometimes known as "intricate") by which the whole surface of the wall is covered with a flat, white, neutral ground painted with an allover, latticelike pattern of fantastic architectural elements, arabesques, grotesques, small figure motifs, or small panels containing pictures. This third type of Fourth-style painting came into vogue at Pompeii between an earthquake of AD 63 and the catastrophic volcanic eruption of 79, and one of its most impressive exponents is the Golden House of Nero in Rome.

The subjects of the panel pictures of the Second, Third, and Fourth styles are for the most part drawn from Greek mythology. Some of them recall literary descriptions of famous classical Greek and Hellenistic paintings or show motifs that suggest their originals were painted on the Greek mainland or in Asia Minor. It is certain that many masterpieces of Greek painting did make their way to Rome as the booty of Roman generals of republican days, and wall painters could have studied them at first hand. But often those artists must have had to rely only on sketches of the celebrated pictures, and it is not known how faithfully the Roman and Campanian murals reproduce the prototypes. Other panel pictures present scenes from contemporary religious ritual, and a few show themes from Roman legend. Frequently, in the case of the Greek mythological subjects and those taken from rustic religious cults, the artist produced landscape with figures, as in the *Odyssey* frescoes (Plate 1), not figures with landscape, as on Trajan's Column. These late republican and early imperial set pieces are competently executed, remarkably vivid, and extremely naturalistic. But, with a few exceptions, they reveal that the principles of a single vanishing point and unified lighting from a single source of illumination either were not understood by Roman painters or did not interest them.

Late republican and early imperial set pieces

The flat, uniform background of the last phase of the Fourth style remained a constant feature of mural painting in houses, tombs, temples, and other religious shrines throughout the 2nd, 3rd, and 4th centuries. The decoration, which stands out against that ground, takes any of several forms: (1) latticelike, allover patterns, as in many pagan tombs; (2) small groups of figures or figure panels spread out at intervals across the field, as in the Christian catacombs of Rome; (3) a mixture of large human figures and extensive scenes with small-scale figures, as in the early 3rd-century Hypogeum of the Aurelii on the Viale Manzoni in Rome, the interesting painted content of which is Gnostic or crypto-Christian; or (4) large scenes with relatively large figures, such as a group of marine deities in a 2nd-century Roman house under the Church of SS. John and Paul on the Caelian, a late 2nd- or early 3rd-century leopard hunt on the south wall of the frigidarium of the hunting baths at Leptis Magna (on the coast of modern Libya), or the early 3rd-century biblical scenes from a baptistery at Doura-Europus, an important archaeological site on the Euphrates in what is now Syria.

In the case of the Roman tombs, cross- or barrel-vaulted ceilings, where preserved, normally carry out the painted decoration of the walls, showing either a latticelike pattern or a series of small, spaced-out, figured panel pictures. At Trier (in what is now West Germany) remains have been found of a flat, coffered ceiling with panels of painted plaster from an early 4th-century imperial hall destroyed to make room for a Christian basilica. Large portions of eight painted panels are preserved. Four depict female busts—three of them with nimbi—which may be either personifications or portraits of members of the imperial family; the other four show pairs of dancing or sporting cupids. As the skillful modeling and lively naturalism of these figures show, late Roman painting could reach high standards.

Roman portrait painting comes only a short way behind portrait sculpture in technical skill and realism. One of the earliest extant examples is a group of Terentius Neo and his wife, from Pompeii (National Archaeological Museum, Naples). Both figures recall mummy portraits in Egypt, being painted in encaustic (a technique by which colours are mixed with liquid wax and fixed by heat) and ranging in date from the Flavian period to the 3rd century. A circular portrait group of frontal figures painted on wood, probably in Egypt (now in the Staatliche Museen Preussischer Kulturbesitz, West Berlin), seems to have originally depicted the emperors Septimius Severus and Caracalla, Julia Domna (Septimius Severus' wife), and Geta (Caracalla's brother); but Geta (so it seems) was subsequently washed out (perhaps most consequent upon his murder by Caracalla). Particularly attractive are the portraits done on gold-glass medallions, which in the exquisite refinement of their treatment may be compared to 16th-century European miniatures. A medallion in the Museum of Christian Antiquities, Brescia, dating from the 3rd century and carrying a portrait group, is a veritable masterpiece (Plate 2).

Gold-glass medallions

(J.M.C.T./P.J.C.)

Early Christian. It is customary to distinguish early Christian painting of the West or Latin part of the late Roman Empire from the Christian painting of regions dominated by the Greek language and to consider the latter as proto-Byzantine. The Western strain of early Christian painting may be said to have ended with the collapse of the empire in the West at the end of the 5th century. In the East, until the 6th and even the 7th century, painting in many regions followed the paths traced by Christian painting at its beginnings. Exceptions to the above schematization are Doura-Europus and early Christian paintings in Egypt (see below).

Early Christian painting did not have a distinct existence until about the end of the 2nd century AD. There are several reasons for this. First, there can have been few, if any, monumental churches before that time capable of taking decoration showing Christian themes. Second, Christianity did not at first make great headway among those able to afford large painted tombs where examples of Christian iconography might be expected to appear. Third, early Christianity was much closer to Judaism than

in later years and may have retained the Judaic distaste for the painted image, especially if it referred to the Godhead. Lastly, even Christians prized classical education, which was, after all, the only sound basis for a public career, and they could appreciate classical works of art even if they rejected pagan subject matter. This was made easier for them by the concept of myth as allegory, according to which depictions of mythological scenes were not so much statements of a religious position as moral lessons whose messages could be appreciated by any educated man. Because most surviving early Christian painting is funerary, it is hardly surprising that purely Christian subjects at first made little headway in a field already crowded with edifying moral messages based on the Greek myths. These may have been pagan, but they did emphasize the common belief in life beyond the grave. It was only in the 3rd century AD, when the idea of Judaic or Christian allegories gained legitimacy, that any real development could begin. Even so, some rather odd compromises took place: representations of Christ as the victorious Sun God or as a philosopher occur in early Christian tomb paintings. Even the emperor Constantine, who by the Edict of Milan (AD 313) established toleration for Christianity throughout the Roman Empire and who himself professed Christianity, seems to have worshiped both in his lifetime.

Myth as allegory

The new elements, then, consisted not of form but of content. As the power of the church over public and private life grew, these new elements tended to gain in importance, but they never quite ousted the pagan scenes. The latter were often drawn undiminished from plays or epics whose prestige remained long after Christianity became the official religion of the Roman Empire.

With the growth of Christian communities, the catacombs—underground burial places—became veritable subterranean cities, their rooms linked by corridors. The most important extant examples are in Rome, with others in Naples, Sicily, Malta, North Africa (specifically in what is now Tunisia), and Egypt. Pictorial decoration of the catacombs, limited to only a few rooms, followed pagan models. Delicate lines on the ceilings and walls trace circles and squares in which decorative motifs are inserted: garlands, birds, four-legged animals, cupids, images of the seasons, and figures of ambiguous significance (pagan or Christian)—praying figures and a shepherd carrying a sheep on his shoulders, generally called "The Good Shepherd" (Figure 6).

Catacomb paintings and the beginnings of Christian iconography

As early as the first half of the 3rd century, however, scenes of purely Christian meaning were added to these neutral subjects. The oldest are located in Rome in the cemeteries of Domitilla (the gallery of the Flavians), of Priscilla (the Greek Chapel), and of Calixtus (the Chapel of the Sacrament). Stories from the Old Testament are joined by stories from the Gospels. These images present examples of the succour brought to the faithful by God the Father and Christ the Son. Even the baptism and the adoration of the Magi can be interpreted in this manner; as revelations of Christ's divinity, they announce man's salvation.

The style and quality of these paintings vary. Some of those from the beginning of the 3rd century are light in touch and charmingly elegant (*e.g.,* gallery of the Flavians), comparable to the best pagan paintings. In others (*e.g.,* cemetery of Priscilla, mid-3rd century) there is a somewhat heavier element, with a passion of expression that seems to match the aspirations of the new faith. In the 4th century the style becomes firmly contoured.

The frescoes in the baptistery of Doura-Europus, executed between 230 and 240, differ only in style from those of the catacombs in the West. Scenes from the Old and New Testaments are used to explain the significance of baptism: the death of the old Adam and his rebirth to a new life through the baptismal bath. The back wall of the baptismal pool bears the images of Adam and Eve, recalling the Fall, as well as that of the Good Shepherd (who in this case is Christ, Saviour of souls). Illustrations on the longitudinal walls are of David fighting Goliath and of incidents from the Gospels—Christ walking on water, the healing of the paralytic, the holy women at the tomb, and Christ and the Samaritan—and were probably inspired by readings that accompanied the rite of baptism. Stylistic elements that recall the paintings in the Roman catacombs are the juxtaposition of scenes without apparent connection and the conciseness of the narrative.

Among the latest examples of early Christian funerary art are paintings dating from the 5th century in the tomb of el-Bagaouat at al-Khārijah, in Egypt.

(He.S./P.J.C./Ed.)

Book illustration in antiquity. That book illustration existed as far back as the late Hellenistic world can be inferred from some of the so-called Megarian bowls, imitations in clay of gold or silver vessels that date from the 3rd century BC to the 1st century AD. They often bear on their exteriors scenes in relief from literary texts that are sometimes accompanied by Greek quotations. They must, in part at least, have served as models for Roman artists. Book illustration is known to have existed in Rome comparatively early—examples include 700 pictures illustrating the early 1st-century-BC scholar and satirist Marcus Terentius Varro's 15 books of *Hebdomades vel de imaginibus* and a portrait of Virgil prefixed to an edition of

By courtesy of Pontifica Commissione di Archaeologia Sacra, Rome

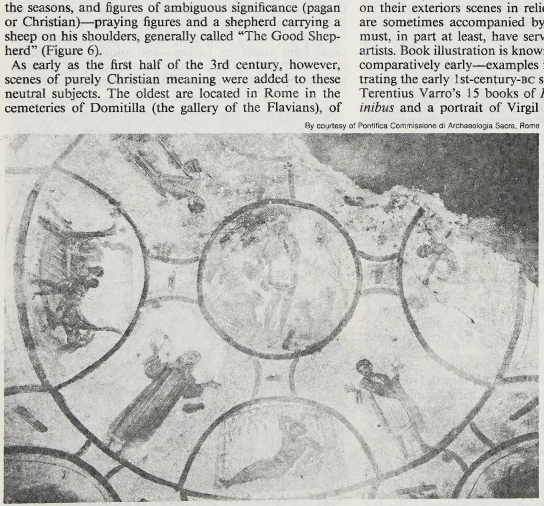


Figure 6: "The Good Shepherd," painted ceiling in the catacomb of SS. Pietro e Marcellino, Rome, beginning of 4th century AD.

his poems. Miniatures in the codex of the *Iliad* in the Biblioteca Ambrosiana, Milan, were painted probably at the end of the 5th or beginning of the 6th century AD but reflect pictures of the 3rd, 2nd, and even 1st centuries AD, as do those of the Codex Virgilius Vaticanus in the Biblioteca Apostolica Vaticana (No. 3225), written about 400. Miniatures in the second great illustrated Codex Virgilius Romanus in the Biblioteca Apostolica Vaticana (No. 3867), written about 500, are still Roman in spirit, if less classical in style.

The tenacious influence of Greco-Roman painting can be traced clearly in the illustrations of certain early Byzantine books. A most remarkable, if aesthetically crude, mid-4th-century mosaic pavement, found in a Romano-British villa at Low Ham, Somerset, and showing scenes from the first and fourth books of the *Aeneid,* is undoubtedly based on the copybook illustrations used for some Virgilian codex. (J.M.C.T.)

The only Christian illuminated manuscript surviving from before the 6th century is a fragment of the Book of Judges from Quedlinburg (Staatsbibliotek Preussischer Kulturbesitz, Berlin). (He.S./Ed.)

Eastern Christian

A new artistic centre was created in the eastern Mediterranean with the foundation in the early 4th century AD of Constantinople (modern Istanbul) on the site of Byzantium. The term Byzantine is normally used to identify the art of this city and of the Orthodox Christian empire that was controlled from it and that survived from 330 until its capture in 1453 by the Ottoman Turks. From the reign of Justinian I (527–565) there were relatively clear political and ecclesiastical differences between the Byzantine world and the West, and the term Byzantine art from this period onward broadly reflects these differences. In practice, the division of Mediterranean art into two polarities is not always easy to maintain, as artistic contacts were frequent and each "sector" influenced the other. For instance, by the 12th century, Byzantine influence had made itself felt outside the empire, as, for example, in the mosaics of Sicily and Venice; and the Byzantine style had been adopted by the Orthodox states that were growing up in Russia, Bulgaria, and the western Balkans. During the first half of the 13th century, when Constantinople was in Latin crusader hands, it was in these outlying areas that the most important developments in painting took place. Once the Greek emperors had returned to Constantinople in 1261, developments there began anew, and fine Byzantine mosaics and paintings date from this last phase. In the study of Byzantine art, mosaics are frequently included with painting, but here painting is treated alone; for mosaics, see DECORATIVE ARTS AND FURNISHINGS.

Byzantine influence in the West

By 1460 or a short while after, the little that remained of the empire following Constantinople's fall in 1453, together with the independent Orthodox states (except Russia), was in Turkish hands. Nevertheless, painting in the Byzantine tradition continued in Greece, the western Balkans, and Bulgaria, for Orthodox Christian art was not banned by the new Muslim rulers. Indeed, works of great technical sophistication were still produced, and a number of painters of icons and church paintings are known through signed and dated works. In Russia a national art of great quality saw continuous development from a Byzantine basis throughout the Middle Ages and up to the end of the 17th century, when Peter I the Great imposed western fashions.

In general, Byzantine painters may perhaps have retained Greco-Roman traditions more faithfully than did medieval painters in the West. There is so much variation of expression in the history of Byzantine painting, however, that it would be misleading to describe it as a "style"; the term is better seen as the label for a period and for the patronage of an Orthodox Christian society. Because most surviving work is religious in content, Byzantine painting does have some distinctive features. Icons, or painted panels depicting holy figures, were a major item of production, and the most important churches have their walls decorated in mosaic. On the other hand, the production of illuminated books was limited. The range of subject matter in Byzantine works is more restricted than that of the medieval West; scenes and figures from the New Testament and the history of the early church are perhaps the most popular choices.

Byzantine painting was a highly effective Christian art, expressing a new view of the divine and a new spirituality. On the whole, Byzantine emphasis concentrated less on presenting a naturalistic narrative than on suggesting the existence of a supernatural and timeless Christian realm; painters retained the pictorial devices of classical antiquity, even if they aimed at portraying a more abstract version of the world. It has been felt that Byzantine art as a result always contains a tension between naturalistic and abstract modes of expression.

EARLY BYZANTINE PERIOD (330–717)

Icons. Until quite recently very little was known about the icons of this age, but, owing to the cleaning of several in Rome and the discovery of hundreds in the monastery of St. Catherine on Mount Sinai, much material is now available for study. Enough examples are now known to substantiate the significance of the icons.

Early importance of icons

The icons in Rome (large cult images such as that of the Virgin from the Pantheon) represent Christian images at their most formal and monumental. The Sinai icons are more intimate, and many must have been intended for private devotions as well as church display. Among the finest are icons that represent Christ, St. Peter, and the Virgin and saints.

Illuminated manuscripts. It is now thought that illuminated manuscripts were relatively few in number even at the time they were produced. Certainly very few religious or classical texts survive. Of the latter, a copy of the pharmacological treatise *De materia medica* by Pedanius Dioscorides, a Greek physician of the 1st century AD, is certainly Constantinopolitan; it was done for Juliana Anicia, the founder of the church of St. Polyeuktos, and is dated 512. A copy of the *Iliad* at Milan may perhaps have been copied and illustrated in a Byzantine scriptorium. Of the religious manuscripts, the most important is a copy of the book of Genesis (known as the Vienna Genesis) at the Österreichische Nationalbibliothek, Vienna (Plate 3); there is a fragmentary copy of the Gospels in the Bibliothèque Nationale—usually known as the Sinop fragment, for it came from Sinop, in Turkey—and another at Rossano, in southern Italy. There is also another copy of Genesis, the Cotton Genesis, in the British Museum, but it was

By courtesy of the Biblioteca Medicea-Laurenziana, Florence



Figure 7: Miniature of the Ascension from the Rabula Gospels, copied at Zagba, Syria, 586. In the Biblioteca Medicea-Laurenziana, Florence. 35.5 × 25.5 cm.

severely damaged by the fire that destroyed part of the Cotton Collection. There has been dispute as to where these manuscripts were written and painted, but either Constantinople or Syria is the normal attribution. A fifth religious manuscript, the Rabula Gospels, whose text is framed in elaborate architectural and floral motifs (Figure 7), was copied at Zagba, in Syria, in the year 586 and was executed in a more sketchy, informal style.

ICONOCLASTIC AGE (717–843)

By the early 8th century so great an importance had accrued to the depiction of the saintly and divine forms that one body of opinion in the state feared the population was in danger of lapsing into idolatry. As a countermeasure, a decree forbidding representation of saintly or divine forms in religious art was promulgated, and from about 717 until 843 there reigned emperors who are called Iconoclasts. To most of them, representation of the saintly or divine in religious art was genuinely anathema. In spite of the ban, pictorial decoration was not in itself forbidden. The church of Ayía Sophia (literally "Divine Wisdom") at Salonika (modern Thessaloníki, Greece) was decorated under the patronage of Constantine VI (780–797); his monogram survives, and in the apse there are indications that there was a great cross like that which is preserved in the Church of St. Irene (Eirene) at Constantinople and which dates from the 740s. The survival of the 6th- and 7th-century figural mosaics in St. Demetrius at Salonika suggests that the ban was not strictly enforced everywhere. In any case, it was strongly opposed in the monasteries. But in Constantinople the ban seems to have been universal, and religious mosaics and paintings in all the churches were removed, including all those in Hagia Sophia.

Figural church decoration

MIDDLE BYZANTINE PERIOD (843–1204)

"Icon lovers"

Icons. With the return to power of the "icon lovers," as they were called, in 843, figural art once more became important in the churches. Elaborate representational decorations in mosaic were set up in the more important buildings, painted ones in the poorer. The next two or three centuries were an age of great brilliance and represent the acme of Byzantine culture. The empire's frontiers were far-flung, its wealth was enormous, and its general culture was far in advance of the rest of Europe. After the death of Basil II (976–1025), a slow decline set in.

Icons were regularly produced throughout this period. The largest number are to be found in the Sinai monastery. These were mostly for Orthodox use but include a 12th- and 13th-century group done in a mixed East-West style by Western painters who were active in the Latin crusader kingdoms of the region and who copied Byzantine models. Others exist in various museums in the Soviet Union, where they were brought from provincial Russian churches and monasteries for cleaning and conservation. Some of these were imported from Constantinople; one of the finest, an icon of the Virgin known as "Our Lady of Vladimir" (Plate 3), was painted for a Russian patron about 1130. It is of considerable importance in the history of painting, for it not only is a work of outstandingly high quality but also is in a new, more human style, anticipating the late style that flourished between 1204 and 1453. It was at this time that the cult of the icon really came into its own, partly because richer materials became rare but mostly because the interior decoration of churches changed with the introduction of a screen called an iconostasis (*i.e.,* a screen that was to be covered in icons).

Wall painting. Wall paintings were important during this period, but only one decoration by trained artists in a larger building is known, namely that in the Church of St. Sophia at Ohrid, Macedonia (Yugoslavia). The majority of the scenes that survive were drawn from the Old Testament. They date from about 1050. More numerous are the paintings that decorate numerous rock-cut chapels in Cappadocia (in what is now Turkey); these were executed by lay painters for the monks who lived there alone or in small communities. This material is most important for understanding the character and varieties of Byzantine painting and for giving records of the near-complete decoration of churches. Some churches (such as the 10th-century Tokalı kilise in the Göreme Valley, in central Turkey) represent the best achievements of the period. Some artists who painted the Cappadocian churches must have traveled out from Constantinople or other cities; others probably made their living locally. All levels of quality were found there; indeed Cappadocia contains a whole range of the subjects depicted in Byzantine painting of this period.

Illuminated manuscripts. Two magnificent manuscripts of this period survive: the Paris Psalter (Figure 8) and a book of sermons (*Homilies of St. Gregory of Nazianzus*), both in the Bibliothèque Nationale, Paris. The former contains 14 full-page miniatures in a grand, almost classical style, which led scholars at one time to date it to the earliest Byzantine period. The miniatures in the other book are more varied in style, some of them recalling the narrative art of Cappadocia, but this latter book represents nevertheless the grandest type of Byzantine manuscript of the age. It was done for Basil I about 880. During the following centuries many illuminated psalters, octateuchs (the first eight books of the Old Testament), homilies, and copies of the Gospels were produced. (Gospels formed the most numerous category.) Notable examples include the Bible of Leo and the Menologion (a liturgical book

Grandest type of Byzantine manuscript of the age

Hirmer Fotoarchiv, München



Figure 8: David composing psalms, miniature in the Paris Psalter (Gr. 139), 10th century. In the Bibliothèque Nationale, Paris. 26.0 × 35.9 cm.

relating lives of saints) of Basil II (976–1025), both in the Vatican, a psalter done for the same emperor and now in the Biblioteca Nazionale Marciana at Venice, and *The Homilies of St. John Chrysostom* in the Bibliothèque Nationale. A few of them contain many small-scale illustrations, as in a famous set of the Gospels in the Biblioteca Medicea-Laurenziana at Florence. The most common type of Gospel book had only a few illustrated scenes or only portraits of the Evangelists. The work is usually of high quality. Some psalters contained marginal illustrations referring to contemporary events (*i.e.,* the Iconoclastic Controversy). The 10th-century Joshua Roll (Figure 9) is interesting as an example of Byzantine illuminated manuscript that shows the tenacious influence of Greco-Roman painting.

In 1204 Constantinople was sacked by crusaders, its treasures were destroyed or dispersed, and the brilliant middle period of Byzantine art was brought to an end.

LATE BYZANTINE PERIOD (1204–1453)

Icons. Painted panels assumed a new importance in the last phase of Byzantine art. The most sophisticated work was done at Constantinople, some of it for patrons from elsewhere (notably Russia), and a number of icons survive



Figure 9: The attack on Ai, from the Joshua Roll (Josh. 7:3–5), Byzantine, 10th century. In the Biblioteca Apostolica Vaticana.
By courtesy of the Biblioteca Apostolica Vaticana

that can be associated with Constantinople on the basis of literary evidence or inscriptions. A particularly fine double-sided icon, with the Virgin on one face and the Annunciation on the other, now in the museum at Skopje in Macedonia, was brought from Constantinople about 1300.

At this period the Russian school was the most important outgrowth of Byzantine icon painting; after the 13th century the influence of Byzantine models continued to be felt more in Russian icons than in the frescoes, but both wall and icon painting were showing local characteristics as early as the 13th century itself. The rigid Byzantine patterns, the dark colours, and the austere lines gradually became graceful, bright, and less solemn. Novgorod's style of icon painting, for example, gradually strengthened and took shape: the severity of faces was softened, composition was simplified, the silhouette became bold and increasingly important, and the palette was lightened by bright cinnabar, snow-white, emerald-green, and lemon-yellow tones.

Local styles of Byzantine models

Icons were produced in many other places, notably at Salonika, on Mount Athos, and in many other centres in what are now the Balkan states and areas such as Russia and Ukraine. In a few instances icons can be assigned to a definite centre, thanks to inscriptions or other records, but the study of these panels has not progressed far enough to permit any reliable classification under localities on the basis of style alone. After the Turkish conquests of the mid-15th century, icons continued to be painted in large numbers in every part of the Orthodox world. In the 16th century Crete became an important centre, and many Cretan painters worked also in Venice, where there was a large Greek colony; many of the products of this school are to be found there today in the museum attached to the Church of St. George of the Greeks.

Wall painting. The last phase really began in the 12th century with the decoration at Nerezi in Macedonia (1164). It was done for a Byzantine patron and is in the same emotional style as "Our Lady of Vladimir." Work in a similar style is to be found in Russia from the late 12th century, and these models were followed by local craftsmen. In the 13th century new styles predominated in such paintings as those at Mileševa (1235) and the Church of the Trinity at Sopočani (*c.* 1265), in Serbia, and in the church of Hagia Sophia at Trebizond (*c.* 1260; Trabzon), on the Black Sea.

It is probable that artists who had fled the capital after 1204 established themselves in a number of different areas and that wall paintings such as those mentioned above were the work of men they had trained. By the end of the century, the local art in the Byzantine Empire emerged as the regional art of Salonika. Examples of this last school are found in the Chapel of St. Eugenius, attached to the Church of St. Demetrius at Salonika, in the Protaton (*i.e.*, the First Church, in the sense of the first in rank, *c.* 1300), at Kariaí (Karyaes) on Mount Athos, on the north coast of the Aegean, and in some of the monasteries there, as well as in a number of churches in Serbia and Macedonia decorated under the patronage of King Stephen Uroš II Milutin at the end of the 13th century and in the early years of the 14th century. There has been some dispute among authorities as to whether King Milutin's painters were Greeks from Salonika or local Slavs. Throughout the 14th century a great deal of work was done by painters in the Balkan region, notably in Greece and Bulgaria.

In Russia the Mongol invasion about the middle of the 13th century disrupted previous centres of production, such as Kiev and Vladimir-Suzdal. Only in the northern regions of Russiâ—particularly in the Novgorod district—did painting continue to develop. As early as the second half of the 12th century, the city of Novgorod had developed an individual style, combining Byzantine severity with a folk-art picturesqueness. (Examples are the frescoes in the Church of St. George in Staraya Ladoga [*c.* 1180] and the Church of Nereditsa.) Novgorod escaped damage by the Asiatic hordes and became virtually the metropolis and cultural centre of old Rus after the fall of Kiev (1240). Together with the city of Pskov and other northwestern Russian population centres, it harboured many Greek artists, who continued to work in the traditions of Byzantium.

Continuity in northern regions

A prominent figure in Russian painting was Theophanes the Greek, a native of Constantinople who moved to Russia after about 1370. His paintings, though closely adhering to Byzantine styles, show distinctive Russian features, notably elongated proportions and delicacy of detail. Similar characteristics and features can be seen in his Novgorod frescoes and especially in the central part of the iconostasis in the Cathedral of the Annunciation in the Moscow Kremlin.

Among the immediate followers and collaborators of Theophanes was Andrey Rublyov, whose religious types are imbued with a fresh spirituality. His best-known work is the icon "The Old Testament Trinity" (*c.* 1410), painted for the Trinity-St. Sergius Monastery at Sergiyev Posad. The subject—popular in Byzantine iconography—is the visit of three angels to Abraham and Sarah. But the severe symbolism of the old Byzantine tradition is transformed into something more human. It is one of the great creations of medieval Russian painting.

Another inspired Novgorod painter of the 15th century was Dionisi, whose art is marked by the extreme elongated stylizing of his figures as well as a subtle and glowing colour scheme. He and his predecessor Rublyov succeeded in expressing the aura of spirituality that is the essence of the Russian icon.

Elongated stylizing

At Constantinople some paintings of outstanding quality were executed at the Monastery of the Chora, now

known as Kariye Cami (Plate 3), and it is known from the texts that similar paintings existed in a number of other churches there. Several were painted in the third quarter of the 14th century by Theophanes the Greek before he went to Russia. The same style was also introduced to Mistrás, in the Peloponnese, and there the wall paintings of the Brontocheion (early 14th century), the Church of the Peribleptos (*c.* 1350), and the Pantanassa (1428) are all of high quality. Paintings in the monasteries of the Morava Valley in Serbia done at the end of the 14th century and beginning of the 15th are in the same refined style.

Illuminated manuscripts. Illuminated manuscripts of the last Byzantine age are not as numerous as those of the middle period, but their quality is often just as high. A few seem to have been produced during the 13th century, both at Constantinople and in the cities where Orthodox nobles established themselves while the Latin crusaders were in possession of the capital, notably Nicaea and Trebizond. After the return to Constantinople in 1261 the noble families seem to have played a greater role than the emperors as patrons of all arts, and many of the more important works of art of the age were produced on their behalf. A copy of a work attributed to the 5th-century-BC Greek physician Hippocrates, now in the Bibliothèque Nationale, was made for the high admiral Alexius Apocaucos (Plate 3), and a beautiful copy of the Gospels in the same library was made for the emperor John VI Cantacuzenus between 1347 and 1355. Manuscripts were, of course, also copied and illuminated in the monasteries, and this process continued until printing made it obsolete. Few of the later ones contain illuminations of great quality. In the Slavic lands, however, fine work continued, and in Romania excellent manuscripts were executed in the 16th century. (A.Vo./D.T.R./R.S.Co.)

Patronage of the nobility

POST-BYZANTINE RUSSIA

In the 15th century, major changes began to take place in Russian icon painting, leading to the birth of what may justifiably be called a national art. This evolution first became noticeable in the gradual elimination of the Mediterranean setting depicted in the background of icons, notably landscape and architecture. Greek basilicas with their porticoes and atria (patios or courts) were replaced by Russian churches with their cupolas and *kokoshniki* (literally "women's headdresses" but here, by extension, "gables"). Russian saints and episodes from their lives furnished subjects for the Russian artists; Muscovite types and native costumes began to appear in icon painting. The colours were extraordinarily brilliant, and there was particular emphasis on outline.

Many of the great icon and fresco painters in the 16th century worked first at Novgorod and later at Moscow, thus linking Novgorod and Moscow closely in artistic terms and in particular introducing to Moscow features characteristic of the Byzantine and Novgorodian traditions. The literary movement of the 16th century strongly influenced contemporary painting, and artists looked to new subjects. Some illustrated church preoccupations and prayers or expressed the rites of the church in symbolic images; others represented parables and legends.

At the end of the 16th century the Stroganov school made its appearance in Moscow, introducing a small-scale manner of icon painting. The masters of the Stroganov school became famous for the elegant attitudes of their figures, their Eastern choice of colours, and their elaborate detail. Some of them—Prokopy Chirin, Nikifor, and Istoma Savin—were later to join the ranks of the icon-painting studios in the Kremlin armory in Moscow.

Rapid spread of Western influences

Moscow icons of the 17th century constitute the last authentically Russian painting. As early as 1650 much of their Russian character had disappeared. From the end of the century, western European influences spread rapidly. (A.Vo./R.S.Co.)

REGIONAL VARIATIONS IN EASTERN CHRISTIAN PAINTING

Georgia. Christian painting in Georgia dates from the 4th century and shows both Eastern and Western influences, owing to the position of the region as a crossroads of trade between Europe and India. From the beginning of the 5th century the Georgian church approved the representation of the human form in religious painting. Accordingly Georgia was not affected by the wave of iconoclasm in the 8th and 9th centuries—a period that inhibited figural representation in most of Eastern Christendom for more than a century. In addition to a Christian tradition, Georgian painting also drew on a pagan one.

Until the 9th century, mosaics—more or less Byzantine in technique and design—were frequently used in the decoration of Georgian churches. By the 11th century the entire interior of Georgian churches was usually covered with frescoes instead. Many well-preserved examples survive from this period. Although following the Eastern Orthodoxy's general theological interest in church decoration, the Georgian murals deviated somewhat from Byzantine style and iconography, notably in extensive ornamentation between individual scenes.

The art of manuscript illumination flourished in Georgia from the 6th century onward, and numerous examples survive from all periods. Characteristic of the early works are two Gospel books, the Adishi Gospels (897) and the first set of Gospels of Dzhruchi (936–940). These are distinguished by their decorative treatment of draperies and their excellent drawing.

Byzantine influence strengthens

At the end of the 10th century Byzantine influence became strong in Georgia, and until the end of the 15th century Georgian manuscripts generally followed Byzantine models, differing only in an independent approach to the use of colour. These illuminations are of very high quality.

In the 16th century Persian influence from the East transformed Georgian manuscript illuminations. Ornamentation abounded, and the representation of figures and scenes was flat, decorative, and highly skillful. (Ed.)

Armenia. What little remains of the pagan art of Armenia strongly resembles late Greco-Roman art. With the establishment of Christianity as the official religion in the first years of the 4th century, however, a truly national art developed.

From an early period the interiors of Armenian churches were adorned with frescoes and mosaics showing scenes from the Gospels and images of Christ, the Virgin, and saints. Surviving examples are less plentiful than illustrated manuscripts, however. Important specimens of the latter exist in an almost uninterrupted series ranging from the late 9th to the 17th century. They are executed in ornamental designs of great richness and diversity. Floral, geometric, and animal motifs are painted in vivid colours on a gold background around the canon tables of the Gospel manuscripts (concordances of the four Gospels), on the headpieces, and in the margins and are ingeniously adapted to the capital letters.

As regards iconography, the Gospel scenes follow early Christian and Byzantine models, but the Armenian painters, especially those of the medieval kingdom of Little Armenia, often displayed a marked independence and interpreted traditional formulas in a more lively or dramatic manner. Two artistic trends can be discerned in manuscript painting: one, more Eastern in character, tends to simplify the human form and subordinate it to ornamental interest; the other, under Byzantine influence, shows a subtle blending of naturalism and stylization. This latter trend was predominant in Little Armenia, where a flourishing school of painting developed under the patronage of the court and the church. The 13th-century manuscripts, in particular, belong in the first rank of medieval illumination. Through contacts with the crusaders and the Mongols, the painters of this period became acquainted with the art of the Latin West and of the Far East, and as a result they produced richly imaginative works.

Manuscripts continued to be illustrated throughout the Middle Ages in Armenian monasteries and in the various centres outside the area of Little Armenia where Armenians had settled after the destruction of the kingdom in 1375. These works are often inferior to those of the earlier period, but some original schools developed—for instance, in the area of Lake Van, especially at Khizan and on Aghthamar (modern Akdamar). (S.Der N./R.S.Co.)

Coptic Egypt. Coptic painting—strictly speaking, that practiced by Christians in Egypt from the time when

Christianity first took hold there—consists primarily of wall paintings in monasteries, the earliest foundations of which date from the 4th and 5th centuries.

Stylistically, Coptic painting differs from that of pagan Egypt in its emphasis on animal and plant ornamentation; less naturalistic rendering of the human form; simplified outline, colour, and detail; and increasingly monotonous repetition of a limited number of motifs.

In content, the wall paintings resemble other Christian examples of the genre around the eastern Mediterranean. The most usual theme is a frieze of saints with an enthroned figure of Christ or the Virgin. There is little variety of pose, though the features of individual saints are distinguishable. An unusually lively piece is a fragment from Wādī Sarga (now in the British Museum) depicting the Old Testament story of the three Hebrews in the fiery furnace; the Hebrews are dressed in Eastern garb and Phrygian hats and are shown as being protected from death by an angel (Figure 10). A celebrated set of wall paintings are those from Bāwīṭ, now in the Coptic Museum at Cairo.

By courtesy of the trustees of the British Museum

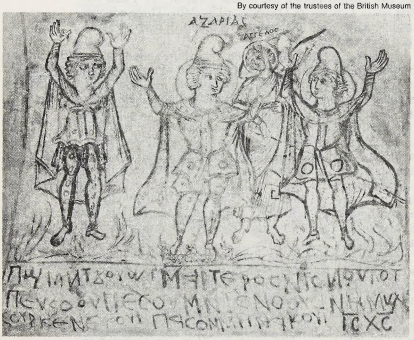


Figure 10: "The Three Hebrews in the Fiery Furnace"; fragment of a wall painting from a Coptic monastery at Wādī Sarga, Egypt, 6th century AD. In the British Museum, London.

Despite the 7th-century Muslim invasion of Egypt, there was no sudden break in the Coptic tradition. Indeed, some of the most notable surviving examples of manuscript illumination were produced during the first five centuries of Islāmic rule. It was only during the later Middle Ages that specifically Coptic painting ceased as Islāmic culture increasingly predominated. (A.F.Sh./R.S.Co.)

Western Dark Ages and medieval Christendom

DARK AGES

Ancient Roman civilization in western Europe foundered and fell apart in the second half of the 6th century, and the changes that took place between late antiquity and the succeeding period, the Dark Ages, were fundamental and catastrophic. Urban life collapsed, patronage of the arts all but ceased, and the centuries-old Mediterranean traditions of artistic training and production died out almost everywhere. It was only in a few places in Italy that artistic production continued unbroken, albeit much reduced. Increasingly the cultural fabric of northern Europe was determined by the various tribal peoples—Franks, Vandals, Goths, Angles, and Saxons—who migrated into the western provinces of the old Roman Empire during the 4th to 6th centuries and who established new patterns of settlement and centres of authority. Painting was not one of the traditional arts of these newcomers, though their craftsmen were expert workers of fine metals, leather, wood, and semiprecious stones (known as hardstones) such as garnet.

The reappearance of painting in northern Europe in the late 7th century was determined by two overriding factors. The first was the conversion of these peoples to Christianity. By the 6th century the Christian church had developed an extensive iconographic repertory, and Christian images were in use everywhere: both as icons, which functioned as focal points of worship, and as symbolic and narrative compositions, which proclaimed the mysteries of the faith and instructed the unlettered in the stories of sacred scripture. Painted images had become an indispensable apparatus of orthodox Christianity, and for the newly converted they would have been one of its most arresting and tangible features. The second factor that induced the new masters of Europe to develop the art of painting and figural imagery was their fascination with, and desire to emulate, the culture of the late Roman world, in which painting had been widely employed.

Apart from a small number of images on wooden panels, two kinds of painting have survived from the early Middle Ages: large-scale painting on the walls of buildings and small-scale painting in manuscripts. These two genres involved differing techniques and, to a large extent, constituted separate artistic traditions. Only a tiny percentage has survived of the wall paintings originally to be found in almost every church and in many public buildings throughout the West. Exposed to the destructive agencies of light, moisture, fire, general wear and tear, and changes in fashion, paintings on walls have little chance of surviving for more than a few hundred years. Illuminated books of this period, on the other hand, have come down in large numbers. Made of resilient animal skin and protected by stout wooden boards, they last almost indefinitely, and their decoration usually remains in a remarkably good state of preservation. It is fortunate that book production and decoration were a major concern of the early medieval church. Christianity was the religion of the Book; the words of Jesus Christ, the Gospel, were written down in this book, which Christ, the Logos (literally the "Word"), and his saints are often represented as holding in their hands. Artists in the Middle Ages expended some of their greatest efforts on the illumination and embellishment of the Gospels, the books of the Old Testament, and the other liturgical, devotional, and instructional texts that the church required.

The history of early medieval painting in the West is best examined in the art produced in five areas: Italy, the British Isles, France, Germany and Austria, and Spain.

Rome and Italy, c. 600–850. Rome, the seat of the pope, was one place in the West where an unbroken tradition of artistic patronage and production endured from late antiquity into the high Middle Ages and beyond. This was of inestimable importance for the history of the period from about 600 to 850, since it was to Italy and to Rome that the people of northern Europe looked for direction and for example.

Rome's unbroken tradition of patronage

The antique tradition of illusionistic naturalism continued in painting in Rome through the early Christian period; but toward 600 it weakened, and figures became flat and insubstantial. Increasingly, Jesus Christ, the Virgin, and the martyred saints of the church are represented alone or in groups, in strict hieratic frontality (in which the figures are arranged facing forward), gazing out to catch the eye of the onlooker. This development accompanied and served the growing cult of saints and the widespread practice of addressing images as focuses of prayer and veneration.

In the 7th and early 8th centuries successive waves of Byzantine influence dominated Roman patronage and artistic production. Rome at this time was still under the rule of the Byzantine emperor, and contacts with the Eastern capital were close. Various distinct Eastern pictorial traditions seem to have flourished side by side: hieratic figures and strictly symmetrical compositions in mosaic at the church of Sant'Agnese (625–638) and the chapel of San Venanzio at the Lateran Baptistery (642–649); faces carefully and vividly modeled to achieve astonishingly lifelike appearances at Santa Maria Antiqua (*e.g.*, the "Pompeian" Annunciation and St. Anne, early 7th century); and elsewhere in the same church figures fleetingly but effectively rendered in delicate washes of colour, so that they seem to scarcely materialize out of a dense, light-

suffused atmosphere (*e.g.,* Eleazar and Solomone and her seven sons, early 7th century).

Another strong and distinctive Byzantine wave hit Rome during the short papacy of John VII (705–707). Under his direct patronage, Eastern artists introduced an iconographic repertory new to the West, compositional schemes that were to endure for more than a century, and a vigorous new figural style (Figure 11).

Alinari/Art Resource, New York City

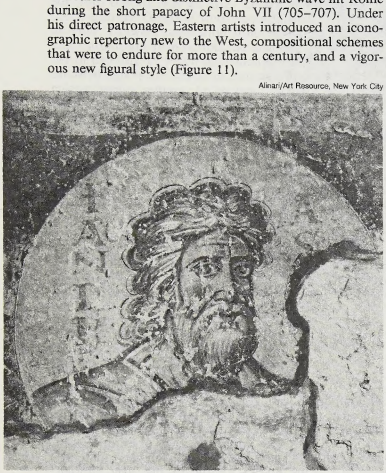


Figure 11: St. Andrew, wall painting in the presbytery of Santa Maria Antiqua, Rome, 705–707.

In the late 8th century a highly effective technique for representing the human figure was developed, in which modeling was almost completely eschewed and an eloquent system of brightly coloured lines was employed to define the clothed body. Examples include the painting of the Ascension (*c.* 850) in San Clemente, Rome, and the crypt (*c.* 830) of San Vincenzo al Volturno, in central Italy. In this technique wall painting was often used in conjunction with elaborate systems of white highlighting (*e.g.,* the Harrowing of Hell in the lower church of San Clemente and paintings [*c.* 870] in the Temple of Fortuna Virile).

Some of the finest work in Italy of the 8th and the first half of the 9th century was done in the north. At Castelseprio, north of Milan, a Byzantine artist painted a wonderfully light and vigorous cycle of the early life of Mary and the Nativity of Christ in a manner that bafflingly recalls the fluid impressionistic painting of early imperial Rome. Other wall paintings of this time, by native Italian masters, are at Cividale del Friuli, in San Salvatore in Brescia, and at Müstair. Contemporary paintings in the south show clear connections with this new Byzantine-influenced art of northern Italy (*e.g.,* San Vincenzo al Volturno, early 9th century).

England and Ireland, c. 650–850. It is recorded that Roman missionaries, who played a major role in the conversion of England to Christianity in the early 7th century, brought painted images with them; but next to nothing is known about painting on panels or walls in the British Isles during the Dark Ages. There is, however, a good deal of information about the illumination of manuscripts.

In the 6th and 7th centuries monasteries were founded and prospered, first in Ireland, later in England. In their scriptoria (writing rooms) manuscripts were written and decorated in increasingly elaborate fashion. In the Northumbrian double monastery of Monkwearmouth and Jarrow, Italian books and their illustrations were imitated extraordinarily faithfully (*e.g.,* the Codex Amiatinus, a great Bible, *c.* 700). But artists in other Northumbrian centres in the late 7th century began to adapt the standard decorative apparatus of late antique Italian manuscripts to very different effect. Portraits of the Evangelists became brilliant symbols, their bodies and clothes radically abstracted and brightly coloured; and, in the earliest books, they are sometimes shown in the guise of the four apocalyptic beasts, the man, the lion, the bull-calf, and the eagle, which represented the transcendental, celestial aspects of the four authors of the Gospels (*e.g.,* the Durrow Gospels, *c.* 680; the Echternach Gospels, *c.* 700). Artists in the British Isles also introduced other new elements, the most striking being richly ornamented cross-pages, commonly called "carpet pages," filled with ribbon interlace and wonderfully intertwined beasts, and large initial letters. The great full-page initial letters in Gospel books of the British Isles, besides articulating the text, serve as images, almost as icons, of the Word of God (Plate 4). These manuscripts are distinguished by their extraordinary ornamental repertory, drawn from the native Celtic tradition, from the Mediterranean, and from the tradition of fine metalworking introduced by Anglo-Saxon settlers in Britain in the 6th century.

In the 8th century there were flourishing scriptoria also in the south of England, and several manuscripts prepared at Canterbury have been identified (*e.g.,* the Vespasian Psalter, *c.* 730–740; the Stockholm Codex Aureus, or "Golden Gospels," *c.* 750). In early 9th-century books from the south, formal and iconographic elements introduced from Frankish scriptoria across the Channel are in evidence.

It is not yet possible to distinguish between different Irish schools of illumination. The outstanding manuscripts are the St. Gall Gospels (*c.* 750), the great Book of Kells (*c.* 800), the Gospels of Macregol (early 9th century), and a group of little "pocket Gospel books."

The innovations of these early Irish and English scribes and artists left a lasting imprint on the subsequent development of book decoration throughout Europe. The elaborate initial letters that are found in nearly all later decorated manuscripts were first devised in the British Isles, and the decorative vocabulary of later continental illumination owed much to English and Irish invention.

Merovingian Gaul. It was only in the first half of the 8th century that manuscripts began to be elaborately decorated in the Frankish kingdom (an area roughly comprising northern France and southwestern Germany as far as the Rhineland). This production is known as Merovingian, after the Frankish dynasty that ruled, in name at least, until 751. In its subject matter, early Frankish illumination is decorative and symbolic rather than narrative. The idea of stressing the initial letters of a text was adopted from the British Isles, but the results were rather different. The strokes of letters are shaped like doves and fish with swelling bodies, or they are filled with simple ornamental motifs. A favoured frontispiece is a large cross standing within an arch, incorporating or surrounded by animals and birds of all kinds (*e.g.,* the Gelasian Sacramentary [Figure 12]; St. Augustine, *Quaestiones in Heptateuchon,* Laon, *c.* 750). The rare instances of figural composition from mid-8th-century France are usually rather ungainly copies of late antique prototypes (*e.g.,* the Gospels of Gundohinus, Laon, 754). It was only in the second half of the century, probably as a result of English influence, that figure drawing was subjected to a controlled linear discipline (as in the Flavigny Gospels). This development culminated about 800 in the wonderfully inventive historiated (decorated with figures of men or animals) initials in the Corbie Psalter.

EARLY MIDDLE AGES

Carolingian Empire. In the mid-8th century a new Frankish dynasty came to power. Under Charlemagne, whose long reign lasted from 768 to 814 and who was crowned the first emperor of the Romans in 800, a new courtly culture was created to rival those of late antique Rome and of contemporary Byzantium. The achievements of two groups of artists, members of both of which worked for the Emperor and his court, were to determine the overall development of painting in northern Europe for the next three centuries. One group, the so-called Court school, produced a series of splendidly rich Gospel books.

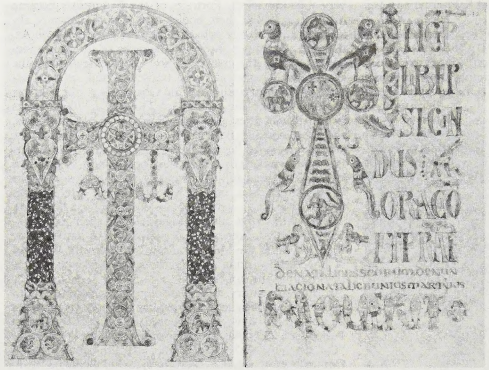


Figure 12: Frontispiece and initial page from the Gelasian Sacramentary, northern France, *c.* 750. In the Biblioteca Apostolica Vaticana, Vatican City (MS. Vat. Reg. Lat. 316, fol. 131v–132r). 26 × 17.1 cm.
Biblioteca Apostolica Vaticana, Vatican City

Their decoration is extremely inventive, even witty, and the figures, with carefully modeled limbs issuing from dense carapaces of brilliantly coloured, elaborately folded drapery, show a completely new mastery of the human form (Plate 4). The second group concentrated on figures dressed in archaic white garments, with faces and limbs modeled in dramatic chiaroscuro (contrasts of light and shade)—a conscious and very successful evocation of the painting of antiquity (*e.g.,* the Coronation Gospels in Vienna, *c.* 795–800).

During the years from about 815 to 835 an extremely active and inventive scriptorium flourished at Rheims, under the patronage of the archbishop, Ebbo. Inspired by the masters of the Coronation Gospels, the Rheims artists aimed at producing work intentionally reminiscent of the art of classical antiquity. However, an extraordinary new spirit of linear excitement pervades their compositions, in such works as the Gospels of Ebbo, the Utrecht Psalter, and the Physiologus at Bern. These are some of the most vital and ecstatic creations of the early Middle Ages.

Leading schools of later Carolingian illumination were located at Tours, Saint-Amand (in what is now Belgium), Metz, St. Gall, and at an unidentified scriptorium from which Charlemagne's grandson, Charles the Bald, commissioned a number of extraordinarily lavishly decorated manuscripts in the 860s and 870s.

The early Carolingian artists reintroduced figurative painting and pictorial narrative to northern Europe. To achieve this, they studied monuments and manuscripts surviving from late antiquity and contemporary works from Italy, the British Isles, and Byzantium. They borrowed freely and exuberantly, but they were rarely mere copyists. Vitality and invention were always paramount. This remarkable achievement was the result of determined and demanding patronage and of intense creative effort.

The 10th century. The late 9th and the first half of the 10th century had been a time of economic depression, social upheaval, and political reorganization throughout western Europe. There had followed a period of reconstruction, with new ruling dynasties emerging and consolidating their power. Although production of wall paintings and manuscripts had continued; the energies of patrons had been directed elsewhere, and there had been a distinct decline in production. Only in the third quarter of the 10th century did renewed patronage lead to an outburst of artistic activity and invention.

Late Anglo-Saxon England. In England a coherent and magnificent style of book illumination was developed in the 960s in the scriptorium at Winchester. Narrative compositions and initial letters are framed in arched and rectangular bossed (articulated with circular and square ornamental motifs) trellises of golden bars filled with rampant foliage; figures are clothed in shells of brittle broken drapery, with elaborate zigzagged contours and fluttering hems (*e.g.,* King Edgar's Charter to the New Minster, Winchester, 966; the Benedictional [a book of episcopal Eucharistic blessings] of St. Ethelwold, 971–984 [Plate 4]). During the following century scriptoria in southern England produced a considerable number of books of this kind, filled with flickering colour and glinting gold and intended for ceremonial liturgical use. Behind this initiative in lavish book production lay a movement of religious reform, instituted by the leading churchmen of the realm and supported by the king.

In the scriptoria at Glastonbury and Canterbury a lively tradition of expressive outline drawing developed, and some of the most arresting Anglo-Saxon works of the period are filled with animated figures in flying ruffled drapery (*e.g.,* the Leofric Missal, 970s; Harley Psalter, early 11th century).

English artists of this time delighted in iconographic invention. The results were sometimes startling, and the innovations often endured: the horns of Moses (for example, see Plate 5, although this is a much later work) and Christ disappearing into clouds at his Ascension were both English inventions of the early 11th century.

France. Continuing Carolingian traditions of illumination can be traced in many centres in France, but it is only at the very end of the 10th century that a new energy is apparent in scriptoria in the north, reflecting a reforming spirit in the church. At Fleury, Saint-Bertin, and Saint-Vaast at Arras, imported works from England and the presence of English artists gave a fresh impetus to manuscript illumination. Spirited outline drawings, inspired by English example, were set alongside frames and initial letters of Carolingian ancestry (*e.g.,* the Psalter and Gospels of Odbert of Saint-Bertin, *c.* 1000; Bible of Saint-Vaast, early 11th century).

Ottonian Germany. In Germany, now under the Saxon Ottonian dynasty, concerted royal and ecclesiastical patronage also brought about a great revival in the arts. As in England, this revival followed a reform movement that touched all the leading monastic communities and revitalized religious life throughout the land.

Ottonian art, like Anglo-Saxon, was solidly based on earlier Carolingian invention; and the illustrations in one of the earliest Ottonian books, the Gospel Lectionary (a book of Gospel lessons for the church year) of Gero (*c.*

960), were copied line for line from a manuscript of Charlemagne's Court school. The dominant figure in the late 10th century was an artist known as the Master of the Registrum Gregorii, who seems to have been based at Trier. Drawing inspiration from both early Christian and Carolingian manuscripts, he developed a new manner of painting, in which meticulously detailed, smoothly modeled figures are placed in elaborate and precisely calculated spatial settings. In his work, volume and planar design interact in dynamic tension (as in the Letters of Gregory the Great, *c.* 983; the Gospel Lectionary of Egbert of Trier, from the 980s; and the Gospels of Sainte-Chapelle, *c.* 1005).

In about 1000, younger contemporaries of this man who had learned much from his art produced, on royal commission, a series of magnificently illuminated books in which brilliantly lighted figures move with a supernatural grandeur against golden grounds and bands of colour (examples include the Gospel Books of Otto III in Aachen and Munich, *c.* 1000; the Gospel Lectionary of Henry II, 1002–14; and the Apocalypse and Commentaries on Daniel and Isaiah, early 11th century). The portraits of the Evangelists and the imperial images in these books are remarkable for their formal subtlety and iconographic ingenuity (Plate 4).

During the first half of the 11th century, manuscript illumination flourished in various monastic scriptoria in Germany. The inventions and example of the Master of the Registrum Gregorii largely determined developments at Echternach and Cologne. At Cologne, Eastern painted books must also have been available as models, since the wonderfully fluid painterly compositions of the early works of the school appear to have been inspired by contemporary Byzantine painting (as in the Gospels of Abbess Hitda of Meschede, early 11th century). At Regensburg the splendid house style was based largely on one grand Carolingian book, the golden Gospels of Charles the Bald, in the possession of the Abbey of St. Emmeram. In this scriptorium, illustrations became vehicles for elaborate theological arguments, laid out in complex schematic compositions and glossed with explanatory inscriptions (*e.g.,* the Sacramentary [a service book typically containing the celebrant's part of the mass together with various prayers] of Henry II, 1002–14; the Gospels of Abbess Uta, early 11th century). At Corvey, on the other hand, book illumination was ornamental and largely aniconic. The ornamentation consisted chiefly of darkly brilliant initial pages, with large gilded capital letters set on densely patterned purple grounds (as in the Wernigerode Gospels, *c.* 970).

From literary sources and fragmentary remains it is known that wall painting was common in Germany during this period. But only one extensive program survives, in the Church of St. George on the island of Reichenau, in Lake Constance. This dates from the late 10th century and consists of a sequence of the miracles of Christ's ministry, narrated with great drama and psychological intensity.

ROMANESQUE

In the second half of the 11th century in many parts of Europe new energies and new initiatives are apparent in painting, sculpture, and architecture. It is impossible to categorize these changes fully or to reduce them to a common denominator, but in many places there was a tendency toward greater schematization and bold configurations in design, in which strong and abstract structures of line and colour predominate. The surfaces of clothed bodies are enlivened by intricate schemes of folds and pleats and highlights in regular patterns of reiterated parallel and converging lines. These developments are partly explained by the arrival in the West of examples of recent Byzantine painting, with its elaborate patterned highlighting. Another factor seems to have been an aesthetic that defined beauty in terms of symmetry and order and the juxtaposition of pure, bright saturated colours.

Italy. In Italy the critical role played by Byzantine art is clearest of all. It is evident both in the north, particularly in Venice, and in the south at Montecassino, where Byzantine artists were summoned by the abbot Desiderius in the 1060s to work on the decoration of his new abbey church. The wall paintings commissioned by the same Desiderius at Sant'Angelo in Formis, near Capua, are the outstanding surviving example of the consequent fusion of Eastern and Western traditions (Plate 5). In Rome and central Italy in the first decades of the century, the dominant fashion was for figures whose garments hung in a multitude of fine parallel pleats (as in the triptych of the Redeemer in the cathedral at Tivoli and in the wall paintings at Castel Sant'Elia di Nepi and in Santa Pudenziana in Rome). In the 12th century Italian artists took an increasing interest in ancient Roman art, nowhere more so than in Rome itself, where there was a veritable renaissance of classical and early Christian compositional formulas, motifs, and even styles.

France. An early Romanesque art emerged in scriptoria throughout France in the late 11th century—at Saint-Omer in the north, at Mont-Saint-Michel in the northwest, at the abbey of Saint-Aubin at Angers in the west, at Limoges in central France, and at Toulouse in the south.

In the early 12th century, major schools of painting flourished in Burgundy, at the great Benedictine abbey of Cluny, and at the newly founded Cistercian house of Cîteaux. From Cluny there is a lectionary in which Byzantine influence is strong and a copy of St. Ildefonsus' treatise on the virginity of Mary, with stiff, gorgeously coloured and gilded compositions owing more to late Ottonian examples than to Byzantium. There are also wonderful wall paintings in the Cluniac chapel at Berzé-la-Ville, where the various compositions are filled with energy and colour, and a tumult of fine sweeping folds and flickering highlights plays over the surface of the drapery. At Cîteaux the early manuscripts show evidence of strong Norman and English influence in their decoration and a satirical delight in observation (as in Gregory the Great's Moralia in Job, 1111). Later, in a group of manuscripts of the second quarter of the century, the illustrations are colour-washed drawings with slender, lyrically conceived figures whose drapery falls in cascades of parallel rounded pleats, apparently inspired by contemporary southern Italian work (*e.g.,* St. Jerome's Commentary on Isaiah, the Cîteaux Lectionary). Cluny

The most complete surviving set of early Romanesque wall paintings in France is in the church of Saint-Savin-sur-Gartempe, where the compositions show great narrative vigour and inventiveness. Quite startling formal mannerisms sometimes occur in provincial French painting of the first half of the 11th century. Two examples are the wonderfully highlighted and emphatically banded and pleated figures at Vicq-sur-Saint-Chartrier and the violently expressive gesturing figures on the vaults of the crypt of Saint-Nicolas at Tavant. In general, wall painters in the early Middle Ages had very limited means at their disposal, and it is remarkable how skilled artists were able to deploy three or four colours to impressive and unifying effect. An example of this is at Montcherand, in the Swiss Jura, where simple hues of brown, ochre, dull blue, and white have been used to depict ecstatically disputing Apostles beneath a huge Christ in Majesty, in a composition of bright abstract subtlety and strength.

England. In the 1120s in England artists at the abbey of St. Albans, drawing on earlier English traditions and Ottonian painting from Germany, devised cycles of full-page scenes with large, emphatically gesturing figures set off against rectangular panels of colour, often within architectural settings. In structural density, in their use of accumulated motifs and bright areas of colour, and in the intensity of their storytelling, these images (*e.g.,* the Psalter in St. Godhard, Hildesheim, and the Life of St. Edmund) have few parallels in earlier English art.

In the second quarter of the century acquaintance with contemporary Byzantine painting—probably via illuminated manuscripts—and recent developments on the Continent led English artists to a more organic, if expressively attenuated, conception of the human body. Drapery is now stretched and gathered, with sinuous folds isolating curving islands of taut cloth (so-called damp-fold drapery) to describe three-dimensional forms in torsion. Faces are more heavily modeled than before, and glances and ges-

tures are even more piercing and insistent. This is first seen about 1130 in the great Bible of the Abbey of St. Edmund at Bury (Plate 5); later stages of the development can be traced in a series of magnificent manuscripts from southern English scriptoria (*e.g.*, the Dover Bible, the Lambeth Bible, the Psalter of Henry of Blois, and the Bodleian Terence) and in the wall painting of St. Paul and the viper in St. Anselm's Chapel in Canterbury cathedral (1160s).

In the late 11th century in southern England and in northern France a type of initial letter emerged in which men, monsters, beasts, and birds climb and struggle in "tanglewoods" of rinceaux (ornamental motifs consisting of sinuous and scrolling foliate branches). These ingenious constructions, full of movement and variety, fired the imaginations of artists throughout Europe. On the surface they are an expression of that love of joyously outlandish, grotesque, and even warring imagery that is a ubiquitous feature of 12th-century art; but at a deeper level they are concerned with man's unending conflict with sin and the Devil (Figure 13).

The Bodleian Library, Oxford



Figure 13: Initial letter U to the Book of Job in a Bible from Winchester, *c.* 1160. In the Bodleian Library, Oxford (MS. Auct. E. Infra I, fol. 304*r*). 14.3 × 14 cm.

Spain. An extraordinary and idiosyncratic tradition of manuscript illumination evolved in Spain in the 10th and 11th centuries. The chief vehicle for this art was the commentary on the book of Revelation of Beatus of Liebana, a text that seems to have been taken by contemporaries as a symbol of Christian resistance to the Muslim Arabs who dominated much of the Iberian Peninsula in the early Middle Ages. The Arab cultural presence in Spain was all-pervasive, and—even if it did not account for the strongly patterned, sometimes barbaric compositions and for the brilliant jarring use of colour—it was responsible for particular motifs adopted by these illuminators (such as the horseshoe arch) and for the common practice of recording in a manuscript's colophon the scriptorium, the scribe and artist, and the date of the manuscript itself.

Northern Spain also produced some of the most splendid Romanesque wall paintings. Spanish artists favoured formal symmetrical and hieratic compositions and strong, barely modulated colours. The human form and the stiff, banded drapery that encases it are consistently more idealized and abstracted than in other European painting of the time. At their finest, these works possess a hypnotic numinous power (Plate 5).

The Meuse Valley. The results of the great increase in artistic production, the sudden intensification of patronage, and the wealth of artistic invention found throughout Europe in the late 11th and early 12th century are nowhere more clear than in the valley of the Meuse, in what is now eastern Belgium. One of the leading centres of artistic production was the abbey of Stavelot. The decoration of the outstanding early manuscript from its scriptorium, the Stavelot Bible, of about 1094–97, is the work of various hands and is a perfect microcosm of the influences and interests that gave rise to the first Romanesque painting. The majestic enthroned Christ clearly has his ancestry in Ottonian compositions from the nearby scriptorium at Echternach. Some of the historiated initials are inhabited by delicately drawn figures that seem to stem from the old English tradition of outline drawing. Others incorporate large, darkly modeled figures that look strikingly Byzantine. And the great initial of the book of Genesis has a complex program in which scenes from the Old and New Testaments are juxtaposed in a tree of medallions to demonstrate the scheme of Redemption. The concept of expounding a theological argument in a composition of diagrammatic complexity is something that was dear to the 12th century.

Germany and Austria. Full-page compositions of complex iconography in elaborate formal settings are also a characteristic of north German manuscripts of the 12th century. They are found above all in a group of books associated with the all-powerful duke of Saxony Henry the Lion (1142–95) and prepared in the abbey of Helmarshausen on the Weser River. This scriptorium's masterpiece is a Gospel book presented by Henry and his wife Matilda to Brunswick cathedral in 1173–75. The illumination is extraordinarily rich and dense, with a solemn and magisterial palette of gold, purple, dark green, azure, ochre, and white. The elaborate iconographies are glossed in long scrolls, which form undulating accents across the pages (Plate 5).

A very different art was practiced in the southeast, where Salzburg was the leading centre. A strong Italian element is detectable in the illustrations in books of the first half of the 12th century, such as the giant Bible at Michaelbeuern and the Admont Bible of 1140–50. The latter manuscript—which features large, full-page compositions dominated by tall turning figures, unreal landscapes, and bright colours—is a parallel phenomenon to the great contemporary English books, such as the Lambeth Bible and the Psalter of Henry of Blois. Each shows a preoccupation with Byzantine models for figures and faces. But the strong Italian influence in the Salzburg scriptorium ensured that the German figures are calmer and more solid than their exuberant English cousins. In the middle of the century a wonderfully elegant art of pen drawing emerged at Salzburg, with expressive swaying and gesticulating figures set against backgrounds of blue and green (the Antiphonary of St. Peter's at Salzburg).

A number of early wall paintings survive in Austria and Germany, but many of those in Germany have suffered disastrously from over-restoration. In Austria the major monument is the late 11th-century Christological cycle in the west choir of the abbey Church at Lambach, apparently by artists from Salzburg. This work was strongly influenced by the contemporary Byzantinizing art of the Veneto. Salzburg painting of the 1150s can be seen in a lyrical female figure personifying the Third Hour, in the monastic Church of St. Peter.

In Germany well-preserved paintings of the early 12th century at Idensen, in lower Saxony, have strong four-square compositions and clearly contoured, stern-faced figures, which stem from late Ottonian tradition. Half a century later, on the lower Rhine, a new spirit and mentality were expressed in two splendid but drastically repainted cycles at Schwarz Rheindorf and at Brauweiler, near Cologne, where elegantly drawn figures play against panels and frames of blue and green, illustrating recondite and complicated iconographic programs.

LATE 12TH CENTURY

In the late 12th century two broad developments took place in wall painting and manuscript illumination throughout the West. On the one hand, forms became smoother and more fluent, and a less abstract and less aggressively patterned interpretation was put on nature. On the other

Minoan and Greek painting



Youthful fisherman with his catch, fresco on the wall of the West House at Akrotiri, Thera, Greece, 15th century BC.



"The Toreador Fresco," restored wall painting from the Palace at Knossos, Crete, *c.* 1550 BC. In the Archaeological Museum, Iráklion, Crete. Height (including borders) 81 cm.

"Three Companions of Odysseus Encountering the Daughter of Antiphatis, King of the Laestrygones," Roman Second-style wall painting from a house on the Esquiline Hill, Rome, possibly based on a Hellenistic painting, *c.* 1st century BC. In the Biblioteca Apostolica Vaticana, Vatican City. Height 1.4 m.



Plate 1: (Top) Ministry of Culture, Archaeological Receipts Fund, Greece. (centre) SCALA/Art Resource—EB Inc., (bottom) Hoile Bildarchiv, Baden-Baden, W.Ger.

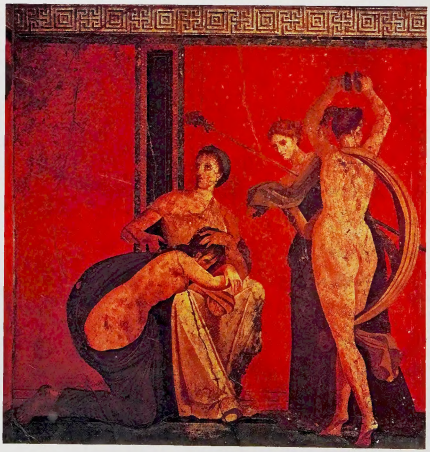
Greek and Roman painting



Hades abducting Persephone, wall painting in the small royal tomb at Vergina, Macedonia, 4th century BC.



The family of Vunnerius Keramus, medallion of gold and glass, *c.* AD 250. In the Museum of Christian Antiquities, Brescia, Italy. Diameter 6 cm.



Dionysiac initiation rites and prenuptial ordeals of a bride, wall painting, Second style, *c.* 50 BC. In the Villa of the Mysteries, Pompeii, Italy.

Plate 2: (Top) Manolis Andronicos, (centre) EDI Studio, Barcelona, (bottom) SCALA/Art Resource—EB Inc



"Anastasis" ("Christ Ascending from Hell"), apse fresco, *c.* 1320. In the Church of St. Saviour at the Monastery of the Chora (Kariye Cami), Istanbul.



Rebekah at the well, illuminated miniature from the Vienna Genesis, 6th century. In the Österreichische Nationalbibliothek, Vienna (Cod. theol. graec. 31, fol. 7).



"Our Lady of Vladimir," tempera on wood, from Constantinople, *c.* 1130. In the State Tretyakov Gallery, Moscow. 78 × 55 cm.

Byzantine painting



Portrait of the high admiral Alexius Apocaucos, illuminated manuscript page from the Hippocrates Manuscript, *c.* 1342. In the Bibliothèque Nationale, Paris (Cod. Gr. 2144). 41.5 × 35 cm.

Plate 3: By courtesy of (top) the Dumbarton Oaks Field Committee, Washington, D.C., (centre right) the Österreichische Nationalbibliothek, Vienna, (bottom right) Bibliotheque Nationale, Paris; photograph, (left) Novosti, Moscow

St. Mark, illuminated manuscript page from the Gospel Book of the Court school of Charlemagne, *c.* 810. In the Stadtbibliothek, Trier. W.Ger. (Cod. 22, fol. 59v). 36.6 × 24.5 cm.





Nativity, Benedictional of St. Ethelwold, Winchester, Hampshire, 971–984. In the British Library, London (Add. MS. 49598, fol. 15v). 29.3 × 22.5 cm.

Early Medieval and Romanesque painting



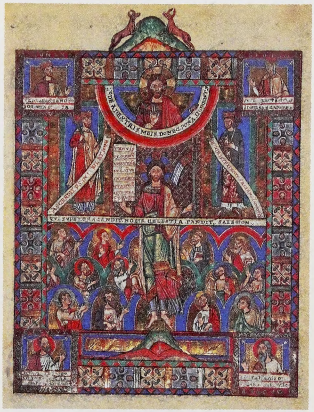


Liber Generationis, initial page from the beginning of St. Matthew in the Lindisfarne Gospels, *c.* 700. In the British Library, London (Cotton MS., Nero D.IV, fol. 27r). 34 × 24.5 cm.

St. Luke, illuminated manuscript page from the Gospel Book of Otto III, school of Reichenau, *c.* 1000, in the Bayerische Staatsbibliothek, Munich (CLM. 4453, fol. 139v). 24 × 19 cm.

Plate 4: By courtesy of (top left) the Stadtbibliothek, Trier, W.Ger., (top right, bottom right) The British Library, London, (bottom left) the Bayerische Staatsbibliothek, Munich

Ascension, illuminated manuscript page from the Gospels of Henry the Lion, from the abbey of Helmarshausen, c. 1173–75. In the Herzog August Bibliothek, Wolfenbüttel, W.Ger. (Cod. Guelf. 105 Noviss. 2°). 34.2 × 25.5 cm.





"Christ in Majesty," apse fresco from San Clemente, Tahull, Spain, by the Master of Tahull, 1123. In the Museum of Catalan Art, Barcelona.

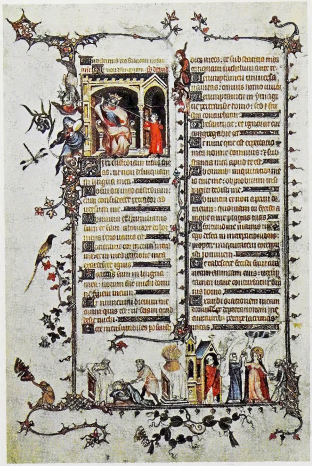


Moses expounding the law of the unclean beasts, illuminated manuscript page from the Bury Bible, about 1130. In Corpus Christi College, Cambridge (MS. 2, fol. 94r). 51 × 35.7 cm.

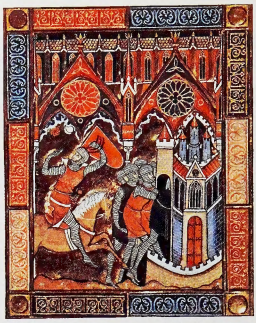


Jesus at Bethany in the house of Simon the leper, detail of a fresco in Sant'Angelo in Formis, near Capua, Italy, second half of the 11th century.

Plate 5: By courtesy of (top right) the Herzog August Bibliothek, Wolfenbuttel, W.Ger., (bottom right) the Master and Fellows of Corpus Christi College, Cambridge, photographs, (top left) SCALA/Art Resource, New York City—ES Inc., (bottom left) Giraudon/Art Resource, New York City



David before Saul, page from the Belleville Breviary, illuminated manuscript attributed to Jean Pucelle, c. 1325. In the Bibliothèque Nationale, Paris (MS. lat. 10 483, fol. 24v). 24 × 18 cm.



Joshua bidding the Sun stand still, illuminated miniature from the Psalter of St. Louis, c. 1260. In the Bibliothèque Nationale, Paris (MS. lat. 10 525, fol. 46). 12 × 9 cm.



Angel blowing the trumpet, illuminated miniature from the Douce Apocalypse, c. 1270. In the Bodleian Library, Oxford (Douce MS. 180, p. 26). 15 × 11 cm.

Gothic painting in Italy, France, and England



"Annunciation," tempera on wood by Simone Martini, 1333 (saints on either side of the central panel by Lippo Memmi). In the Uffizi, Florence. 3.1 × 2.7 m.

Plate 6: By courtesy of (top left, top right) the Bibliotheque Nationale, Paris, (centre right) Bodleian Library Filmstrips, Oxford; photograph, (bottom) SCALA/Art Resource—EB Inc.



October, from the *Très Riches Heures du duc de Berry*, illuminated manuscript page by the Limburg brothers, c. 1416. In the Condé Museum, Chantilly, Fr. (fol. 10v). 29 × 21 cm.



"Adoration of the Magi," tempera on wood by Gentile da Fabriano, 1423. In the Uffizi, Florence. 3 × 2.8 m.

Initial letter of the Magnificat, from the Book of Hours of Gian Galeazzo Visconti, illuminated manuscript page by Giovannino de' Grassi, c. 1385. In the Biblioteca Nazionale Centrale, Florence (Fondo Landau-Finaly MS. 22, fol. 147v).





Cuer reading the inscription on the magic fountain, page from *Le Livre du coeur d'amours éspris*, illuminated manuscript attributed to René of Anjou, 1465. In the Österreichische Nationalbibliothek, Vienna (MS. 2597, fol. 15r). 29 × 21 cm.

Plate 7: By courtesy of (bottom left) the Österreichische Nationalbibliothek, Vienna; photographs, (top left, bottom right) SCALA/Art Resource—EB Inc., (top right) Giraudon/Art Resource—EB Inc.



"Descent from the Cross," tempera on wood by Rogier van der Weyden, *c.* 1435–40. In the Prado, Madrid. 2.2 × 2.7 m.



"The Resurrection," open right panel of the "Isenheim Altarpiece," oil on wood by Matthias Grünewald, completed before 1516. In the Unterlinden Museum, Colmar, Fr. 2.7 × 1.4 m.

Gothic painting in Germany and the Low Countries



"Hell," open right panel of the "Garden of Delights" triptych, oil on wood by Hieronymus Bosch, *c.* 1505–10. In the Prado, Madrid. 2.2 m × 97 cm.

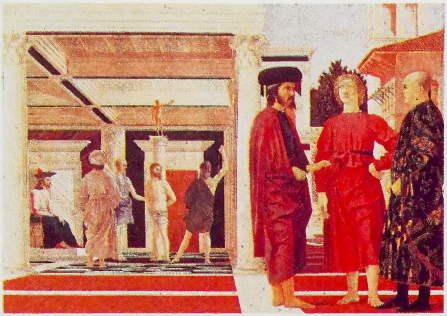


"Madonna with Chancellor Rolin," oil on wood by Jan van Eyck, *c.* 1435. In the Louvre, Paris. 66 × 82 cm.

Plate 8: By courtesy of (top left) the Museo del Prado, Madrid; photographs, (top right, bottom right) SCALA/Art Resource—EB Inc., (bottom left) Archivo Mas, Barcelona



"Arrival of Cardinal Francesco Gonzaga," fresco by Andrea Mantegna, completed 1474. In the Camera degli Sposi, Palazzo Ducale, Mantua, Italy.



"Flagellation of Christ," oil on wood by Piero della Francesca, late 1450s. In the National Gallery of the Marches, Urbino, Italy. 59.5 × 82 cm.

Renaissance painting of the 15th century in Italy



"The Tribute Money," fresco by Masaccio, 1425. In the Brancacci Chapel, Santa Maria del Carmine, Florence. 2.5 × 6m.



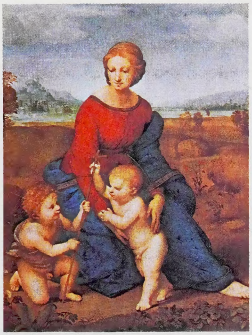
"The Primavera," tempera on wood by Sandro Botticelli, 1477–78. In the Uffizi, Florence. 2.1 × 3.2 m.

Plate 9: (Top left) SCALA/Art Resource, New York City—EB Inc., (others) SCALA/Art Resource, New York City



"The Bacchanal," oil painting by Titian, 1518–19. In the Prado, Madrid. 1.8 × 1.9 m.

"The Tempest," oil painting by Giorgione, *c.* 1505. In the Gallerie dell'Accademia, Venice. 77 × 74 cm.



"Madonna of the Meadows," oil on wood by Raphael, 1505. In the Kunsthistorisches Museum, Vienna. 1.1 m × 87 cm.





"The Last Supper," fresco by Leonardo da Vinci, 1495–97. In the refectory of the convent of Santa Maria delle Grazie, Milan. 8.8 × 4.6 m.

Plate 10: By courtesy of (centre left) the Kunsthistorisches Museum, Vienna, (bottom) the Pinacoteca di Brera, Milan; photographs, (top) Archivo Mas, Barcelona, (centre left) Erwin Meyer, (centre right) SCALA/Art Resource—EB Inc., (bottom) Angelo Coi



"The Last Judgment," fresco by Michelangelo, 1533–41. In the Sistine Chapel, Vatican, Rome. 14.5 × 13.5 m.



"Assumption of the Virgin," oil painting by El Greco, 1577. In the Art Institute of Chicago. 4 × 2.3 m.

Renaissance painting of the 16th century in Italy and Spain

"The Last Supper," oil painting by Tintoretto, 1594. In the chancel, S. Giorgio Maggiore, Venice. 3.5 × 5.6 m.





"Madonna dal Collo Lungo" ("Madonna of the Long Neck"), oil on wood by Parmigianino, *c.* 1535. In the Uffizi, Florence. 2.2 × 1.3 m.

Plate 11: By courtesy of (top left) the Art Institute of Chicago; photographs, (top right, bottom left, bottom right) SCALA/Art Resource—EB Inc.



"Peasant Dance," oil on wood by Pieter Bruegel the Elder, *c.* 1568. In the Kunsthistorisches Museum, Vienna. 1.1 × 1.6 m.

Renaissance painting of the 16th century in Germany and the Low Countries



"Venus in a Landscape," oil on wood by Lucas Cranach the Elder, 1529. In the Louvre, Paris. 38.1 × 25.9 cm.



"Four Apostles," oil on two wood panels by Albrecht Dürer, 1526. In the Alte Pinakothek, Munich. Each panel 2.2 m × 77 cm.



"Portrait of the Merchant Georg Gisze," oil and tempera on wood by Hans Holbein the Younger, 1532. In the Staatliche Museen Preussischer Kulturbesitz, West Berlin. 96.3 × 86 cm.

Plate 12: By courtesy of (top left) the Kunsthistorisches Museum, Vienna, (top right) Musee du Louvre, Paris, (bottom left) the Alte Pinakothek, Munich, (bottom right) the Staatliche Museen Preussischer Kulturbesitz, West Berlin; photographs, (top left) Erwin Meyer, (top right) The Bridgeman Art Library, London, (bottom left) Joachim Blauel

hand, the perennial interest that Western artists had shown in contemporary Byzantine art grew more intense, and this sometimes led to the opposite extremes of turbulent and mannered design. Both of these tendencies probably aimed at representing human actions and interactions with greater conviction and increased psychological power.

In England a new soft style is apparent in the later hands responsible for illuminating the great Winchester Bible in the 1170s. There, all traces of the elaborately patterned damp-fold drapery of mid-century painting have vanished, to be replaced by material that falls in tiny ripples and soft irregular undulations to reveal firm limbs beneath. A later, simplified, mannered, and frenzied version of this style is found in the illustrations of a bestiary from the Midlands of the early 1200s (Figure 14). But the rounded, billowing drapery of the enthroned Christ in the contemporary Westminster Psalter seems to have left the 12th century far behind. This is pure Early Gothic painting.

The Bodleian Library, Oxford



Figure 14: "The Unicorn," miniature from an English Bestiary, early 13th century. In the Bodleian Library, Oxford (MS. Ashmole 1511, fol 14v). 10.4 × 11.5 cm.

A similar evolution can be traced in northern France, in books such as the Capucin's Bible from Champagne and in the Souvigny Bible from central France, in which Byzantine influence is strong. A variation, which originated in the Meuse Valley, was the so-called *Muldenfaltenstil,* named after the small, troughlike folds into which drapery breaks (*e.g.,* the Psalter of Queen Ingeborg, northern France, *c.* 1200). In Germany this style is found in manuscripts made on the middle Rhine and at Regensburg.

Byzantine influence in Italy

The other major factor in European art about 1200 was a widespread interest in Byzantium. Byzantine mosaicists in the late 12th century undertook vast commissions in Venice and Sicily, and these provided Western artists with the opportunity of studying monumental Byzantine art of the finest quality at first hand. Imported Byzantine illuminated manuscripts and panel paintings, enamels, and ivory carvings were also available as models. The purest and most striking instances of Byzantinizing painting are found in Italy. In Venice local craftsmen, trained by Byzantine masters, designed and laid mosaics that are almost indistinguishable from genuine Byzantine work; and in the cathedral at Aquileia the standing prophets and saints painted on the vaults of the crypt look as if they had walked straight out of an Eastern atelier. In Rome and its environs, too, the development of painting in the first half of the 13th century was determined largely by the extensive new programs of mosaics in Norman Sicily, at Palermo, Cefalù, and Monreale.

This swirling, contrived Byzantine art of the middle to late 12th century gave rise to many experiments in northern Europe. It strongly affected artists at Salzburg (*e.g.,* the drawing of Christ in Majesty in Vienna, Österreichisches Nationalbibliothek, MS. 953) and on the upper Rhine (*e.g.,* the Gospel Lectionary from Speyer of 1196, in Karlsruhe), and it underlies the many figures in the great Tree of Jesse on the ceiling of the Church of St. Michael at Hildesheim, figures conceived in elaborate three-dimensional attitudes, with angular broken drapery. Finally, the *Zackenstil*—the new, elegant, early Gothic, jagged style of early 13th-century Germany, most magnificently exemplified in the Saxon Gospels in Goslar—was directly inspired by contemporary Byzantine painting.

In the early Middle Ages, wall painters had largely been laymen, whereas the illumination of manuscripts had been practiced almost exclusively in monastic scriptoria. In the late 12th century the production of books began to be taken up by lay scribes and painters working in their own shops. At the same time, illuminated books of private devotion and both religious and secular illustrated texts became increasingly popular. This process continued in the 13th century, when growing literacy and learning among laymen and the rise of the universities created a demand for illuminated and illustrated texts of all kinds. (Jo.B.M.)

GOTHIC

Gothic is the term generally used to denote the style of architecture, sculpture, and painting that developed from the Romanesque during the 12th century and became predominant in Europe by the middle of the 13th century. The many variations within the style are usually distinguished by the use of chronological or geographical terms (for example, early, high, Italian, International, and late Gothic).

Pictorial style

Early Gothic. One of the moves away from Byzantine influence took the form of a softer, more realistic style whose general characteristics survived until the middle of the 13th century. In France the style is particularly noticeable in a series of magnificent Bibles Moralisées (books of excerpts from the Bible accompanied by moral or allegorical interpretations and illustrated with scenes arranged in eight paired roundels, resembling stained glass windows) done probably for the French court *c.* 1230–40. In England the new style appears in numerous manuscripts—for instance, the psalter done for Westminster Abbey (British Museum, London; Royal MS. 2a XXII) and the Amesbury Psalter (*c.* 1240; All Souls College, Oxford). A particularly individual application of it is found in the manuscripts attributed to the chronicler Matthew Paris and in a series of illustrated manuscripts of the Apocalypse (Plate 6).

In Germany the graceful pictorial style did not become popular. Instead the successor to the Byzantine conventions of the 12th century was an extraordinarily twisted and angular style called the *Zackenstil.* In the Soest altar (*c.* 1230–40; now in the Staatliche Museen Preussischer Kulturbesitz, Berlin), for example, the drapery is shaped into abrupt angular forms and often falls to a sharp point, like an icicle.

High Gothic. Certain characteristics of high Gothic sculpture spread to influence painting about 1250–60. Probably the first place where this became evident was Paris, where Louis IX (St. Louis) was a leading patron (Plate 6). In an evangelary (a book containing the four Gospels) prepared for use at the Sainte-Chapelle (Louis IX's palace chapel), one can see the early Gothic pictorial style superseded quite abruptly by a drapery style incorporating the large, rather angular folds of the Joseph Master (Bibliothèque Nationale). Combined with this style was a growing emphasis on minute detail almost as an end in itself; faces, in particular, became tiny essays in virtuoso penmanship.

Although details such as faces and hands continued to be described chiefly by means of line, in a subsequent development drapery and other shapes were modeled in terms of light and shade. This "discovery of light," partial and piecemeal as it was, began around 1270–80 but is particularly associated with a well-known Parisian royal illuminator called Master Honoré, who was active about 1288–1300 or later.

It is possible that this new use of light was stimulated

New use of light

by developments in Italian painting. However that may be, Italian influence emerged quite clearly in the second quarter of the 14th century, in the workshop of the Parisian artist Jean Pucelle (Plate 6). More than a dozen books have been associated with this artist; most show an awareness of the recent Italian discovery of perspective in the portrayal of space and some an awareness of Italian iconography.

The French style was introduced fairly rapidly into England. Although Henry III apparently was not a bibliophile, various manuscripts executed for his immediate family contain echoes of the dainty and minute style of Louis IX's artists. Some large-scale paintings that demonstrate similar stylistic traits, notably the "Westminster Retable," survive in Westminster Abbey.

Subsequent changes in English painting involved greater decorative elaboration. A number of large psalters, such as the Queen Mary Psalter (in the British Museum), survive from the first half of the 14th century, many of them done for East Anglian patrons and almost all laying heavy emphasis on marginal decoration. Although some books with elaborate border decorations date from as early as the 13th century, such decorations became much more lavish in the 14th. There are occasional indications of Italian influence in figure poses and compositions but nothing really comparable to that found in books from Jean Pucelle's Parisian workshop.

Italian influence reached other European countries. An Italianate style of painting developed in Spain in the 14th century and, to a lesser extent, parts of German-speaking Europe—in Austria, for instance, paintings in the Italianate style were added around 1324–29 to make up the present Klosterneuburg altarpiece.

Italian Gothic. In the 13th century both Rome and Tuscany had flourishing pictorial traditions, and both, until the middle of the century, were strongly influenced by Byzantine art. The transitional period 1250–1300 is poorly documented. Since much of the Roman work was subsequently destroyed, evidence for what was happening in Rome must be sought outside the city. The most important location where such evidence exists is Assisi, where the upper church of St. Francis was decorated by Roman-trained fresco painters between about 1280 and 1300. In Tuscany the stylistic changes are probably best revealed by Duccio di Buoninsegna's "Maestà" (1308–11), formerly the high altarpiece of Siena cathedral.

As with all Gothic decorative art, the changes are in the direction of greater realism. By the end of the 13th century, painters in Rome, such as Pietro Cavallini and probably Duccio in Tuscany, had discovered, like their contemporaries in Paris, the use to which light could be put in figure modeling. The Italian painters also made sudden and unexpected advances in the manipulation of perspective to describe the space of the scenes they were painting. More than this, the best painters developed an extraordinary ability to create figures that really look as if they are communicating with each other by gesture and expression; the work of the Isaac Master in the upper church at Assisi is an especially good example.

How far the Italian tradition of painting on a large scale magnified problems such as perspective, it would be hard to say. The survival of a large-scale mural tradition certainly marks Italy off from the north. Italian mural paintings were executed with a technique involving pigment applied to, and absorbed by, lime plaster that was still fresh (hence the name of this type of painting—fresco). It was with work in this medium as much as in tempera (a substance binding powdered pigments, usually made from egg at this date) on panel that artists in Italy won their reputations. The typical subjects of fresco painting were series of biblical or hagiographic narratives. The painting of such fresco narratives (in Italian, *istorie,* hence "history painting") was to be regarded in the 15th century as the most important part of an artist's work by Leon Battista Alberti, an architect, painter, sculptor, and the founder of "modern" or "Renaissance" art theory. In making such claims, Alberti had in mind the work of the painter Giotto di Bondone, better known as simply Giotto, of the late 13th to early 14th century.

Giotto

Trained in Rome, Giotto executed his first important surviving work for the papal financier Enrico Scrovegni at the latter's family palace in Padua. The palace chapel, called the Arena Chapel (decorated *c.* 1305–13), is a masterpiece in which all the lessons of Roman mural painting were translated into a narrative sequence of great economy and expressiveness. In spite of the apparent realism of Giotto's work, however, the Byzantine past makes itself felt in the extremely strong sense of pattern and design noticeable throughout the compositions.

In Tuscany somewhat similar developments took place. Duccio's altarpiece, the "Maestà," contains a large number of small narrative scenes reminiscent of Giotto's fresco paintings. The figures, which have firmly modeled faces and expressive gestures, are arranged in buildings or landscapes that convincingly enclose them. Duccio's interest in realistic space, however, was much weaker than Giotto's. Although Duccio's scenes feature a variety of action and wealth of detail that, on the whole, is lacking in Giotto's early work, they do not make the same simple but dramatic impact.

These conflicts are inherent in all realistic painting. In Giotto's work a shift in the balance between the two conflicting elements takes place. He completed two chapels in Santa Croce, Florence (*c.* 1315–30), of which one, the Bardi Chapel, is smaller but not unlike the Arena Chapel. The other, the Peruzzi Chapel, tends toward greater detail and less stability in the settings.

Subsequent Florentine and Sienese painters also moved in this direction. Of the Sienese, Simone Martini was probably the most famous, since he worked outside Italy at the papal court in Avignon and was a friend of the great Italian poet Petrarch. His painting has strong suggestions of northern influence in its elegance and grace, but his care over detail is reminiscent of Duccio, and the careful structure of his setting recalls Giotto and the Roman painters. His major surviving work is now in Siena and Assisi, but some impressive remains have been recovered at Avignon (Plate 6).

The Lorenzetti brothers

Among other Tuscan painters were the brothers Pietro and Ambrogio Lorenzetti, who worked for almost their entire lives in that part of Italy. Their major works are in Siena, but, again, there are important frescoes at Assisi, where, probably, it was Pietro Lorenzetti and his workshop who decorated a transept in the lower church (*c.* 1330). Ambrogio Lorenzetti is especially famous for an enormous landscape, illustrating the effect of good government, painted in the Palazzo Pubblico Siena (1338–39). Historically, it is the first large, realistic landscape in which Byzantine conventions were entirely discarded. It had strangely few imitators, suggesting that the process of discarding convention and using the evidence of the eye is a slow one.

By the middle of the 14th century, Italian painters had achieved a unique position in Europe. They had made discoveries in the art of narrative composition that set them quite apart from painters anywhere else. Their achievements in capturing reality were not easily ignored. Many subsequent changes in northern painting consist of the adaptation of Italian compositional realism to northern purposes.

International Gothic. The style of European painting prevalent during the last half of the 14th century and the early years of the 15th is frequently called International Gothic. There were certainly at that time features common to European painting generally. In particular, figures were elegant and graceful, yet at the same time there was a certain artificiality about such figures, and a taste grew for realism in detail, general setting, and composition. The degree of internationalism about this phase of Gothic painting owes something to the fact that much of the most important work was executed under court patronage, and most European royal families were closely linked by marriage ties. Local idiosyncracies, however, persisted; seldom can the art of Paris, for example, be mistaken for that of Lombardy.

The main European courts were those of the Holy Roman emperors (who had nominal suzerainty over central Europe and who at this time had their capital at Prague),

the Visconti of Milan, the Valois of France, and the Plantagenets of England. But other sources of patronage existed—in Florence, for example, where the art of Lorenzo Ghiberti and Lorenzo Monaco merged with that of the early Renaissance. And an extraordinary number of important painters were associated about 1350–1400 with the linguistic area of Low Germany—the Low Countries and Westphalia especially—and the Rhineland.

The patronage of Charles IV

Under the Holy Roman emperor Charles IV and his son Wenceslas, Prague was the seat of a flourishing and enlightened court for about 60 years. Brought up in Paris, Charles had also traveled in Italy. Indeed, his main palace chapel at Karlštejn Castle near Prague, which is the chief monument to Charles's patronage, had an altarpiece by an Italian painter called Tommaso da Modena. The chapel itself was decorated chiefly by a local painter called Theodoric of Prague, whose work is Italianate. A group of his panel paintings, especially the altar of Vyšší Brod (*c.* 1350), shows a curiously Sienese character, though he did not achieve the delicacy associated with paintings from Siena. The emphasis instead is on heavily modeled faces and thick, heavy drapery. Theodoric's style seems to have initiated the "soft style" that remained a part of German painting well into the 15th century. He certainly determined the character of Bohemian panel painting up to the outbreak of the disastrous Hussite wars (1419).

Charles IV apparently did not collect manuscripts. His ministers and courtiers, however, stimulated an important school of manuscript painting, influenced by French and Italian styles but with distinctive decorative characteristics. Two of the more important manuscripts were a missal (a book containing the office of the mass) done for the chancellor Jan of Streda (*c.* 1360; Prague, National Museum Library, MS. XIII. A. 12) and a huge Bible begun for Charles's son Wenceslas (1390s; Vienna, Österreichische Nationalbibliothek, Cod. 2759–2764).

Styles similar to this Bohemian painting soon appeared elsewhere—the paintings of Master Bertram of Minden at Hamburg (*c.* 1380), for example.

In Paris a style appeared that had some of the characteristics of Bohemian work, especially a strong emphasis on faces and facial expression. An early example, probably executed before 1364, is a portrait of John II (Louvre, Paris), which is firmly modeled in a rather Italianate manner. More important, however, is the workshop of the master of the "Parement de Narbonne" (1370s; Louvre; Figure 15), an altar hanging (*parement*) found at the Cathedral of St. Justin Narbonne. These artists, who were active *c.* 1370–1410, worked in a very distinctive style: their figures, while graceful, have markedly heavy heads and expressive faces. That some interest in settings had developed is suggested by the care that must have been taken to render them reasonably three-dimensional. In this respect the works have much in common with earlier Italian painting.

An interest in the settings of paintings was shared by panel painters such as Melchior Broederlam, who executed the Dijon altar wings (1390s; Museum of Fine Arts, Dijon). The interest quickly spread during the early 15th century to the manuscript painters, who produced a series of extremely impressive landscape and architectural settings. Especially fine are the so-called Brussels Hours (Brussels, The Belgian National Library, MS. 11060–1) and the *Hours of the Maréchal de Boucicaut* (Jacquemart-André Museum, Paris). The best of the manuscript painters worked for the royal family, among whom Jean, duc de Berry, the brother of King Charles V of France, has achieved permanent fame as a patron. The most notable painters who enjoyed his patronage were Pol de Limburg and Pol's two brothers. Their illuminations are frequently reminiscent of contemporary Italian painting. The largest and most sumptuous work, the *Très Riches Heures du duc de Berry* (left unfinished in 1416, Condé Museum, Chantilly, Fr.), includes calendar pictures representing each month in terms of the seasonal activities of nobility and peasants (Plate 7). At least one Italian artist—identified tentatively as Zebo da Firenze—was painting in Paris at this period (*c.* 1405). Manuscripts associated with him are usually sumptuously, if erratically, decorated. In-

The *Très Riches Heures du duc de Berry*

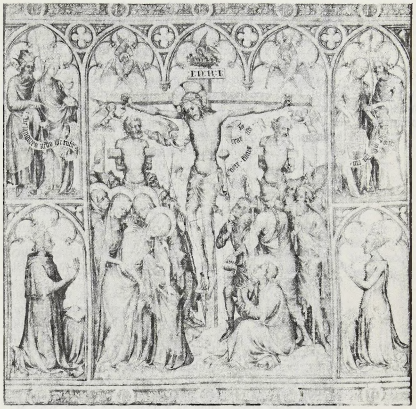


Figure 15: Detail from the "Parement de Narbonne," scenes of the Passion of Christ, with portraits of Charles V of France and his queen, brush drawing in grisaille on a white-silk hanging, 1370s. In the Louvre, Paris. 77.5 cm × 2.86 m.
J.E. Bulloz

deed, in the matter of erratic decoration they seem to have had a baleful influence. The border decoration of Parisian manuscripts *c.* 1410–25, such as those of the artist called the Master of the Duke of Bedford, often seems to run wild and to lack the restraint characteristic of Parisian painting up to this date.

The most eminent Italian artist of this period was perhaps Gentile da Fabriano. Trained probably in Venice, he painted there in the Doges' Palace (first decade of the 15th century) and also at Brescia. Subsequently he moved to Florence and thence to Rome, where he died. Most of his north Italian work has been destroyed, and his style must be assessed chiefly by the work done in Tuscany (Plate 7), the "Adoration of the Magi" altar (1423; Uffizi, Florence). His faces and drapery tend to have a soft, rounded modeling, somewhat reminiscent of the northern "soft style." The subject matter of his painting includes detailed studies of birds, animals, and flowers.

His style forms an interesting contrast to that of Lorenzo Monaco in Florence, who, though equally an International Gothic artist, tended to draw figures with finer, more incisive lines. In many ways Gentile's style resembles painting done at the Milanese court during this period. Many illustrated manuscripts survive, giving an impression of a transition about 1370–1410 from a strongly traditional Lombard style to something that has much in common with northern work. In particular, Michelino da Besozzo seems as court artist to have worked in a soft style similar to that of Gentile. Also dating from around 1400 is a distinguished group of illuminated manuscripts including the Book of Hours of Gian Galeazzo Visconti (Plate 7), herbals (manuals containing botanical drawings), and a famous sketchbook (*c.* 1395) containing a large number of drawings of animals (Bergamo, Municipal Library, A VII 14) from the workshop of an earlier court artist, Giovannino de' Grassi.

In England the decoration of the royal Chapel of St. Stephen's (*c.* 1360) was apparently, for the period, outstandingly Italianate. (Surviving fragments are in the British Museum, London.) Subsequently, however, in the Chapter House of Westminster Abbey (probably executed *c.* 1370) there was strong Germanic influence, which has been tentatively compared with the work of Master Bertram at Hamburg.

The court style of the second half of the 14th century is

best illustrated by a series of manuscripts done for members of the Bohun family and by a sumptuous missal given to Westminster Abbey by its abbot, Nicholas Litlington, in 1383–84. The work is decoratively lavish, but the figure style conveys only distant reflections of Italian painting.

Change in English manuscript painting

A great change in English manuscript painting occurred about 1400 and is associated with an artist named Herman Scheerre, who seems to have come from the region of Cologne. His figures have a rather plump softness that brings them into line with stylistic developments elsewhere; he also had a command of perspective and compositional structure lacking in the work of most previous artists in England. The style of John Siferwas, another painter active during this period, is similar, but his page decoration is usually more lavish; he produced a series of beautiful bird studies reminiscent of Lombard work. It should be noted, however, that this sort of realistic observation had long been a feature of English work—in the 14th-century East Anglian manuscripts, for example, and in English embroidery from about 1300.

In view of the number of good painters who came from the region of the Low Countries, Westphalia, and the Rhineland, it is puzzling that these areas should themselves have produced little important painting from the period about 1350–1410. Judging from the surviving works, easily the most distinguished of the painters active in this part of Europe was the Duke of Burgundy's painter, Melchior Broederlam, who lived and worked at Ypres. Other artists, such as Konrad von Soest, who executed the "Niederwildungen Altar" about 1403, seem to have reflected developments elsewhere without pioneering anything strikingly new. It was not until the 1420s that the Low Countries became the centre of intense pictorial development.

Late Gothic. The key to much 15th-century painting in northern Europe lies in the Low Countries. The influence of Paris and Dijon decreased, partly because of the renewal of the Hundred Years' War between England and France and partly because of the removal of the Burgundian court, after the mid-1420s, from Dijon to Brussels, which subsequently became the centre of an extensive court patronage.

The founder of the Flemish school of painting seems to have been Robert Campin of Tournai. The works of Campin, his pupil Rogier van der Weyden, and Jan van Eyck remained influential for the whole century. One of the most important discoveries of the period of about 1430—especially in the work of van Eyck—was the multifarious effects a painter can achieve by observing the action of light (Plate 8). These early Flemish artists found that light can define form, shape, and texture and that, when captured in a landscape, it can help convey a mood. Rogier van der Weyden also explored the problems of conveying emotion (Plate 8). A development in the rendering of the drapery—the so-called crumpled style of hard angular folds—is particularly clear in the paintings of Campin. Portraiture made dramatic progress during this period. Portraits were obviously not new; sculptors were already experimenting in the 14th century with life—and death—masks. But the brilliant use of lighting gives the portraits of Jan van Eyck, for instance, a vivid life hitherto quite unknown.

Use of light in Flemish painting

A great deal of later 15th- and 16th-century Flemish painting seems to play variations on these themes. Although there were painters with distinctly individual styles, such as Hugo van der Goes, with his highly accomplished technique and somewhat contemplative depictions, Hans Memling was more typical (despite having been born in the Rhineland).

The influence of van Eyck's paintings was felt to a limited extent outside the Low Countries—for example, by Konrad Witz of Basel, Switz., by the Master of the Aix Annunciation (1442) of Aix-en-Provence, Fr., and by the Neapolitan artist Colantonio and his illustrious pupil Antonello da Messina. In the course of the century, however, the style of Rogier van der Weyden and his immediate successors, such as Dirck Bouts, became more influential, being felt in Germany, England, Spain, and Portugal. Evidence of Rogier van der Weyden's influence can be seen in the works of Hans Pleydenwurff of Nürnberg, in the wall paintings in Eton College Chapel (*c.* 1480), and in the paintings of Nuno Gonçalves in Portugal. This new "international style" also influenced the great German engraver Martin Schongauer and, ultimately, the outstanding representative of the German Renaissance school of painting, Albrecht Dürer.

Any individualists at this time were usually painters who chose to go to the extreme of emphasizing the bizarre or the horrifying. Hugo van der Goes veered in this direction. Much more disquieting is the painting of Hieronymus Bosch, whose strange scenes still puzzle and perplex (Plate 8). The work of Matthias Grünewald, whose main surviving work is the altarpiece for a monastery at Isenheim, Ger. (Unterlinden Museum, Colmar, Fr.; Plate 8), is grotesque and horrifying.

Hiëronymus Bosch

A very different sort of extreme individuality is found in the work of the Tirolean painter and sculptor Michael Pacher. His pictorial work is so strongly marked by a concern with the structure of the composition and with effects of perspective—particularly foreshortening—that it seems clear he knew the work of Andrea Mantegna of Padua. Although virtually free of antique motifs, Pacher's painting demonstrates the growing fascination of Italian Renaissance art for northern artists.

Rather different were the French painters of the 15th century (Plate 7). Court art revived, especially during the reign of Louis XI (1461–83), as exemplified by the illuminated manuscript *Le Livre du coeur d'amours éspris* (1465; Österreichische Nationalbibliothek, Vienna). The most interesting painter was probably Jean Fouquet, who, apparently early in his career, visited Italy. Italian details certainly appear in his work, but, as is evident in the *Hours of Étienne Chevalier* (Condé Museum, Chantilly) and the "Melun Diptych" (now divided between the Staatliche Museen Preussischer Kulturbesitz, Berlin, and the Museum of Fine Arts, Antwerp), he still painted within the northern tradition. The restrained and somewhat reticent character of much French painting is interestingly similar to much of the sculpture. (A.Ma./Ed.)

Renaissance

The term Renaissance was first used by French art historians of the late 18th century in reference to the reappearance of antique architectural forms on Italian buildings of the early 16th century. The term was later expanded to include the whole of the 15th and 16th centuries and, by extension, to include sculpture, painting, and the decorative arts. There is still considerable disagreement among art historians as to whether the term should be restricted to a phenomenon that had its origins in Italy and then spread through western Europe (the point of view taken here) or whether directly contemporary developments north of the Alps, and especially in the Low Countries, should be included on an equal footing with what was happening in Italy.

The controversies that raged after the publication of Jacob Burckhardt's *Civilization of the Renaissance in Italy* (English translation, 1878) have abated, and the time span of the Renaissance is generally accepted as the period from roughly 1400 to about 1600, although certain geographical areas and certain art forms require greater latitude. This period is characterized as a rebirth or, better, the birth of attitudes and aims that have their closest parallel in the art of classical antiquity. Classical literature and, less often, classical painting were invoked as a justification for these new aims. The theoretical writings on art from the period indicate that man was the dominant theme. In religious painting, drama and emotion are expressed in human terms. From the late Middle Ages the theme of the Madonna enthroned with Christ Child is presented in an earthly setting peopled by mortals. This strongly humanistic trend serves to explain, at least in part, the development of portraiture as an independent genre and the ever-increasing number of profane, usually classical mythological, subjects in the art of the Renaissance. The painting of landscapes, as the earthly setting of man's activity, has its first modest beginnings in this period.

The role of art and of the artist began to take on modern form during the Renaissance. Leon Battista Alberti's *De pictura* (*Della pittura*), a treatise on the theory of painting, as opposed to the techniques of preparing and applying colours, appeared in Florence in 1435–36. The directions that art and art theory were to follow for the next 470 years are already present in this little book. The artist is considered to be a creator rather than a technician because he uses his intellect to measure, arrange, and harmonize the elements of his creation. The intellectual activity of art is demonstrated, by a series of comparisons, to be equivalent to that of the other liberal arts. Influences such as Alberti's book led to a new evaluation of the artist, with painters and their works being sought after by the rulers of Europe (Michelangelo and Titian were actually ennobled); the result was that great collections containing the works of major and minor masters were formed. At the same time the artist slowly began to free himself from the old guild system and to band together with his colleagues, first in religious confraternities and later in academies of art, which, in turn, were to lead to the modern art school. During the Renaissance, practitioners of all the arts evolved from anonymous craftsmen to individuals, often highly respected ones. Painting became more intellectual, sometimes to its own disadvantage, and changed from serving as a vehicle for didacticism or decoration to becoming a self-aware, self-assured form of expression.

For the sake of convenience, painting of the Renaissance is divided into three periods, although there is considerable overlap depending upon the painter and the place. The early Renaissance is reckoned to cover the period from about 1420 to 1495. The High Renaissance, or classic phase, is generally considered to extend from 1495 to 1520, the death of Raphael. The period of Mannerism and what has more recently been called late Renaissance painting is considered to extend from the 1520s to approximately 1600.

EARLY RENAISSANCE IN ITALY

The early Renaissance in Italy was essentially an experimental period characterized by the styles of individual artists rather than by any all-encompassing stylistic trend as in the High Renaissance or Mannerism. Early Renaissance painting in Italy had its birth and development in Florence, from which it spread to such centres as Urbino, Ferrara, Padua, Mantua, Venice, and Milan after the middle of the century.

The political and economic climate of the Italian Renaissance was often unstable; Florence, however, did at least provide an intellectual and cultural environment that was extremely propitious for the development of art. Although the direct impact of humanist literary studies upon 15th-century painting has generally been denied, three writers of the 15th century (Alberti, Filarete, and Enea Silvio Piccolomini, later Pope Pius II) drew parallels between the rebirth of classical learning and the rebirth of art. The literature of antiquity revealed that in earlier times both works of art and artists had been appreciated for their own intrinsic merits. Humanist studies also fostered a tendency, already apparent in Florentine painting as early as the time of Giotto, to see the world and everything in it in human terms. In the early 15th century Masaccio emphasized the human drama and emotions in his painting "The Expulsion" (Brancacci Chapel, Santa Maria del Carmine, Florence) rather than the theological implications of the act portrayed. Masaccio in his "Trinity" (Santa Maria Novella, Florence) and Fra Angelico in his San Marco altarpiece seem to be much more concerned with the human relations between the figures in the composition than with the purely devotional aspects of the subject. In the same way, the painter became more and more concerned with the relations between the work of art and the observer. This latter aspect of early 15th-century Florentine painting relies in great part on the invention of the one-point perspective system, which derives in turn from the new learning and the new vision of the world. The empirical system devised through mathematical studies by the architect Filippo Brunelleschi was given theoretical form and universal application by Alberti in *De pictura*.

Invention of a perspective system

In this system all parts of the painting bear a rational relation to each other and to the observer, for the observer's height and the distance he is to stand from the painting are controlled by the artist in laying out his perspective construction. By means of this system the microcosm of the painting and the real world of the observer become visually one, and the observer participates, as it were, in what he observes. To heighten the illusion of a painting as a window on the world, the Italian artists of the early 15th century turned to a study of the effects of light in nature and how to represent them in a painting, a study of the anatomy and proportions of man, and a careful observation of the world about them. It is primarily these characteristics that separate early Renaissance painting from late medieval painting in Italy.

Father of Renaissance painting

Masaccio has rightly been called the father of Renaissance painting, for every major artist of the 15th and 16th centuries in Florence began his career by studying Masaccio's murals in fresco. Masaccio is the artistic heir of Giotto, yet there is no indication of direct borrowing from the older master. He was also a friend of Brunelleschi and from him may have learned perspective and the concept of a clear and rationally articulated space. He was a friend, too, of the Florentine sculptor Donatello and may have learned from him the effectiveness of simple drapery folds over a full and powerful figure. Whatever his artistic sources, Masaccio's extant work reveals a concern with large and simple figures clad in simple draperies. He was concerned with light and the way it gives the illusion of solidity to the painted figure. He created a deep and clearly articulated space in his paintings, and he was above all concerned with his actors as humans carrying out some purposeful human activity. The only extant work by Masaccio that can be clearly dated is the Pisa altarpiece of 1426 (the central panel depicting the Madonna enthroned with Christ Child and angels, now in the National Gallery, London, is the largest surviving section). Although Masaccio continued the medieval tradition of using a gold background, the architectural elements of the throne indicate his awareness of the influence of Roman antiquity on the architecture of his friend Brunelleschi. The Madonna is no longer an elegant queen of heaven but an earthly mother with a human child on her lap. The figure of the Christ Child is a clear demonstration for future generations of the way light and shade can be manipulated in a painting to give the illusion of a solid three-dimensional body. In this painting Masaccio laid the foundations for one major current in all Florentine painting. His concern with the sculpturally conceived figure, bathed in light and presented in a strong and simple manner, created a work of quiet dignity and great monumentality in that it appears to be larger than it really is.

Masaccio's great fresco series in the Brancacci Chapel of Santa Maria del Carmine in Florence adds yet another dimension to early Renaissance painting. In this narrative sequence devoted to the life of St. Peter, he chose the most important moment in the narrative and then emphasized the drama by the human reactions to it. "The Tribute Money" is a simple yet powerful illustration of Christ's words, in which each apostle reacts individually to the tax collector's claim and Christ's reply (Plate 9). In this same chapel Masaccio also demonstrated his awareness of the real world, for the light of the paintings, indicated by the cast shadows, is the same as the natural light falling into the chapel.

Masaccio's awareness of the real world

"The Trinity" in Santa Maria Novella, Florence, provides a summary of Masaccio's brief career and indeed of the aesthetic principles of early Renaissance painting generally. The simple sculptural figures have acquired even greater dignity. The drama and emotion are presented in touching human terms as the Madonna turns to the observer to point out her crucified Son. In addition to the use of light to unite the space of the painting with the space of the observer, Masaccio also employed what appears to be the earliest practical example of the one-point perspective system, later to be formulated in words by Alberti. All the highest aims of early Renaissance painting are here: simplicity, strength, monumentality; man as observer, as actor, and as participant in the work of art.

Masaccio had no true followers or successors of equal stature, though there was a group of other Florentine painters who were about the same age as Masaccio and who followed in his footsteps to a greater or lesser degree: Fra Filippo Lippi, Fra Angelico, and Paolo Uccello.

Fra Filippo Lippi was a Carmelite monk who spent his youth and early manhood at Santa Maria del Carmine, where Masaccio's work was daily before his eyes. His earliest datable work, the "Madonna and Child" (1437) from Tarquinia Corneto, relies on the Madonna from the Pisa altarpiece, but in his Christ Child Fra Filippo already reveals an earthiness and sweetness unlike anything by Masaccio. "The Madonna and Child with Two Angels" (Uffizi, Florence)—with its urchin-angels, lumpy Christ Child, and elegant Madonna—is perhaps one of his best-known late works; the placement of the Madonna before an open window is one of the key sources for later Renaissance portraiture, including Leonardo da Vinci's "Mona Lisa," while the elegance and sweetness of the Madonna were to have their greatest reflection in the work of Fra Filippo Lippi's student, Botticelli.

Born about the same time as Masaccio, Fra Angelico was a Dominican monk who lived at Fiesole (just outside Florence) and at San Marco in Florence. His earliest documented work, the "Linaiuoli Altarpiece" (Museum of San Marco, Florence) of 1433, continues much that is traditional to medieval art, although the male saints in the wings (side pieces of a composite painting, typically a tripartite altarpiece) already reveal the influence of Masaccio. The altarpiece that he executed between 1438 and 1440 for the high altar of San Marco is one of the landmarks of early Renaissance art. It is the first appearance in Florence of the *sacra conversazione,* a composition in which angels, saints, and sometimes donors occupy the same space as the Madonna and Christ Child and in which the figures seem to be engaged in conversation. In addition to inaugurating a new phase of religious painting, the altarpiece reveals the influence of Masaccio in the sculptural treatment of the figures and an accurate awareness of the perspective theories of painting expressed by Alberti in his treatise. At about the same date, Fra Angelico was commissioned to decorate the monks' cells in San Marco. The nature of the commission—traditional devotional images whose execution required assistants—apparently turned Fra Angelico toward the religious and didactic works that characterize the end of his career; *e.g.,* the Chapel of Nicholas V in the Vatican.

Introduction of the *sacra conversazione*

Paolo Uccello's reputation as a practitioner of perspective is such that his truly remarkable gifts as a decorator tend to be overlooked. Studies of his extant works suggest that he was more interested in medieval optics than in the rational perspective system of Alberti and Brunelleschi. His earliest documented work, the "Sir John Hawkwood" fresco of 1436 in Florence cathedral, is a decorative work of a very high order and one that respects the integrity of the wall to which it is attached. Uccello is perhaps best known for the three panels depicting "The Battle of San Romano," executed about 1456 for the Medici Palace (now in the National Gallery, London; the Louvre, Paris; and the Uffizi). The paintings were designed as wall decoration and as such resemble tapestries: Uccello is concerned only with creating a small boxlike space for the action, for he closes off the background with a tapestry-like interweaving of men and animals. His primary concern is with the rhythmic disposition of the elements of the composition across the surface, an emphasis that he reinforces with the repetition of arcs and circles. Uccello's concern with the decorative and linear properties of painting had a great impact on the cassone (chest) painters of Florence and found its greatest reflection and refinement in the work of Botticelli.

Masaccio's greatest impact can be seen in the works of three younger painters, Andrea del Castagno, Domenico Veneziano, and Piero della Francesca. Castagno was the leader of the group. His "Last Supper" of about 1445, in the former convent of Sant'Apollonia in Florence, reveals the influence of Masaccio in the sculptural treatment of the figures, the painter's concern with light, and his desire to create a credible and rationally conceived space. At the same time Castagno betrays an almost pedantic interest in antiquity, which roughly parallels a similar development in letters, by the use of fictive marble panels on the rear wall and of sphinxes for the bench ends, both of which are direct copies of Roman prototypes. In the last years of his life, Castagno's style changed abruptly; he adopted a highly expressive emotionalism that paralleled a similar development in the work of his contemporaries. His "The Trinity with Saints" in the church of the Santissima Annunziata, Florence, was originally planned with calm and balanced figures, as the underpainting reveals. In the final painting, however, the figures, though sculpturally conceived, project an agitation heightened by the emaciated figure of St. Jerome and the radically conceived figure of the crucified Christ. The optimism, rationality, and calm human drama of earlier Renaissance painting in Florence were beginning to give way to a more personal, expressive, and linear style.

Expressive emotionalism among Florentine painters

One aspect of this new direction is met in the work of the enigmatic Domenico Veneziano, the second of the three principal painters who looked to Masaccio. His name indicates that he was a Venetian, and it is known that he arrived in Florence about 1438. He was associated with Castagno, and perhaps Fra Angelico, and helped to train the somewhat younger Piero della Francesca. His St. Lucy altarpiece of about 1445–50 (Uffizi) is an example of the *sacra conversazione* genre and contains references to the painting of Masaccio and the early 15th-century sculpture of the Florentine Nanni di Banco. The colour, however, is Domenico's own and has no relation to the Florentine tradition. His juxtaposition of pinks and light greens and his generally blond tonality point rather to his Venetian origins. In the painting he has lowered the vanishing point in order to make the figures appear to tower over the observer, with the result that the monumentality of the painting is enhanced at the expense of the observer's sense of participating in the painting.

Domenico Veneziano's unique use of colour

Piero della Francesca received his early training in Florence but spent the active part of his career outside the city in such centres as Urbino, Arezzo, Rimini, and his native Borgo San Sepolcro, in Umbria. His "Flagellation of Christ" (late 1450s), in the National Gallery of the Marches, Urbino, is a summary of early 15th-century interest in mathematics, perspective, and proportion (Plate 9). The calm sculptural figures are placed in clear, rational space and bathed in a cool light. This gives them a monumental dignity that can only be compared to early 5th-century-BC Greek sculpture. Much the same tendency can be seen in Piero's great fresco cycle in the church of San Francesco in Arezzo.

A hiatus occurred in Florentine painting around 1465–75. All the older artists had died, and the men who were to dominate the second half of the century were too young to have had prolonged contact with them. Three of these younger artists, Antonio Pollaiuolo, Sandro Botticelli, and Andrea del Verrocchio, began their careers as goldsmiths, which perhaps explains the linear emphasis and sense of movement noticeable in Florentine painting of the later 15th century.

As well as being a goldsmith, Antonio Pollaiuolo was a painter, sculptor, engraver, and architect. His work indicates his fascination with muscles in action, and he is said to have been the first artist to dissect the human body. In the altarpiece "The Martyrdom of St. Sebastian" (1475; National Gallery, London) he presents the archers from two points of view to demonstrate their muscular activity. His painting (formerly in the Uffizi but now lost) and small sculpture (Bargello, Florence) of "Hercules and Antaeus," like the engraving of "The Battle of the Nudes," depict struggle and violent action. "The Rape of Deianira" (Yale University Art Gallery, New Haven, Conn.) emphasizes yet another new element in Florentine painting, the landscape setting, in this case a lovely portrait of the Arno Valley with the city of Florence in the background.

Pollaiuolo's concern with anatomy

A similar concern with moving figures, a sense of movement across the surface of the panel, and landscape is found in the earlier works of Sandro Botticelli. In his well-known painting "The Primavera" (Uffizi) he uses line in depicting hair, flowing draperies, or the contour of an arm

Elegance and grace of Botticelli's works

to suggest the movement of the figures (Plate 9). At the same time the pose and gesture of the figures set up a rising and falling linear movement across the surface of the painting. Botticelli's well-known paintings of the Madonna and Child reveal a sweetness that he may have learned from Fra Filippo Lippi, together with his own sense of elegance and grace. A certain nervosity and pessimistic introspection inherent in Botticelli's early works broke forth about 1490. His "Mystic Nativity" of 1501 (National Gallery, London) is even, in one sense, a denial of all that the Renaissance stood for. The ambiguities of space and proportion are directed toward the unprecedented creation of a highly personal and emotionally charged statement.

Florentine painters active in the closing decades of the 15th century include Andrea del Verrocchio, who is best known as the master of Leonardo da Vinci and Perugino. There was also Filippino Lippi, who was apparently apprenticed to Botticelli when his father, Fra Filippo Lippi, died; he painted a group of madonnas that are easily confused with Botticelli's early work. By 1485, however, he had developed a somewhat nervous and agitated style that can be seen in the highly expressive "Vision of St. Bernard" in the Badia, Florence. His last works, such as the series of frescoes he painted in Santa Maria Novella (1502), reveal a use of colour and distortion of form that may have influenced the later development of Mannerism in Florence a generation or so later. Another painter active at this time was Domenico Ghirlandajo, whose artistic career was spent as a reporter of the Florentine scene. The series of frescoes on the "Life of the Virgin" in Santa Maria Novella (finished 1490) can be viewed as the life of a young Florentine girl as well as a religious painting. His art was already old-fashioned in his own time, but he provided a large number of Florentine artists, among them Michelangelo, with training in the difficult art of fresco painting.

Diffusion of the innovations of the Florentine school

The discoveries and innovations of the early 15th century in Florence began to diffuse to other artistic centres by mid-century. Siena painters in general continued the traditions of the 14th century except for such artists as Matteo di Giovanni, Neroccio di Bartolomeo, and Vecchietta, who alone in that city were to a certain degree under Florentine influence. In Ferrara, Cosimo Tura, Francesco del Cossa, and Ercole de' Roberti felt the influence of Florence as transmitted by Piero della Francesca. Only in Padua and Venice, however, did painters arise who could actually challenge the preeminence of Florence.

Andrea Mantegna was influenced by the sculpture executed by Donatello in Padua, the art of antiquity around him, and the teaching of his master, Francesco Squarcione. The frescoes he completed in 1455 in the Ovetari Chapel of the Eremitani Church in Padua (destroyed in World War II) grew out of the traditions of Florence, traditions to which Mantegna gave his own special stamp, however. His space is like that devised by the Florentines except that he lowers the horizon line to give his figures greater monumentality. His sculptural and often stony figures descend from Donatello and from ancient Roman models. His use of decorative details from antiquity reveals the almost archaeological training that he had received from Squarcione. By 1460 Mantegna had moved to Mantua, where he became court painter for the Gonzaga family, executing a number of family portraits (Plate 9) and pictures depicting ancient myths. His altarpieces, interpretation of antiquity, and engravings made him preeminent in northern Italy and a strong influence on his contemporaries and successors.

The Bellini family

The Bellini family of Venice forms one of the great dynasties in painting. The father, Jacopo, who had been a student of Gentile da Fabriano, adopted a style that owed something to both that prevailing in the Low Countries and that in Italy; he also compiled an important sketchbook (British Museum; Louvre). A daughter of Jacopo's was married to Mantegna, and the two sons—Gentile and, more especially, Giovanni Bellini—dominated Venetian painting until the first decade of the 16th century. Gentile followed more closely in his father's footsteps and is perhaps best known for his portraits of doges and sultans of Constantinople and his large paintings of Venetian religious processions. Giovanni early fell under the influence of Mantegna. The paintings each executed of "The Agony in the Garden" (both in the National Gallery, London) indicate how close they were stylistically and also their common reliance on Jacopo Bellini's sketchbook. At an unknown point in his career, Giovanni was in addition introduced to Flemish painting. These different influences permitted him about 1480 to evolve a highly personal style that greatly influenced the work of subsequent Venetian painters. This style consists above all of a softly diffused Venetian light that can only be achieved in an oil medium. Giovanni's work in the traditional medium for painting on panels—egg tempera—retains the crispness of contour and tightness of composition that the medium seems to require. The oil paintings, however, emphasize by their use of light the textures of the objects represented, softening the outlines and creating an elegiac mood. The "Madonna and Child with Saints" of 1488, in Santa Maria dei Frari, Venice, derived its composition from the Florentine *sacra conversazione* and two earlier altarpieces by Mantegna in which the Madonna and attendant saints are located in a unified but compartmentalized architectural setting. Giovanni's greatest innovation is the way in which the soft light suffuses the entire space, an effect particularly remarkable where it strikes the golden half dome of the apse and the ample draperies of the figures, which seem almost palpable. The "Enthroned Madonna from San Giobbe" (Gallerie dell'Accademia di Venezia) of about the same date goes even further in defining a composition and a way of painting that endured in Venice for centuries. The painting of "St. Francis in Ecstasy" (*c.* 1480; Frick Collection, New York City) adds yet another dimension to Giovanni's art. The observer's eye tends to wander from the saint and his cell into the distant landscape, for Giovanni was one of the greatest 15th-century masters of landscape painting. Figures, animals, trees, and buildings provide a series of guideposts leading the eye back into space. Giovanni influenced several Venetian painters: Lorenzo Lotto and Vittore Carpaccio and also, more importantly, Giorgione and Titian.

The richness, the variety, and even the inherent contradictions of 15th-century Florentine painting are both embodied and transformed in the art and the person of the multifaceted genius Leonardo da Vinci. Although he devoted a great deal of his career to a theoretical treatise on the art of painting, he was above all interested in the appearance of things and in the way they operated. This curiosity led him to a study of the flight of birds, the movement of water, the features of the land, the mechanical advantage obtainable in gears and gear trains, the growth of plants, the anatomy of man, and many other things. His consummate skill as a draftsman made it possible for him to record these discoveries as no man before him had done. All the knowledge that he gained was directed toward enriching his art, for Leonardo thought of himself primarily as a painter.

Leonardo's interest in the appearance and operation of nature

As a youth Leonardo was apprenticed to Verrocchio, in whose shop he learned to draw, prepare and mix colours, and paint. He probably also learned how to model in wax and clay and how to cast bronze. He may even have been introduced to the art of sculpting in marble, although he clearly stated in his writings that he did not relish this difficult craft. Leonardo's genius is already apparent in his collaboration with Verrocchio in the "Baptism of Christ" (*c.* 1474–75; Uffizi), in which his contributions to the landscape and his figure of an angel clearly reveal his superiority. The unfinished "Adoration of the Magi" (Uffizi) for the monastery of San Donato a Scopeto outside the walls of Florence, together with the preparatory drawings (Louvre; Uffizi), is at once a summary of 15th-century Florentine painting and a forecast of the High Renaissance style. In his studies Leonardo reveals his debt to Pollaiuolo and Botticelli and his awareness of the rational compositions of the first half of the century, with their strong sense of perspective. In the painting, however, he makes a synthesis of these divergent tendencies and creates a composition that is at once ordered and free, calm and full of movement, simple and varied. Pose, gesture, and glance in the attendant figures create a movement leading

toward and coming to rest in the figure of the Madonna and Christ Child. The figures are placed in a free yet ordered space that gives a sense of grandeur and expansion, the expansion in turn being balanced by the concentration of the group around the Madonna. The appearance of the finished painting and the final direction Leonardo's planning would have taken can only be guessed, yet the nature of the composition and the preparatory underpainting of the figures and landscapes clearly demonstrate that Leonardo had advanced so far beyond his contemporaries that his innovation would only be comprehended by the group of younger painters who emerged some 20 to 25 years later. (J.R.Sp./Ed.)

In 1481 Leonardo wrote a famous letter to the Duke of Milan offering his services. The offer was accepted, and for the next 18 years he remained in Milan, where he executed a number of paintings and innumerable drawings, worked on a never-completed equestrian monument to the Sforza dynasty, planned additions to the canal systems of the city, designed costumes for ducal entertainments, and wrote extensively. "The Virgin of the Rocks" (Louvre), painted in Milan about 1483, stands at the threshold of the High Renaissance. In this painting Leonardo introduced the pyramidal composition that was to become a hallmark of the High Renaissance. The placement of the Madonna, the Christ Child, the young St. John the Baptist, and the angel creates a movement that the eye willingly follows, yet the movement is contained within the implied pyramid, giving a sense of stability and calm grandeur to the composition. The mysterious landscape that surrounds them implies adequate space in which the figures can exist and move and an extension into depth that the eye cannot follow. The light that falls on the figures delicately models them in a subtle juxtaposition of light and shade. The contours of the figures seem to dissolve into the background, and the light seems to flow gently over a surface. The subtle and delicate modeling and the suggestive smoky atmosphere are known as sfumato and were much imitated, but what was more important and eventually more influential was Leonardo's use of light and shade as a unifying compositional factor. This was unprecedented in painting. It was achieved by the tonal continuity of the shadows—a tonal continuity conditional upon a severe restriction of local colour.

Introduction of the pyramidal composition

These effects, as well as the softly diffused light characteristic of Venetian painting, were only possible in the oil medium, which, because of its lengthy drying period, enables all parts of a painting to be advanced and adjusted together and the transparent glazes of which make possible unity of atmosphere and chiaroscuro. The rich effects of impasto (deliberately rough and thick paint textures) were also made possible in oil and were particularly exploited in Venice, where the use of canvas as a support first became truly popular. But there is no doubt that oil painting is a technique that originated in the Low Countries.

(J.R.Sp./N.B.P.)

"The Last Supper"

Leonardo's attempts to transfer this new concept of painting to the difficult genre of murals led to the triumph and the tragedy of "The Last Supper." Because the traditional technique of fresco painting was too final for Leonardo's method of working, he invented a new technique—still not fully understood—that permitted him to revise in the manner of oil painting. The technique was not permanent, and the painting began to deteriorate in Leonardo's own lifetime. Despite its deterioration the painting stands as one of man's greatest achievements. All elements of the painting lead the eye to the calm and pyramidal figure of Christ. The room is depicted according to the rules of perspective, with all the direction implied by the lines of the architecture meeting at the vanishing point in the head of Christ. In this painting Leonardo has combined the sense of drama of the groups of disturbed apostles, the sculptural figure of Christ, and the rationally constructed space of the first half of the 15th century with the movement and emotion of the second half, achieving a new synthesis that goes far beyond anything his predecessors had dreamed was possible. Leonardo's "The Last Supper" marks the actual beginning of the High Renaissance in Italy (Plate 10).

HIGH RENAISSANCE IN ITALY

In painting, the style called High Renaissance or classic is, in a sense, the culmination of the experiments of the 15th century, for it is above all characterized by a desire to achieve harmony and balance. Movement is important and necessary, yet the eye is always given a point of focus and rest. The composition is self-contained and conforms to Alberti's definition of beauty as "that harmony of parts to which nothing can be added or taken away without destroying the whole." Although there is movement implied in the poses of the figures and movement across the surface of the composition, it is always dignified movement, giving the impression of calm. The style exhibits variety and richness, yet maintains simplicity and unity. It is never as self-conscious as 15th-century painting had been, nor is it as laboured as much of Mannerist painting. It is frequently compared to Greek art of the 5th century BC for its calm and monumentality. Its greatest practitioners were the Florentines Leonardo da Vinci (although Leonardo's earlier work is usually assigned to the early Renaissance) and Michelangelo, the Urbino-born Raphael, and the Venetian Titian. Other artists, such as Andrea del Sarto and Fra Bartolomeo in Florence, Correggio in Parma, and Giovanni Bellini and Giorgione in Venice, were more or less attracted to the style at some point in their careers.

Alberti's definition of beauty

The new style of painting that Leonardo had invented in Milan was continued with modifications by Bernardino Luini and others. It had no immediate repercussions in his native Florence, although the example of his unfinished "Adoration" remained there. It is true that Botticelli, Ghirlandajo, and Filippino Lippi borrowed the broad outlines of the composition, but they did not penetrate completely to the innovational features inherent in it. The full impact of Leonardo's art was felt only upon his return to Florence in 1500. Crowds flocked to the church of the Santissima Annunziata to see his cartoon (a full-scale study for a finished painting) of "The Virgin and Child with St. Anne." Leonardo's great mural of the "Battle of Anghiari" (1503–06) pitted him against his rival Michelangelo in a competition to record the history of the city in the seat of city government. Neither painting was finished. Yet, despite the inconclusive nature of the works partially executed during his brief Florentine stay, Leonardo left a deep impression on that city. The "Mona Lisa" (Louvre) revolutionized portrait painting. Leonardo's drawings encouraged fellow artists to make more and freer studies for their paintings and encouraged connoisseurs to collect those drawings. Through the drawings his Milanese works were made known to the Florentines. Finally, his reputation and stature as an artist and thinker spread to his fellow artists and assured for them a freedom of action and thought similar to his own.

The painter who benefited most from the example of Leonardo was undoubtedly Raphael. Born Raphael Sanzio, he was as a youth under the influence of Perugino. He was already a successful and respected artist when, at the age of 21, he came to Florence only to discover that all he had learned and practiced was old-fashioned and provincial. He immediately set about learning from the Florentines. His drawing style changed from the tight contours and interior hatching he had learned from Perugino toward the freer, more flowing style of Leonardo. From Leonardo's "Virgin of the Rocks" he evolved a new Madonna type seated in a soft and gentle landscape, such as "The Madonna of the Goldfinch" in the Uffizi or those in the Louvre and the Kunsthistorisches Museum, Vienna (Plate 10). He adopted the "Mona Lisa" format for his portraits, and he also studied closely the sculpture of Michelangelo. By 1509, when he departed for Rome, Raphael had assimilated all Florence had to offer and was ready to make his own unique statement.

Evolution of the Raphael type of Madonna

The Stanza della Segnatura (the first of a series of rooms in the Vatican constituting Pope Julius II's apartments), particularly the "School of Athens," which Raphael painted between 1508 and 1511, is one of the clearest and finest examples of the High Renaissance style. In the "School of Athens" Raphael, like Leonardo before him, made a balance between the movement of the figures and the ordered

and stable space. He peopled this space with figures in a rich variety of poses yet controlled poses and gestures to make one group lead to the next in an interweaving and interlocking pattern, bringing the eye to the central figures of Plato and Aristotle at the converging point of the perspective construction. The unity, variety, and harmony of High Renaissance felicitously combine in the frescoes that decorate the Stanza della Segnatura.

The painting of the Sistine Chapel ceiling

At about the same time Raphael was working in the papal apartments in the Vatican, Michelangelo had undertaken the formidable task of decorating the ceiling of the Sistine Chapel (1508–12), also for Pope Julius II. In 1481–82 under Pope Sixtus IV, the uncle of Julius, the chapel had been completed and the walls decorated with frescoes depicting scenes from the life of Moses and the life of Christ executed by Botticelli, Ghirlandajo, Perugino, and others. Against his will Michelangelo was assigned to paint in fresco scenes from the creation. Although he had been trained in fresco painting in the shop of Ghirlandajo and although he had already executed a few paintings of considerable power, Michelangelo thought of himself as a sculptor. He engaged a group of his former colleagues from the shop of Ghirlandajo and with them began to paint the "Drunkenness of Noah" above the entrance to the chapel. Michelangelo had little patience with his less gifted associates, dismissed them, and executed the entire ceiling alone. The scenes were painted in reverse chronological order, beginning with the "Drunkenness of Noah" over the door and ending with the act of creation over the altar. In the first three frescoes Michelangelo seems to be feeling his way. With the second three—"Temptation and Expulsion," "Creation of Eve," "Creation of Adam"—he returns to the models of his youth (Masaccio for the "Expulsion" and Jacopo della Quercia for the "Creation of Eve") to create a powerful High Renaissance composition. The balance between the kinetic energy of God the Creator with his whirlwind of figures around him and the flaccid lifeless form of Adam comes to a focus in the two hands and the significant void between them. In the final three scenes of creation, Michelangelo moves beyond his contemporaries to a highly personal statement without parallel in the art of the 16th century. The Sistine ceiling was recognized as a masterpiece in its own time. The artist was judged to be a superhuman being and earned the title "the divine Michelangelo." Contemporaries spoke of the *terribilità*, or awesome power, of the frescoes and their creator. Michelangelo, Leonardo, and Raphael raised the artist and his art to a position of esteem perhaps never enjoyed (or deserved) either before or since. Certainly their soaring levels of achievement made it difficult for succeeding artists to follow in their footsteps and impossible to surpass them. Hence, the "anticlassic style," as it has been called, emerged in their own lifetime—even in some of Raphael's late works—and provided one of the sources of Mannerism.

By 1513, when Julius II died and Leo X was elected pope, the three great painters of the Florentine-Roman High Renaissance style became involved in projects that diverted them from the paths they had hitherto been following. Leonardo had been since 1506 at the French court in Milan, where he continued to refine his portrait of "Mona Lisa" and where he devoted a great deal of time to his treatises and his projects for the French king. Michelangelo returned to the sculpture of the tomb of Julius II and in 1516 began to work in Florence on a number of architectural and sculptural projects for its Medici rulers. Raphael became more and more involved with purely administrative duties as architect in charge of the new St. Peter's and as surveyor of antiquities. The great number of commissions he received led him more and more to rely on talented assistants, such as Giulio Romano, and on the workers in the shop for the actual execution of his later works. In the few works in which Raphael's hand clearly appears, he seems to be moving away from the "School of Athens" toward a new style that had not completely developed at the time of his death in his 37th year.

Other projects of the High Renaissance

Raphael's frescoes (1512–14) in the Stanza d'Elidoro ("Heliodorus Room," the second of the rooms in Julius II's apartments) already reveal Mannerist tensions in "The Expulsion of Heliodorus from the Temple" and an almost Baroque concern with light and shade in the "Liberation of St. Peter." The succeeding rooms were decorated largely by assistants. It is only in such works as the "Triumph of Galatea" (1511; Villa Farnesina, Rome), the "Sistine Madonna" (1513?; Gallery of Old Masters, Staatliche Kunstsammlungen Dresden, Ger.), or Raphael's final unfinished work, the "Transfiguration" (1517; Vatican Museum), that he can be seen to move toward a nascent Mannerism and then away from it toward a more relaxed, more personal, and deeply moving reconsideration of High Renaissance ideals.

Michelangelo and Mannerism

When Michelangelo returned to painting in 1534, he had already had some experience of Mannerism in Florence. The past few years had not been entirely a matter of aesthetic enrichment, however, for they had witnessed the sack of Rome and the siege of Florence. Some of the horror of those events emerges in "The Last Judgment," painted in fresco on the altar wall of the Sistine Chapel (Plate 11). Commissioned by Pope Clement VII, the work was executed during the pontificate of Paul III. Rather than a Christian saviour, the Christ in Judgment is a thundering god in the way the pagan supreme god of the Romans, Jupiter, was, more concerned with damning the human race than welcoming the blessed into heaven. The figures falling into hell and the torments of the damned attract the eye by their power and the emotions they reveal. It is as though a new Dante had been born to depict hell with a brush rather than a pen. Though the fresco met with strong but mixed reactions when it was unveiled in 1541, Pope Paul was pleased enough to commission two frescoes representing the "Conversion of St. Paul" and the "Crucifixion of St. Peter" for his own private chapel, the Pauline Chapel. Since this chapel has never been open to the general public and since Michelangelo had already moved into his highly personal late style, these frescoes had little impact on the painting of the time.

The ideals of the High Renaissance as they appeared in the works of Leonardo, Raphael, and Michelangelo continued to develop independently in areas outside Rome and Florence. In Parma the painter Antonio Allegri, better known as Correggio, formed his art under the influence of Mantegna and Leonardo's Milanese followers. His "Rest on the Flight into Egypt" (Uffizi) and the "Madonna of the Bowl" (*c.* 1525; National Gallery, Parma) are clearly painted in the High Renaissance idiom. Yet Correggio is perhaps best known for his frescoes at Parma in the cathedral and in the church of San Giovanni Evangelista, which seem to prefigure the style of painting found in the Baroque, and for his late series of sensuous paintings on the loves of Jupiter, all executed between about 1530 and 1534, consisting of "Danae" (Borghese Gallery, Rome), "Leda" (Staatliche Museen Preussischer Kulturbesitz, Berlin), and "Jupiter and Io" and "The Rape of Ganymede" (Kunsthistorisches Museum, Vienna).

In the late 15th century, painting in Venice traveled much the same paths toward the High Renaissance as in Florence, while still maintaining a purely Venetian flavour. Giovanni Bellini's Madonnas of 1505–10 (Pinacoteca di Brera, Milan; and the National Gallery, London) are stylistically similar to the Madonnas that Raphael was painting in Florence at about the same time. The San Zaccaria altarpiece ("Enthroned Madonna with Four Saints") of 1505 carries the *sacra conversazione* fully into the High Renaissance. Inasmuch as Giovanni Bellini dominated Venetian painting, his style influenced the younger painters Giorgione and Titian, yet he was receptive enough to learn in turn from them and inventive enough to maintain his position of dominance.

Giorgione's treatment of landscape

Giorgione, having learned from Bellini, went beyond his master to bring to Venetian painting a treatment of landscape that can only be compared to pastoral poetry. In his brief career (all his extant paintings date from the last five years of his life) this highly inventive young artist taught his contemporaries and successors how to exploit the medium of oil paint to create the illusion of textures, light, and air in their paintings. His earliest known painting, the "Madonna and Child with SS. Francis and Lib-

erale" (*c.* 1504; Castelfranco cathedral, Italy), derives from the style of the mature work of Bellini. In only a few years Giorgione passed from this style of painting, through the turbulence of the dramatic landscape with storm clouds of "The Tempest" (*c.* 1505; Gallerie dell'Accademia di Venezia; Plate 10), to the dreamy landscape of the "Pastoral Concert" (*c.* 1510; Louvre). In the latter, Giorgione reveals the Venetians' love of textures, for he carefully renders almost palpable the appearance of flesh, fabric, wood, stone, and foliage. The soft, typically Venetian diffused light, together with the landscape, its hills stretching into the distance and all harsh or sharp contours removed (whether of landscape or figures), creates a gently pastoral mood. The use of landscape to create a mood and the use of figures in the landscape to reflect or intensify that mood is an innovation characteristic of Venetian painting of the 16th century and one of great importance to the development of Baroque art. (J.R.Sp./Ed.)

Whatever the significance of the figures in Giorgione's painting "The Tempest," it is certain that the picture soon came to be regarded merely as a landscape, and it is one of the first European paintings, certainly the first by a great artist, to make this sort of "background" material its subject matter. Landscape painting was to become a specialization of artists only toward the close of the 16th century, and even then chiefly in northern Europe. (N.B.P.)

The impact of Giorgione on Venetian art was immediate and direct. Bellini's last works, such as "The Feast of the Gods" (National Gallery of Art, Washington, D.C.), owe much to Giorgione, but Giorgione's greatest impact was on Titian (Tiziano Vecelli). Although Titian was never a student of Giorgione, he worked with him on one project and finished a number of his paintings. The so-called "Sacred and Profane Love" of 1512–15 (Borghese Gallery, Rome) is, in a sense, Titian's trial piece in which he shows himself capable of rivaling and surpassing Giorgione in Giorgione's own terms. The influence of Giorgione is especially marked in the profane paintings just as the religious paintings are marked by the influence of Bellini, Titian's teacher and rival until his death, when Titian himself emerged as the leader of Venetian painting.

Titian's reputation as Bellini's successor

Titian's great masterpiece, the "Assumption of the Virgin" (1516–18; Santa Maria dei Frari, Venice), established his reputation as Bellini's successor. In the painting he exhibits the Venetians' love of colour and texture, but he succeeds in achieving a balanced and moving composition that can only be compared to Raphael's "School of Athens" in its grandeur. The environment is both earth and heaven, yet it is created and defined by light and atmosphere in a typically Venetian way, rather than by architecture, as would have been more common in Florence. This painting, together with the "Sacred and Profane Love," "Entombment" (Louvre), and the "Pesaro Madonna" (1519–26; Santa Maria dei Frari), typifies Titian's contribution to the High Renaissance.

Upon the completion of the "Assumption" Titian undertook to execute a series of paintings on mythological themes for the court of Ferrara. "The Bacchanal" (1518–19; Prado, Madrid) was soon joined by the "Worship of Venus" (1518–19; Prado) and "Bacchus and Ariadne" (1520–23; National Gallery, London). In "The Bacchanal" Titian reveals his mastery in treating mythological subjects (Plate 10). The bacchants are disposed about the miraculous stream of wine that flows through an island, dancing, singing, and drinking. The movement of the figures, the juxtaposition of nude and clothed, of male and female, creates a revel in which even the landscape seems to participate—only a Venetian could have created such a pagan, earthy, and hedonistic glorification of life.

On a trip to Rome in 1545 Titian succeeded in rivaling Raphael's "Portrait of Leo X with Cardinals Giulio de' Medici and Luigi de' Rossi" (Uffizi) and Michelangelo's "Leda and the Swan" (copy; National Gallery, London), while demonstrating that Venetian painting in its own unique way was the equal of the Florentine-Roman tradition. The portrait of "Pope Paul III and His Grandsons Ottavio and Cardinal Alessandro Farnese" (1546; Museo e Gallerie Nazionali di Capodimonte, Naples) is a free variation on Raphael's "Leo X," but Titian altered the composition and introduced a narrative dimension more dramatic and compelling than that in any of his earlier portraits. The "Danae with Nursemaid" (1553–54; Prado), from the same period, poses the colourism and sensuousness of Venetian painting against the sculptural and restrained tradition of Michelangelo for succeeding centuries to judge.

Titian's use of the oil medium

In his late works Titian carried the oil medium to new heights. He used loosely juxtaposed patches of colour, sometimes allowing the prepared canvas to show through. He applied paint freely and loosely with the brush and sometimes reworked it with his fingers. Although his paintings have a fresh quality that makes them appear to have been painted quickly in the heat of inspiration, it is known from his biographers and friends that each work had been carefully studied, criticized, and reworked before the artist was satisfied.

Two works from this late period reveal the scope of Titian's genius. The "Martyrdom of St. Lawrence" (Church of the Jesuits, Venice) was begun about 1548 after his return from Rome and before a trip to Augsburg and the court of the Holy Roman emperor Charles V. Although Titian was nearly 60, he painted with all the enthusiasm of youth. There is a certain amount of Mannerist foreshortening and exaggeration, but Titian used these aspects of Mannerist vocabulary, together with the Venetians' skill in the use of light, to emphasize the dramatic and emotional content of the painting. Drama, light, and colour were used in the late "Rape of Europa" (*c.* 1559–62; Isabella Stewart Gardner Museum, Boston) in a way that prefigured the work of Rubens and the Baroque.

Titian's genius, given full rein in his long and productive career, deeply influenced Venetian painting. The two most outstanding painters of the end of the 16th century, Veronese and Tintoretto, each took a different aspect of Titian's style and developed it. Paolo Caliari, called Veronese (he was born in Verona), is best known for the rich colour and interweaving compositions he learned from Titian. His frescoes at the Villa Barbaro at Maser northwest of Venice are important for Venetian Mannerism and for landscape painting, but the richness of his palette is best seen in the mythologies, such as "Mars and Venus United by Love" (Metropolitan Museum of Art, New York City), or the "Marriage of St. Catherine" (Gallerie dell'Accademia di Venezia). With Tintoretto he decorated the chambers of the Doges' Palace in Venice, partially supplanting the aging and busy Titian as official painter of the city; his "Apotheosis of Venice" (*c.* 1585) in the Sala del Maggior Consiglio in the Doges' Palace is a bold and successful use of dramatic foreshortening and rich colour to express the vigour and vibrancy of Venice. Splendid also are his extremely large paintings crowded with figures, such as the "Feast in the House of Levi" (1573; Gallerie dell'Accademia di Venezia).

Veronese as a rich colourist

Jacopo Robusti, called Tintoretto, was most interested in Titian's use of dramatic light and heightened emotion. By 1548 he had established his reputation as a leading artist of the younger generation with his "San Marco Freeing the Slave" (Gallerie dell'Accademia di Venezia). He decorated several chambers of the Doges' Palace with a number of inventive mythological scenes. A great part of his career and energy was devoted to the decoration of the Great School of San Rocco, Venice (1564–*c.* 1588). Perhaps the crowning achievement of his career can be found in "The Last Supper" of 1594, painted for the church of San Giorgio Maggiore, Venice (Plate 11). In this painting Tintoretto made use of all the rapidly receding diagonals and dramatic foreshortenings of the Mannerist vocabulary, but he brought to the painting the Venetians' love of light effects to define the forms and heighten the drama. The head of Christ is bathed in light that is repeated in the smoky lamps so that its true source cannot be known. Light is used, as in the work of Titian, to pick out certain forms, throw others into darkness, and create a sense of movement within the composition. The comparison of this painting with Leonardo's "Last Supper" of 100 years earlier reveals the differences between the High Renaissance and the late 16th century.

Tintoretto's use of dramatic light

With the death of Titian, Tintoretto, and Veronese, Venice became a school for 17th-century painters, where great works of the past were studied but few great works were produced until the 18th century. The influence of 16th-century Venetian painting on such diverse Baroque artists as Annibale Carracci, Peter Paul Rubens, and Nicolas Poussin was not negligible, but they came to Venice only to learn; they made their major contributions in other centres. (J.R.Sp./Ed.)

ITALIAN MANNERISM AND LATE RENAISSANCE

The first reaction against Leonardo, Michelangelo, Raphael, and Andrea del Sarto occurred in Florence between 1515 and 1524, during which time the painters Giovanni Battista (called Rosso Fiorentino) and Jacopo Carucci Pontormo decisively broke away from the harmony and naturalism of the High Renaissance style. Their movement, particularly what might be called their aesthetic anarchy, attracted the sympathetic attention of some 20th-century art historians, largely because of affinities such art historians saw between their work and modern trends, particularly Expressionism. After the lead given by the German art historian Max Dvořák in his book *Über Greco und der Manierismus* (1921), these 16th-century nonconformists came to be known as Mannerists. Recent historians have suggested, however, that the term Mannerism can more accurately be applied to a very different style initiated in Rome about 1520. Roman Mannerism, which subsequently spread throughout Europe, is characterized by a display of the artificiality of art, a thoroughly self-conscious cultivation of elegance and facility, and a sophisticated delight in the bizarre.

The term Mannerism is ultimately derived from the Italian word *maniera* (literally "style"). It was in the 16th century that *maniera* was first consistently used in art criticism to indicate a definable quality—that of stylishness. Giorgio Vasari, who is known chiefly for his biographies of artists (some of whom were his contemporaries) but who was also an architect and painter, indeed a Mannerist himself, attributed this absolute quality of stylishness to Leonardo, Michelangelo, and Raphael, and, above all, to artists of his own day who had learned their styles from studying these great masters. Standing at the head of the enormous representational discoveries of the Renaissance and with an increased knowledge of antiquity, Vasari was convinced that his contemporaries were in a position to understand the secret of true artistic style. This was the *maniera.*

Taking Vasari's quality of *maniera* as the key to Mannerism, it is possible to outline some of its hallmarks. In figure style, the standard of formal complexity had been set by Michelangelo and that of idealized beauty by Raphael. In the art of their followers, obsession with style in figure composition often outweighed the importance of the subject matter. The highest value was placed upon the apparently effortless solution of considerable artistic problems, such as the portrayal of the nude figure in complex poses. Specifically, the finished work was not supposed to betray signs of the labour that lay behind it.

While depending heavily upon ancient Roman art for many of its decorative motifs and for many of its standards of design, Mannerist style commonly exploited a certain degree of license within the classical vocabulary—what Vasari and contemporary literary theorists called "a departure from the normal usage."

It was in the intellectualizing atmosphere of the Italian courts that Mannerism met with the greatest favour. There the conscious intricacies of Mannerist compositions and the eloquent quotations from antiquity were well appreciated; court literature of this period displayed many analogous features. Mannerism was first and foremost a connoisseur's art—certainly not one that appealed to a churchman. It is not surprising that the later Mannerist painters were censured by the church during the Counter-Reformation for painting altarpieces that were intended to demonstrate the virtuosity of their creators rather than illustrate a religious story. Even Michelangelo was attacked, one critic calling him "the inventor of obscenities, who cultivated art at the expense of devotion."

Factors such as these caused the style to fall into general disrepute, and, when in 1662 the French writer on architectural theory Fréart de Chambray coined the word *Maniériste* (translated six years later as "Mannerist" by the English diarist John Evelyn), he applied it in disparaging fashion to Vasari and his contemporaries, the practitioners of the *maniera.* If, therefore, Mannerism is identified with the *maniera,* it can be historically related to a particular 16th-century style; but if it is applied strictly to early Rosso and Pontormo, as it was by Dvořák, it has no firm grounding in the way people in the 16th century thought about painting.

During the second decade of the 16th century, Andrea del Sarto had emerged as the foremost practitioner of High Renaissance naturalism in Florence. The subtle and ambiguous emotional tension present beneath the harmony of Andrea's forms and colours was greatly accentuated by one of his pupils, Jacopo da Pontormo. In Pontormo's Visdomini altarpiece (1518), the tension approaches the breaking point; the composition is vertical and lacking in a sense of space; and a host of similar but clashing centres of action create an impression of agitation. Pontormo persisted with this expressive style, becoming increasingly influenced by the angular forms of Albrecht Dürer's German engravings and by the more tortured aspects of Michelangelo's figure style. Vasari made it quite clear that Pontormo's development was in direct contradiction to the later ideals of Mannerism.

The second of Andrea's important pupils, Rosso Fiorentino, began in a not dissimilar spirit of expressive rebellion. His highly unconventional "Madonna with SS. John the Baptist, Anthony Abbot, Jerome and Stephen" for Santa Maria Nuova (1518; Uffizi) displays an aesthetic anarchy bolder than anything by Pontormo, and by the 1520s he was creating works of savage emotionality (*e.g.,* the Volterra "Deposition," 1521). In 1523 Rosso journeyed to Rome. There he was overwhelmed by three experiences: Michelangelo's ceiling in the Sistine Chapel, the late style of Raphael, and the art of the newly arrived Parmigianino.

Parmigianino brought with him from Parma three sample pictures to display his virtuosity to Roman patrons. His style, based originally upon that of Correggio, already possessed much of the attenuated elegance for which he became famous (Plate 11). In Rome Parmigianino was hailed as the new Raphael and specifically as a painter capable of reproducing the sophisticated grace of Raphael's late "St. Michael" (Louvre). Raphael had died in 1520, but his most authoritative late work in the Vatican *stanze* (papal apartments) was continued and developed by his foremost pupils, Giulio Romano (who left for Mantua in 1524) and Perino del Vaga. Their Roman styles rely upon a direct though refined use of the art of classical antiquity as a source of inspiration and upon an ingenious exploitation of different levels of pictorial reality within a single decorative scheme. The underlying artificiality of their manner was reinforced by the latent academicism of Michelangelo's Sistine ceiling nudes.

Rosso's encounter with the latest painting Rome offered resulted in a radical realignment of his style. His "Dead Christ with Angels" (*c.* 1526; Museum of Fine Arts, Boston), a subject that he would earlier have been inclined to treat with exceptional angularity of form, is executed with a new feeling for rarefied beauty. Emotion is now expressed less overtly, and his handling of paint is less aggressive. Rosso, Parmigianino, and Raphael's pupils undoubtedly influenced each other during the mid-1520s, but in 1527 Rome was sacked, and the artists of Pope Clement VII's court became scattered. Parmigianino fled to Bologna, returning after four years to his native Parma, where he continued to develop his personal form of mannered beauty (*e.g.,* "Madonna of the Long Neck," Uffizi). Perino found employment with the ruling family at Genoa, and Rosso visited a number of Italian cities before settling in France.

Artistic dispersion resulting from the sack of Rome

The sophisticated Mannerism that evolved in Rome before 1527 became the chief formative influence upon the styles of a number of important younger artists. Vasari and Francesco Salviati had passed their period of ap-

prenticeship in Andrea del Sarto's Florence. They parted in 1527 but resumed their close acquaintance in Rome (1531), and it was the Roman style that influenced their subsequent development. Vasari, Salviati, and Jacopino del Conte, who worked with Salviati on the frescoes for San Giovanni Decollato, Rome, attempted to combine the formal and narrative artifice of the late Raphael decorations with the complex figure style of Michelangelo. The result in Vasari's case is undeniably eclectic, but Salviati created an individual *maniera* of enormous facility and inventiveness (*e.g.,* "Peace," Palazzo Vecchio, Florence). The Raphaelesque element in the Roman style was reinforced by Perino's return to Rome in about 1538.

Salviati's career was unsettled—he worked in Florence, Rome, Venice, and France—but Vasari returned to Florence to the court of the Medici duke Cosimo I, who had replaced in 1537 the unpopular (and assassinated) Alessandro de' Medici. Cosimo and his Spanish wife, Eleonora de Toledo, whose formal Iberian tastes influenced the artistic life of the court, shrewdly embarked on an ambitious series of propagandistic projects to consolidate his political position. Vasari became the "stage manager" for much Medicean propaganda. His success as a painter and architect after 1555 was considerable, but his most important contribution to Mannerism was undoubtedly his advocacy of Mannerist ideals in his *Lives,* first published in 1550 and revised and extended in 1568. As Vasari realized, the most important painter in Cosimo's court was Il Bronzino, a pupil of Pontormo.

Aloof formality of Bronzino's portraits

Bronzino had, from the first, reduced the emotional content that had been an important feature of Pontormo's style, and, during the 1530s in Florence, he began to establish a reputation as a court portrait painter. His mature portraits are elegant, perfectly finished, ingenious in detail, and aloofly formal, reflecting the Spanish etiquette of Cosimo's court. Bronzino was adopted as favourite artist by Eleonora, receiving the commission to decorate her small private chapel in the Palazzo Vecchio. The resulting frescoes are in no sense spiritually expressive, but they are brilliantly stylish, with references to antiquity, which displayed the erudition of both artist and sitter, and to Raphael and Michelangelo. Bronzino was later influenced by the teachings of the Counter-Reformation, adopting a more modest narrative style, but his underlying aesthetic art remained a sense of *maniera.*

A number of later Mannerists responded similarly to the Counter-Reformation—Santi di Tito is particularly important in this respect—but it was only with Federico Barocci that the ideals of Mannerism were abandoned in favour of an all-pervasive piety in religious painting. Barocci's attractively fluent and softly coloured style, based largely upon Correggio, may be considered as an exceptional precursor of the Baroque style. Barocci abandoned his Roman Mannerism as early as 1575, but the majority of his contemporaries in Rome and Florence continued to develop the eclectic aspects of the original *maniera.* Daniele da Volterra and Pellegrino Tibaldi painted in an explicitly Michelangelesque manner, while Cavaliere d'Arpino (Giuseppe Cesari) and Federico Zuccari, at the end of the century, investigated the complex intellectual conceits of the Raphael studio style. Zuccari—a painter, designer, and theorist—is the most representative figure of this late phase, and his travels (to Rome, Venice, Spain, England, France, and Antwerp) underline the internationalism of late Mannerist style.

Spreading influence of Mannerism

Outside Florence and Rome, many of the major Italian cities succumbed to the spreading influence of Mannerism after 1527. Siena, under the lead of Domenico Beccafumi, developed a bizarre form of emotional Mannerism, but only Venice maintained a steady, independent Mannerism. Venice was certainly receptive to Mannerist influence—as seen in the works of Titian after 1530, Tintoretto, and Veronese—but, with the exception of Andrea Meldolla (Schiavone), Venetian painting continued to be dominated by non-Mannerist ideas in colouring and expression. Vasari's disparaging remarks about Tintoretto's lack of good design show clearly that the differences between Romano-Florentine and Venetian painting remained fundamental.

The early and High Renaissance style as developed in Italy did not immediately dominate all European painting. A few northern artists adopted Renaissance motifs but used them in a piecemeal manner without full comprehension of Italian compositional methods. After 1520, however, northern and Spanish artists came increasingly to understand and adopt Mannerist ideas, and highly individual schools of Mannerism began to appear in various centres outside Italy. Regional styles of considerable decorative flamboyance resulted from the fusion of the intricacies of the late Gothic style with the complexities of Mannerism.

RENAISSANCE OUTSIDE ITALY

France. Francis I, despite his military reverses in Italy, was enamoured of all things Italian. He commissioned the celebrated goldsmith Benvenuto Cellini to execute both tableware and sculpture and prevailed upon friendly rulers in Italy to send him works by Titian and Bronzino and casts of sculpture. He also imported Italian artists to design, build, and decorate his palaces, the Château de Madrid and Fontainebleau, both outside Paris. Rosso arrived in France in 1530, followed two years later by his fellow Italian, the Mannerist Francesco Primaticcio. In the gallery of Francis I at Fontainebleau, Rosso initiated a new and intricate decorative system in which stucco and painting form a richly luxuriant complex—the plastic realization of the late Raphaelesque decorative manner. Primaticcio, who had been trained by Giulio Romano at Mantua and influenced by Parmigianino, took over Rosso's leading position on the latter's death in 1540. His Ulysses Gallery at Fontainebleau continued and refined Rosso's elaborate system of painted narratives surrounded by convoluted strapwork, elegant figures, and swags in stucco. French artists at the court, such as the two Jean Cousins and Antoine Caron, quickly adopted aspects of Italian Mannerism to create a style of painting characterized as the school of Fontainebleau. Less of a tendency to mimic the fashion was noticeable in Corneille de Lyon and Jean and François Clouet, whose portraits, while exhibiting some Mannerist qualities, recalled 15th-century court portraiture. (M.J.Ke./Ed.)

Spain. During the first decade of the 16th century, Fernando Yáñez, who may have assisted Leonardo da Vinci on the "Battle of Anghiari" in 1505, executed works showing a good knowledge of Italian Renaissance developments. Further Italianate tendencies emerged strongly in the Valencian works of Juan de Macip and his son Juan de Juanes. Full-fledged Mannerism made its appearance in the Seville cathedral in the "Descent from the Cross" (1547) by Pedro Campaña (Pieter de Kempeneer), an artist from Brussels, and subsequently in the refined court portraiture of Anthonis Mor (Sir Anthony More) and Alonso Sánchez Coello, whose royal portraits possess an elegance reminiscent of Bronzino's Florentine style. Although Campaña's paintings are Mannerist in composition, they also foreshadow the expressiveness characteristic of Spanish style in the hands of Luis de Morales and El Greco.

From 1546 until his death in 1586, Morales remained almost exclusively in the provincial isolation of Badajoz, developing a highly individual art of great spiritual intensity, radically separated from the Mannerist mainstream. El Greco, though born in Crete, was more fully conversant with Italian painting, having studied with Titian in Venice and later residing in Rome for two years. His Spanish paintings exploit the anatomical attenuations of Roman Mannerism, but the vividly emotional qualities of his colour and paint handling depend almost entirely upon Venetian precedents—Tintoretto and Jacopo Bassano in particular. Under the influence of Counter-Reformation mysticism in Toledo after 1575, he developed an increasingly personal and nonrealistic manner, indulging in space and supernatural light effects (Plate 11). The narrative fervour of his style stands in sharp contrast to the stylish formalism of international Mannerism.

El Greco's exploitation of anatomical attenuations

Germany. Albrecht Dürer was the first important German artist who displayed a profound understanding of Italian Renaissance art and theory. Trained in Nürnberg in the late Gothic tradition, he had ambitions even as a

youth far beyond the narrow confines of his native city and the late medieval style. He traveled to Switzerland and the Rhine Valley and may have been in the Low Countries. Shortly after his marriage in 1494 he made a brief trip to Italy, where he studied the works of Mantegna and the Venetians. In 1505–07 he was again in Italy and was on intimate terms with Giovanni Bellini. Dürer was interested in what he felt to be the "secrets" of Italian art and in the new humanism carried north by such friends as the German humanist Willibald Pirkheimer. As a result, his paintings maintain the northerners' love of detail, rendered meticulously in oil, but he joined to it the Italian interest in broadly conceived compositions. In "The Paumgärtner Altarpiece" of 1502–04 (Ältе Pinakothek, Munich), for example, the saints in the wings are depicted with the scrolls and a complexity of composition more reminiscent of a heraldic achievement, while the broad planes of the architecture and the large, simple figures of the adoration shown on the central panel suggest an Italianate conception. The "Four Apostles" (Alte Pinakothek) of 1526 (Plate 12) ultimately derives from the wings of Bellini's Frari altarpiece. Dürer's close association with the Venetian painter and admiration for his art can be seen in the broad simple folds of the drapery, the breadth of handling of the heads, and the quality of the light depicted.

Dürer's interest in Italian Renaissance culture

Although he executed a large number of important paintings, Dürer is perhaps best known for his woodcuts and engravings, by which he raised printmaking from a minor to a major art (see PRINTMAKING). Dürer's prints, paintings, and writings had such a profound influence on 16th-century art in Germany that it is sometimes difficult to realize that he died in 1528.

In the 16th century the Renaissance, as far as German painting was concerned, tended to follow the lines established by Dürer. Two artists of note do emerge, but their styles are so individual that they do not represent a national school.

Lucas Cranach the Elder was deeply influenced by Dürer and the Danube school, an early 16th-century tradition of landscape painting that was in some ways a transition between the styles of Gothic and Renaissance painting. By 1505 he had moved to Wittenberg and become court painter to the electors of Saxony. There his style changed radically, epitomizing the dichotomy that existed in 16th-century northern European painting. He developed in Wittenberg the full-length portrait in which the sitter is rendered with consummate skill and fidelity. Cranach was a personal friend of Martin Luther and is probably best known for his portraits of the great reformer. At the same time, his "Reclining River Nymph at the Fountain" of 1518 (Museum of Fine Art, Leipzig) illustrates his knowledge of Giorgione and Venetian painting and points the way to the group of highly erotic female nudes of his later works (Plate 12).

Hans Holbein the Younger was trained by his father in Augsburg but took up residence in Basel, Switz., about 1515. He early developed a portrait style that was greatly sought after by the burghers of Basel. His portraits of Burgomaster Meyer and his wife (1516; Kunstmuseum-Öffentliche Kunstsammlung Basel) or of Bonifacius Amerbach (1519; Kunstmuseum-Öffentliche Kunstsammlung Basel) show his gift for characterization. In 1526 he made his first trip to London, where he painted "Sir Thomas More with His Household" (1527). In 1532 religious troubles in Basel were so intense that he accepted a position at the English court and left the city forever. He is perhaps best known for his portraits of Henry VIII, Henry's bride Anne of Cleves (1539; Louvre), and Christina of Denmark (1538; National Gallery, London), at one time considered by the King as a possible bride. "Jean de Dinteville and Georges de Selve" ("The Ambassadors," 1533; National Gallery, London), which depicts two French ambassadors to the English court, is probably the greatest tour de force of his years in England. The two sitters are rendered faithfully in a well-defined room and are surrounded by the trappings of 16th-century humanism—*e.g.*, books, globes, musical instruments. Holbein's portraits were all painted with a great understanding of the sitter and often have a note of Italian elegance. His surfaces tend to be tight and hard, yet there is a certain expansiveness created by the positioning within the frame (Plate 12). He established a portrait tradition in England and also contributed to the popularity of the miniature in that country.

Holbein's first trip to London

Low Countries. In the Low Countries there emerged early in the 16th century a group of painters misleadingly lumped together as the Antwerp Mannerists. Their exaggerated and fanciful compositions descend in great part from the decorative excesses of late Gothic art, generally with some Italianate details probably transmitted by architects' and goldsmiths' pattern books.

The Antwerp Mannerists

The Flemish painter Jan Gossaert, called Mabuse, visited Rome in 1508. At first he continued his ornate late Gothic style, but by 1514 he began to adopt the great innovations occurring in Italian painting. His mythological paintings, such as the "Neptune and Amphitrite" (Staatliche Museen Preussischer Kulturbesitz, Berlin) of 1516, indicate that he was able to understand only the superficialities and not the motivation and *terribilità* of Michelangelo's nudes. Bernard van Orley remained in Brussels and learned of Italy through Raphael's cartoons, which were sent to Brussels to be woven into tapestries. Before the end of the century, painters such as Jan van Scorel, Maerten van Heemskerck, and Sir Anthony More (a Utrecht-born portraitist knighted by Queen Mary I of England) were absorbing Italian influences. Van Scorel demonstrated a specifically Venetian influence, yet all three created an art that was distinctly their own. Joachim Patinir's depiction of the world around him, particularly of landscape, parallels Italian developments in northern terms and greatly influenced Pieter Bruegel the Elder.

Pieter Bruegel the Elder visited Italy in 1551–53 but was more influenced by the Italian and particularly Alpine landscape than by Italian painting. His two-dimensional sources are to be found rather in the popular prints of the time, the landscapes of Patinir, and the fantasies of Bosch. He was also a great observer of peasant life. Bruegel spent his adult life in the company of learned humanists, yet he showed no real interest in classical mythological subjects or antiquity. His paintings illustrating Low Countries' proverbs, children's games, or "The Fight Between Carnival and Lent" (1559; Kunsthistorisches Museum, Vienna) reveal an interest in popular themes and common life rather than in the pedantic Romanizing compositions of some of his contemporaries. This choice of subject matter, latent from the early 15th century in the Low Countries, was given new dimensions by Bruegel. His series of depictions of the months is at once a revival of the labours of the months found in the portal sculptures of Gothic cathedrals and medieval books of hours and at the same time a new treatment of rural landscape and the peasants who work the land. His "Harvesters" (1565; Metropolitan Museum of Art, New York City) displays a remarkable sensitivity to colour and pattern. The intense golden yellow of the ripe wheat sets up a bold pattern across the lower half of the picture and contrasts with the cool greens and blues of the limitless plain stretching off into the distance. Some figures move across a lane cut through the wheat, while others cut into what seems a solid space. The sleeping peasants resting after their noon meal are disposed in patterns and poses that make one feel the heat and calm of the summer's day. This sympathetic view of peasant life, with its bold geometric patterns, runs throughout the series of the months and recurs in "The Wedding Dance" (1566; Detroit Institute of Arts) and "Peasant Dance" (Plate 12) and "Peasant Wedding" (both in the Kunsthistorisches Museum, Vienna).

Bruegel's synthesis of new and old artistic and cultural traditions

Bruegel brought to an end the 16th century in the north and prepared the way for the Baroque. His sons and grandsons were important painters who helped to train some of the leading artists of the 17th century in the Low Countries. It was the elder Bruegel, however, who made landscape and peasant life an accepted subject for painting in the Renaissance. (J.R.Sp.)

The Mannerist style was not comprehended as soon in the 16th century in the Low Countries as it had been in France or Spain. With the notable exception of Frans Floris, it was not until the generation of artists born during the middle years of the century that Mannerism was

fully assimilated. This generation of Flemish and Dutch Mannerists was influenced chiefly by the Italian Mannerists of the second half of the century: Frederik Sustris studied with Vasari; Hendrik Goltzius was an associate of Taddeo Zuccari, Federico's older brother; Johann von Aachen remained in Rome between 1574 and 1588; and Bartholomaeus Spranger collaborated with Federico Zuccari. Haarlem and Amsterdam became the early centres for northern Mannerism. Spranger's style was diffused throughout Europe by the engravings of his colleague Goltzius. Finally, as a late flowering of international Mannerism, Carel van Mander founded a Vasarian academy in Haarlem, in 1604 publishing his biographies of Netherlandish artists in direct emulation of Vasari.

Dutch centres of Mannerism

(M.J.Ke./Ed.)

Bohemia. By far the most ambitious patron of Mannerist art in Europe north of Italy was the Holy Roman emperor Rudolf II, who in the late 1570s established his court at Prague. Between the end of the 16th century and the beginning of the 17th, Rudolf employed architects, sculptors, and painters to create impressive artistic works for his court, much as Cosimo de' Medici had done in Florence. Spranger's "Allegory of Rudolf II" indicates the quality of Rudolf's court art and its clear Mannerist sympathies—sensually graceful figures clad in the dress of classical antiquity and a cultivated facility in composition and execution. (J.R.Sp./Ed.)

Baroque

Baroque is a term loosely applied to European art from the end of the 16th century to the early 18th century, with the latter part of this period falling under the alternative stylistic designation of Late Baroque. The painting of the Baroque period is so varied that no single set of stylistic criteria can be applied to it. This is partly because the painting of Roman Catholic countries such as Italy or Spain differed both in its intent and in its sources of patronage from that of Protestant countries such as Holland or Britain, and it is partly because currents of classicism and naturalism coexisted with and sometimes even predominated over what is more narrowly defined as the High Baroque style.

The Baroque style in Italy and Spain had its origins in the last decades of the 16th century when the refined, courtly, and idiosyncratic style of Mannerist painting had ceased to be an effective means of artistic expression. Indeed, Mannerism's inadequacy as a vehicle for religious art was being increasingly felt in artistic circles as early as the middle of that century. To counter the inroads made by the Reformation, the Roman Catholic Church after the Council of Trent (1545–63) adopted an overtly propagandistic stance in which painting and the other arts were intended to serve as a means of extending and stimulating the public's faith in the church and its doctrines. The church thus adopted a conscious artistic program, the products of which would make an overtly emotional and sensory appeal to the faithful. The Baroque style of painting that evolved from this program was paradoxically both sensuous and spiritual; while naturalistic treatment rendered the painted religious image more readily comprehensible to the average churchgoer, dramatic and illusory effects were used to stimulate piety and devotion. This appeal to the senses manifested itself in a style that above all emphasized movement and emotion. The stable, pyramidal compositions and the clear, well-defined pictorial space that were characteristic of Renaissance paintings gave way in the Baroque to complex compositions surging along diagonal lines. The Baroque vision of the world is basically dynamic and dramatic; throngs of figures possessing a superabundant vitality energize the painted scene by means of their expressive gestures and movements. These figures are depicted with the utmost vividness and richness through the use of rich colours, dramatic effects of light and shade, and lavish use of highlights. The ceilings of Baroque churches thus dissolved in painted scenes that presented convincing views of the saints and angels to the observer and directed him through his senses to heavenly concerns. (Ed.)

EARLY AND HIGH BAROQUE IN ITALY

By the last decades of the 16th century the Mannerist style had ceased to be an effective means of expression. Indeed, in Florence a conscious reassessment of High Renaissance painting had taken place as early as mid-century. This tendency gathered momentum in the last decades of the century, particularly with the Bolognese painters Lodovico Carracci and his cousin Annibale. The Roman Catholic Church's reaction to the Reformation, known as the Counter-Reformation, reaffirmed the old medieval concept of art as the servant of the church, adding specific demands for simplicity, intelligibility, realism, and an emotional stimulus to piety. For the zealots of the Counter-Reformation, works of art had value only as propaganda material, the subject matter being all important; and in Rome there was as a result a sharp decline in artistic quality. Under austere Counter-Reformation popes such as Paul IV and Pius V, most official patronage favoured the dry and prosaic; this late 16th-century style is best called Counter-Reformation Realist. A similar process took place in Florence, where a strong movement away from Mannerist conventions is seen in the paintings of Ludovico Cigoli, and in Milan, where the dominant artistic personalities were the painters Giovanni Crespi (known as Il Cerano) and Pier Francesco Mazzucchelli, known as Il Morazzone.

Effects of the Counter-Reformation

In contrast, late 16th-century Venetian painting was as little influenced by the Counter-Reformation as it had been by Mannerism; and the workshops of Tintoretto, Paolo Veronese, and Palma Giovane remained active until the plague of 1629–30.

Michelangelo Merisi, better known by the name of his birthplace, Caravaggio, a small town near Milan, was active in Rome by about 1595. His earliest paintings are conspicuous for the almost enamel-like brilliance of the colours, the strong chiaroscuro called Tenebrism, and the extraordinary virtuosity with which all the details are rendered. But this harsh realism was replaced by a much more powerful mature style in his paintings for San Luigi dei Francesi, Rome, begun in 1597, and Santa Maria del Popolo, Rome, executed about 1601 (Plate 13). His selection of plebeian models for the most important characters in his religious pictures caused great controversy, but the utter sincerity of the figures and the intensity of dramatic feeling are characteristic of the Baroque. Although Caravaggio had no direct pupils, "Caravaggism" was the dominant new force in Rome during the first decade of the 17th century and subsequently had enormous influence outside Italy.

Parallel with Caravaggio's was the activity of Annibale Carracci in Rome. During Annibale's years in Bologna, his brother and cousin had joined with him in pioneering a synthesis of the traditionally opposed Renaissance concepts of *disegno* ("drawing") and *colore* ("colour"). In 1595 Annibale took to Rome his mature style, in which the plasticity of the central Italian tradition is wedded to the Venetian colour tradition. The decoration of the vault of the gallery in the Palazzo Farnese, Rome (1597–1604), marks not only the high point in Annibale's career but also the beginning of the long series of Baroque ceiling decorations. The third important painter active in Rome during the first decade of the 17th century was the Low Countries' painter Peter Paul Rubens, who became court painter to the duke of Mantua in 1600. He came under the influence of Raphael and Titian, as well as that of Caravaggio, during a journey to Spain in 1603. The rich colours and strong dramatic chiaroscuro of his altarpieces for Santa Maria in Vallicella (New Church), Rome (1606–07), show how much he contributed to the evolution of Italian Baroque painting.

The Carracci synthesis of drawing and colour

Just as the first decade tended to be dominated by the "Caravaggist" painters, the second decade in Rome was the heyday of the Bolognese classicist painters headed by Guido Reni (Plate 13), Domenichino, and Francesco Albani, all of whom had been pupils of the Carracci. The crucial developments that brought the High Baroque into being took place in the third decade.

The little church of Santa Bibiana in Rome harbours three of the key works that ushered in the High Baroque,

all executed in 1624–26: Gian Lorenzo Bernini's facade and the marble figure of Santa Bibiana herself, over the altar, and Pietro da Cortona's series of frescoes of Bibiana's life, painted on the side wall of the nave. The rich exuberance of the compositions is a prelude to the gigantic "Allegory of Divine Providence and Barberini Power," which Pietro was to paint on the vault of the Great Hall of the Palazzo Barberini, Rome (1633–39). Pietro continued with this style of monumental painting for the remainder of his career, and it became the model for the international grand decorative style, which by the close of the 17th century was to be found in Madrid, Paris, Vienna, and even London.

Despite the continued triumph of High Baroque illusionism and theatricality in the hands of Bernini and Pietro da Cortona from the 1630s, the forces of classicism, now headed by the painter Andrea Sacchi and the Flemish-born sculptor François Duquesnoy, gained the upper hand in the 1640s after the death of Pope Urban VIII; and for the remainder of the century the Baroque-versus-classicism controversy raged in the Academy in Rome. Sacchi and the classicists, including the Frenchman Nicolas Poussin, held that a scene must be depicted with a bare minimum of figures, each with its own clearly defined role, and compared the composition to that of a tragedy in literature. But Pietro and the Baroque camp held that the right parallel was the epic poem in which subsidiary episodes were added to give richness and variety to the whole, and hence the decorative richness and profusion of their great fresco cycles. The lyrical landscapes of the French painter Claude Lorrain are among the finest expressions of High Baroque classicism; and they exerted a continual influence throughout the 18th century, particularly in Britain. Even in Rome itself, however, a number of painters of importance succeeded in remaining more or less independent of the two main camps. Sassoferrato (1609–85), for example, painted in a deliberately archaizing manner, carefully reproducing Raphaelesque formulas. The cryptically romantic movement, centred on Pier Francesco Mola, Pietro Testa, and Salvator Rosa, was more important and, together with the landscapes of Gaspard Dughet, was to have considerable repercussions in the 18th century. Claude Lorrain also adopted an independent stand, despite the highly developed classicism of his poetic landscapes and seascapes (Plate 17), both of which, but especially the latter, featured much splendid architecture.

Controversy between the Baroque and classicism

The first two-thirds of the 17th century in Italy were dominated by the Roman Baroque, and few painters elsewhere provided serious competition. Reni, who returned to Bologna from Rome in 1614 and remained there until his death in 1642, remained the strongest artistic personality in that northern city but steadily abandoned the strong plasticity of the Carracci for a much looser style with a pale tonality (Plate 13). When Guercino, in turn, left Rome in 1623, he returned to his native Cento, just north of Bologna, and not until the death of Reni did he decide to settle in Bologna. Guercino's early, fiery style slowly gave way to a much more calm and classical outlook. Venetian painting took a new direction with the rich colours and free brushwork of Domenico Fetti, who had worked in Mantua before moving to Venice. In the hands of Johann Liss (or Jan Lys) the groundwork was laid for the flowering of the Venetian school of the 18th century. Venetian painting was also enriched by the pale colours and flickering brushwork of Francesco Maffei from Vicenza, whereas Bernardo Strozzi in 1630 carried to Venice the saturated colours and vigorous painterly qualities of the Genoese school. Giovanni Benedetto Castiglione also began his career in Genoa and, after a period in Rome, worked from 1648 as court painter in Mantua, where his brilliant free etchings and brush drawings anticipated the Rococo. Naples, under its Spanish viceroys, remained strongly influenced by the "Caravaggesque" tradition, particularly in its best-known painter, a Spaniard, José de Ribera, who settled there in 1616; the two most important native painters of the period, Massimo Stanzione and Bernardo Cavallino, both died in the disastrous plague of 1654.

New currents of painting in Venice

The most conspicuous aspect of the last phase of the High Baroque in Italy is provided by the series of great fresco cycles, which were executed in Rome during the last decades of the 17th century. Pietro da Cortona's decoration of Santa Maria in Vallicella (1647–55) is the link with the earlier phase of the Baroque, and his decoration of the gallery of the Palazzo Pamphili in Rome (1651–54) points the way to the decorations of Giovanni Coli and Filippo Gherardi in the Palazzo Colonna (1675–78) and to those of the vault of the gallery of the Palazzo Medici-Riccardi in Florence by Luca Giordano (1682). Bernini's dynamic and theatrical schemes of decoration reached their climax in the nave vault of the Gesù, Rome, painted in 1674–79 by Giovanni Battista Gaulli (Baciccia) under the direct tutelage of Bernini. The fresco bursts out of its frame and creates an overwhelming dramatic effect, with painted figures flooding over the gilt stucco architectural decoration of the ceiling into the space of the church (Plate 13). After this, the "Allegory of the Missionary Work of the Jesuits," painted by Andrea Pozzo on the nave vault of San Ignazio, Rome (1691–94), seems almost an anticlimax, despite its gigantic size and hypertrophic illusionism. Concurrently, the Baroque-versus-classicism controversy took on a new lease on life, with Gaulli heading the Baroque party in opposition to Sacchi's pupil Carlo Maratta. By the last decades of the century the Baroque was triumphant, and Maratta's Baroque classicism appears almost to be a compromise between Pietro da Cortona and Sacchi. Maratta's style, however, was to provide one of the most important sources for the grand manner of the 18th century.

The essential characteristics of Late Baroque painting can be identified first in the frescoes (1661) of Mattia Preti at the Palazzo Pamphili, Valmontone (southeast of Rome); but the transition between the High Baroque and the Late Baroque was a continuous process and occurred at different dates with different artists. At Valmontone the sense of dynamic structure characteristic of the High Baroque frescoes of Pietro da Cortona yields to a more decorative scheme in which the figures are scattered across the ceiling, giving the painting an overall unity without identifying any specific area as the focal point. Francesco Cozza used this scheme in the Pamphili Library, Rome (1667–73), but among the finest Late Baroque decorations of this type are ceilings painted in Genoa by Gregorio de' Ferrari and Domenico Piola, while Giordano took the style to Spain. The breakdown of any sense of direction in the composition is paralleled by a loosening in the design of individual figures; once again the unity is decorative rather than structural.

Transition between High and Late Baroque

LATE BAROQUE AND ROCOCO

Symptomatic of the changing status of the papacy during the 17th century was the fact that the Thirty Years' War was ended by the Peace of Westphalia in 1648 without papal representation in the negotiations. Concurrently, the influence of Spain also declined. The commencement of the personal rule of Louis XIV in 1661 marked the beginning of a new era in French political power and artistic influence, and the French Academy in Rome (founded 1666) rapidly became a major factor in the evolution of Roman art. Late Baroque classicism, as represented in Rome by Maratta, was slowly transformed into a sweet and elegant 18th-century style by his pupil Benedetto Luti, while Francesco Trevisani abandoned the dramatic lighting of his early paintings in favour of a glossy Rococo classicism. In the early 18th century, Neapolitan painting under Francesco Solimena developed from the brilliant synthesis of Pietro da Cortona's grand manner and Venetian colour that Giordano had evolved in the late 17th century. The impact, also, of Preti is revealed by his predilection for brownish shadows; but, compared to the pupils and followers of Maratta in Rome, Solimena's style has a greater strength and vitality despite the characteristic Late Baroque fragmentation of the composition. He himself supplied large paintings to patrons all over Europe, and his pupils occupied key positions in the mid-18th century. Francesco de Mura took the style to Turin, where he was court painter; Corrado Giaquinto, as court painter in Madrid, turned increasingly toward the Rococo, and Sebastiano Conca worked in Rome, falling increas-

Co-existence of Late Baroque classicism and the Rococo

ingly victim to the academic classicism dominant there. Anton Domenico Gabbiani practiced a particularly frigid classicism in Florence, and it was mainly in Bologna and Venice that real attempts were made to break away from the confines of Late Baroque classicism.

Giuseppe Maria Crespi (called Lo Spagnolo, "The Spaniard") turned instead toward the early paintings of Guercino and evolved a deeply sincere style, remarkable for its immediacy and sensibility. In Bologna he had no real successors, but in Venice his work provided one of the bases for the brilliant flowering of Venetian painting in this period. While Giovanni Battista Piazzetta looked toward Crespi for the basis of his expressive Tenebrist style, Sebastiano Ricci took his cue from Giordano. The brilliant lightness and vivacity of his frescoes in the Palazzo Marucelli-Fenzi, Florence, mark the beginning of a great tradition of Venetian decorative painting, a tradition that was to be carried all over Europe by Giovanni Antonio Pellegrini, Giambattista Pittoni, and, above all, Giovanni Battista Tiepolo. The vast majority of the finest decorations (*e.g.,* frescoes) carried out by the Venetian 18th-century painters were executed outside the Veneto (the region of which Venice is the principal city), but the opposite is true of the flourishing Venetian school of landscape, *vedute* ("views"), and genre painters. Giovanni Antonio Canal, called Canaletto, developed the views of Venice painted by Luca Carlevaris into an industry almost entirely dependent upon foreign tourists; and his nephew Bernardo Bellotto spent most of his career painting views in central Europe. Francesco Guardi avoided the cool precision of the *vedute* of Canaletto and Bellotto and instead evolved a much lighter and more lyrical Rococo style with a strong sense of the picturesque (Plate 13) and, occasionally, the bizarre. In Rome a similar contrast existed between the brilliant, precise *vedute* of Giovanni Paolo Pannini and the strange, almost Romantic *vedute* in the form of etchings by Giovanni Battista Piranesi.

Revitalization of the Venetian school

Spain and Portugal. Two fundamental and ostensibly opposed streams permeate Spanish painting and separate it from that of the rest of Europe—ecstatic mysticism and sober rationalism. These qualities are essentially Gothic in spirit, and the Iberian Peninsula is remarkable for the tenacity with which Gothic ideas were retained and for the relatively small influence of Renaissance humanist ideas. The early 17th-century still lifes of Sánchez Cotán, with their strong realism and harsh, mysterious lighting, illustrate these contrasts admirably, whereas Luis Tristán abandoned the Mannerist style of his master El Greco for a much more careful realism. Francisco Pacheco, the teacher and father-in-law of Velázquez, was a more important writer than painter, and his writings laid down a theoretical basis for the Spanish approach to spirituality through naturalism. The early works of José de Ribera show a synthesis of Spanish realism and ideas drawn from both Annibale Carracci and Caravaggio; the fierce darkness of these paintings formed the basis of the Tenebrist style that dominated Neapolitan painting during the first half of the 17th century. Ribera himself, however, developed away from this style in his later paintings and moved toward a softer and more even handling of light (Plate 18). Francisco de Zurbarán was active mainly in Seville until his removal to Madrid in 1658, and unlike Ribera he painted throughout his life in the stark Spanish realist style. The massive solemnity of his figures and simple, clear-cut compositions are wholly in sympathy with the demands of the Counter-Reformation, and only in Madrid did he come under substantial Italian influence (Plate 18).

Tenacity of Gothic ideas in the Iberian Peninsula

Diego Velázquez was almost the exact contemporary of Zurbarán, but, unlike Zurbarán, who spent almost all his life in the company of monks in the provinces, Velázquez' time from 1623 was spent in the Spanish court in Madrid. His early *bodegones* (scenes of daily life with strong elements of still life in the composition) were painted in Seville and belong to the Spanish realist tradition, but at court he saw the Titians collected by Philip II and also Rubens' paintings. After he visited Italy in 1629–31, there was greater freedom in the way he handled paint, more interest in colour, and increased depth to his analyses of character (Plate 18).

The early works of the Seville painter Bartolomé Esteban Murillo again follow the Spanish realist tradition in their cool detachment, but in his late works his style softened and sweetened into a sentimentality that proved immensely popular (Plate 18). Alonso Cano formed his early painting style in Seville on the simple monumentality of Zurbarán, but after he moved to Madrid in 1638 his paintings took on a new elegance and gracefulness. (Cano was also active as a sculptor and architect in Granada [1652–57]). Antonio del Castillo and Juan de Valdés Leal were the most important painters active in Andalusia after Murillo, and the works of both reveal that liveliness of handling, with accents of strong local colour, which replaced the sober realism popular in the first half of the century.

Portugal was ruled by Spain until 1640, when John IV was proclaimed king. But economic conditions hampered serious patronage of the arts until the reign of John V, when the most distinguished painter was Francisco Vieira de Matos. Unfortunately, the Lisbon earthquake of 1755 destroyed much of the best art collected in the Portuguese capital at that time.

Low Countries. *The Spanish Netherlands.* The year 1566 saw the Netherlands in open revolt against Philip II of Spain, and, inasmuch as this revolt had a Protestant as well as a nationalist aspect, a wave of iconoclasm swept across the area. By 1600 the area had become divided into the Spanish-dominated, Catholic, southern provinces—broadly modern Belgium—and the independent, predominantly Calvinist United Provinces of the north—broadly the modern Netherlands, or colloquially Holland; the boundary between the two remained fluid, however. In the southern provinces throughout the 16th to 18th centuries Brussels, headed by viceroys, remained the centre of court patronage, while Antwerp, with its great patrician families, was the commercial centre.

Painting in the southern provinces before 1610 was intensely conservative; the Mannerist conventions were never accepted as fully as in the north. Instead, Italianate ideas were joined with the late Gothic tradition.

Peter Paul Rubens arrived back in Antwerp from Italy late in 1608. In the following year he was appointed court painter to the archduke Albert and the archduchess Isabella, with special permission to reside in Antwerp, to help repair damage caused by the iconoclasm of 1566. The necessary ingredients were present for a brilliant flowering of the Baroque art that Rubens had evolved in Italy, and his studio became an artistic centre not only for the Netherlands but for England, Spain, and central Europe as well. The monumentality of Rubens' forms, with their impulsive drawing, restless movement, and dramatic lighting, provided the touchstone for the High Baroque in the Catholic areas of northern Europe (Plate 14). By Rubens' death, Philip IV of Spain had acquired more than 130 paintings by him. A diplomatic visit to England (where he found so much favour with Charles I that the latter knighted him) in 1630 had resulted in the commission to decorate the ceiling of the Banqueting Hall in Whitehall, one of the most monumental commissions of Rubens' last period.

Rubens' studio: a European artistic centre

Anthony Van Dyck, a pupil and assistant of Rubens, was a much less forceful personality than his master; and this is reflected in the quieter, more introspective note characteristic of his paintings. His greater sympathy for the sitter made him the most successful portrait painter of his time. Between 1625/26 and 1632 he was active, mainly as a portrait painter, in the entourage of Rubens, but the last years of his life (1632–41) were spent in England as court painter to Charles I, from whom he, too, received a knighthood (Plate 15). The elegant, relaxed, aristocratic portrait style he introduced was outstandingly successful and rendered obsolete the stiff portraits of Daniel Mytens and the straightforward, unpretentious portraits of Cornelius Johnson, two other painters of Low Countries origin active in England at this time. Van Dyck's death coincided with the outbreak of the Civil War in England; and the portraitists William Dobson and Robert Walker, in the troubled years 1641–60 the only painters of note active in England, reveal a considerable debt to him. Jacob Jor-

daens also worked as an assistant in Rubens' workshop in Antwerp and took it over after his death. His handling of the Rubensian idiom moved increasingly away from the control of Rubens himself toward a much more boisterous and vulgar style with an emphasis on large genre scenes populated with rough plebeian types.

The remaining members of Rubens' studio, such as Cornelis de Vos and Caspar de Crayer, were much weaker artistic personalities, and one of the few painters of genius relatively independent of Rubens was Adriaen Brouwer, who painted in the tradition of Pieter Bruegel the Elder. Best known for his low-life pictures (Plate 14), Brouwer also painted very expressive landscapes; his work is characterized by the sensitive use of a heavily loaded brush. In comparison, David Teniers the Younger was a minor master, and with him the influence of Dutch painting became increasingly strong. The impact of Rubens' landscape style is felt in the paintings of Jan Wildens and Lucas van Uden, while in contrast Jan Brueghel the Younger turned the making of copies and pastiches of his father's works into something approaching an industry. Still-life and animal painting reached new heights in the works of Frans Snyders as a result of the influence of Rubens, and in a much quieter vein Snyders' pupil Jan Fyt continued the tradition, which was to last into the 18th century. Jan Davidsz de Heem was also active in Holland, but he is important as one of the creators of the elaborate, fully developed Baroque still life, and as such he had a host of followers and imitators.

The United Provinces. Dutch painting of the 17th century shares roots with that of the Spanish Netherlands. Holland, however, was independent, rapidly prospering, and almost entirely Protestant. In the last decades of the 16th century the great port of Haarlem was the most active artistic centre, and the remarkable flowering of Mannerist painting there, as exemplified by Cornelis van Haarlem and Hendrik Goltzius, is without a parallel south of the border. In the later pictures of Abraham Bloemaert, Mannerism gave way to the much more straightforward realist style characteristic of the earliest phase of Dutch 17th-century painting. The influence of the figure paintings of Adam Elsheimer on this generation of artists was considerable; his particularly Italianate style, with sharply delineated forms painted in rich, deep colours and with a pronounced element of fantasy, is reflected by the early paintings of Leonard Bramer and, even more importantly, Pieter Lastman, the master of Rembrandt. Elsheimer's poetic little landscapes were also extremely important for the group of Dutch artists active in Rome about 1620. This group was headed by Cornelis van Poelenburgh and Bartolomeus Breenbergh, and back home it provided an additional source of Italian influence. The most striking influence of Italy was provided, however, by the Dutch followers of Caravaggio, who had seized eagerly upon the harsh dramatic lighting and coarse plebeian types they had seen in his paintings during their stays in Italy and brought the style to the north to form the so-called Utrecht school. Gerrit van Honthorst, Hendrik Terbrugghen (Plate 15), and Dirck van Baburen were leading champions of this style, but after 1628 Honthorst turned away in the direction of Van Dyck.

Frans Hals was born in Antwerp, but almost all of his life was spent in Haarlem, where he evolved his characteristic bravura style of portraiture (Plate 15). The stiff solemnity of earlier Dutch portraits gave way to the capture of fleeting changes of expression and superb textural effects, though Hals never succeeded in attaining the degree of psychological penetration characteristic of the portraits painted by Rembrandt.

The early works of Rembrandt van Rijn, painted in Leiden (1625–31), show a progressive lessening of the influence of Lastman, and Rembrandt, together with his associate Jan Lievens, evolved an increasingly Baroque style, with strong contrasts of light and shade derived from the "Caravaggists." After he moved to Amsterdam in 1631, these tendencies developed to an opulent and highly Baroque climax in the late 1630s. Following the death of his first wife, Saskia, in 1642, difficult times and the changing tastes of art collectors culminated in his bankruptcy in 1656. In his later works the dramatic Baroque panache gives way to a deep introspection and sympathy for his subjects (Plate 15), and his series of about 60 self-portraits reveals this process in intimate detail. Parallel to his development as a painter is that of his style as an etcher; Rembrandt is considered by many to be the greatest etcher of all time (see PRINTMAKING). During the years of his financial success, Rembrandt had the largest and most successful painting and printmaking studio in Holland.

The increasing use at this time of portable easel paintings as domestic ornaments, many of them made for sale by dealers rather than on commission by the consumer, is related to the extraordinary range of subjects in which Dutch painters specialized. Nevertheless, certain basic changes in style and taste occurred during the course of the 17th century, and, although many painters long persisted in outdated styles, the same fundamental changes can be traced in the various specialities. The earliest phase of simple realism held sway until the early 1620s; and the characteristic bright local colours, lack of spatial unity, sudden transition between different planes, and tendency toward high viewpoints are to be found in the genre paintings of Willem Buytewech, flower pieces of Jacob II de Gheyn and Roelant Savery, and marine paintings of Hendrick Cronelisz Vroom and Adam Willaerts. This gave way to a much more limited palette in the early 1620s when, by reducing the strength and range of the colours, an atmospheric unity was obtained. In landscapes and marine paintings the horizon tended to drop, and a continuous and coherent recession into depth was attained, particularly in the paintings of Esaias van de Velde, Jan van Goyen, Hercules Seghers, and Jan Porcellis. The same change is seen in still lifes by Pieter Claesz and Willem Claesz Heda, in which the colours are almost monochrome (Plate 14). Atmospheric unity having been mastered, the change to the heroic classical phase of the middle of the 17th century was gradual, but there was a tendency toward ever-increasingly dramatic Baroque contrasts, be they the leaden skies or great oaks of Jacob van Ruisdael (Plate 14), the vast panoramas of Philips de Koninck, the luminous pastures of Aelbert Cuyp, or the heavy gray seas of Simon de Vlieger. The monumentality of these scenes is paralleled by the rich splendour of the still lifes of Jan Davidsz de Heem, Abraham van Beyeren, and Willem Kalff and the classical calm and simplicity of the scenes by Jan Vermeer (Plate 14) and Pieter de Hooch painted in Delft. In the landscapes of Meindert Hobbema, Claes Berchem, and Adam Pijnacker the majesty of Jacob van Ruisdael's landscapes gives way to a much lighter, more picturesque style. Similarly, the vigorous social realism of Adriaen van Ostade yields to a much lighter and more frivolous treatment in the paintings of his younger brother Isack and Jan Steen and the elegant hunting scenes of Philips Wouwerman.

With the French invasion of 1672 and the subsequent Dutch economic collapse, the demand for paintings dropped heavily, and in the last decades of the 17th century many Dutch painters either stopped painting or, like the van de Veldes Willem I and Willem II, left the country to work in England or Germany. Late 17th- and 18th-century taste tended toward the almost enamel-like brilliancy and intricate detail of the still lifes by Rachel Ruysch and Jan van Huysum; the same slightly dated flavour is characteristic of the marine paintings of Ludolf Backhuysen and of the hard figure subjects of Willem van Mieris and Adriaan van der Werff.

Decline of the Dutch school

France. French-speaking painters continued the Mannerist conventions even later than did those at Haarlem, and at Nancy (capital of the independent duchy of Lorraine before 1633 and again from 1697 to 1766) a group of artists around Jacques Bellange and Jacques Callot was responsible for the last great flowering of the Mannerist style in Europe. By comparison, painting in Paris during the first decades of the 17th century was relatively insignificant, with the exception of that of Claude Vignon, who exchanged his Mannerist training for a style based on Elsheimer and to a lesser extent Lastman, and who in the 1620s revealed a remarkable knowledge of the earli-

est paintings of Rembrandt. The return of Simon Vouet to Paris, however, marked the arrival of the Baroque in France. The earliest paintings from his stay in Rome are strikingly vigorous essays in the "Caravaggesque" style, but by 1620 he was painting in an eclectic, classicizing style based on the early Baroque painters active there, including Giovanni Lanfranco and Guido Reni. This style he brought back to France, enjoying until his death an immense success in Paris as a decorator and painter of large-scale altarpieces; even the return of Nicolas Poussin failed to shake his position. Poussin's activity in Paris is of relatively little importance compared with the remainder of his career in Rome, but the large number of works commissioned by French patrons then and subsequently was an important factor in the formation of the French predilection for classicism (Plate 16).

The influence of the highly Baroque paintings depicting the life of Marie de Médicis that Rubens had executed for the Luxembourg Palace in Paris was small. But Philippe de Campaigne evolved a grave and sober Baroque style that had its roots in the paintings of Rubens and Van Dyck rather than in Italy. Clear lighting and cool colours with an austere naturalism provided an alternative to the intellectual and archaeological classicism of Poussin. Georges de La Tour, a painter who had affinities with the Dutch "Caravaggists" of Utrecht, was active in Lorraine; but although he exploited the Caravaggist system of lighting, his figures became increasingly detached and simplified, leading to an uncomfortable hardness (Plate 16). The paintings of the Le Nain brothers—Antoine, Louis, and Mathieu—again look to Dutch painting for their inspiration (Plate 17). Eustache Le Sueur began painting under the influence of Vouet, but after Poussin's brief return to Paris (1640–42) he turned to a much more rigorous classical style influenced by Raphael's tapestry designs, whereas Sébastien Bourdon was capable of painting in almost any current style on request.

In the reorganization of the Academy of Painting and Sculpture in 1648, Charles Le Brun was appointed director and given the position of virtual dictator of the arts in France. An imaginative painter and designer, Le Brun was also a brilliant organizer, and the creation of the Louis XIV style, as exemplified by the Palace of Versailles, was above all due to him. The particular Baroque style that emerged was based on the Roman High Baroque but was purged of all theatricality and illusionism and modified to conform to the classical canons of French taste; this compromise solution struck the keynote for the frescoes of Le Brun and Pierre Mignard. The more full-blooded Baroque style of Pierre Puget received little official recognition, and his attempts to obtain major commissions at Versailles were thwarted, probably because of his difficult nature. During the last decades of the century, the full Baroque style took on a new lease on life, and the decorative paintings of Charles de La Fosse and Antoine Coypel clearly reveal the influence of Rubens. Even more Baroque are formal portraits by Hyacinthe Rigaud and Nicolas de Largillière, in which the strong contrapposto (twisting of the figure so that one half is in opposition to the other), rich settings, and floating masses of drapery reflect the pomp and swagger of this era—which, significantly, came to be known as the Grande Époque.

Creation of the Louis XIV style

The great formal portraits of Largillière and Rigaud are entirely Baroque in their approach, but in the late informal portraits of these masters a new atmosphere prevails. This atmosphere goes by the name of Rococo. The turn of the century marks the victory of Rubens' influence over the severe classicism of Poussin. The evolution of the Rococo style of decoration has been traced from its emergence at the beginning of the 18th century, and it must be emphasized that the Rococo is fundamentally a decorative style. It made relatively little impact on religious painting in France, and painters such as Pierre Subleyras continued to work in a Baroque idiom until the arrival of Neoclassicism in the second half of the century. It took the genius of Antoine Watteau to put together all the ideas current in Paris and to create the new style of painting. Rubens (in particular his oil sketches), the brush drawings and etchings of Castiglione, the naturalism of the Dutch painters, and the fantasy of the French artist Claude Gillot all provided important source material for early Rococo painting. The delicate sketchlike technique and elegant figures of Watteau's wistful fantasies, called *fêtes galantes* (Plate 16), provided the models for the paintings of Jean-Baptiste Pater and Nicolas Lancret, both of whom conveyed a delicately veiled eroticism. Eroticism was more explicit in the sensuous nudes, both mythological and pastoral, of François Boucher (Plate 16). Another painter with whom amorous dalliance is a hallmark was Jean-Honoré Fragonard, in whose soft landscapes flirtation and even seduction are conducted with gallantry. Such paintings formed an intimate part of the decoration of Rococo interiors, and more than any earlier secular paintings they were intended as a kind of two-dimensional furniture.

Fantasy and eroticism

The furniture role also applies to the paintings of dead game and live dogs by François Desportes and Jean-Baptiste Oudry. But in the still lifes and tranquil scenes of domestic life painted by Jean-Baptiste-Siméon Chardin there is a sobriety of colour and composition (although great richness in the handling), an often relatively homely subject matter, and a concern to order the mind rather than dazzle the eye (Plate 17). Some of Chardin's subjects—the labours of the servant class, the care of children—were shared by Jean-Baptiste Greuze, who was, however, more interested in narrative and sentiment. Unlike Dutch painters of lower-class life, Greuze endowed his peasants with the sensibility of their social superiors. The edifying moral sympathy he intended to inculcate was, however, often subverted by a sly erotic interest he could not resist giving expression to.

Despite his great success, Greuze was judged to have failed in his attempt at painting heroic narrative from ancient history. But then it is true that the "higher" class of painting was generally less successfully practiced in France than were the "lower" genres in the 18th century. The mythologies and altarpieces of the Coypel family, Jean-François de Troy, or Jean-Marc Nattier may have been underestimated, but their names are not as familiar as those of still-life and genre painters such as Watteau or Chardin or even those of such accomplished painters of capricious ruin pieces or of landscapes and seascapes as Hubert Robert and Claude-Joseph Vernet.

The middle decades of the 18th century saw more accomplished portrait painters flourishing in France than perhaps ever before in any country. Yet it is the informal, the convivial, and the intimate that are associated with the portraiture of Jacques-André-Joseph Aved, François-Hubert Drouais, Louis Tocqué, Louis-Michel Van Loo, or Étienne Aubry. The heroic was seldom attempted and never achieved.

Britain. *The 17th century.* English painting during the 17th century had been dominated by a series of foreign-born practitioners, mostly portraitists (*e.g.*, Rubens and Van Dyck), even before the Civil War. Sir Peter Lely and Sir Godfrey Kneller continued this trend after the Restoration. The vast majority of the painting executed by native artists remained thoroughly provincial. Lely began his activity in England during the Civil War, probably in 1641, but his portraits of the members of the court of Charles II set the pattern for English portraiture of the second half of the 17th century. British patrons in the 18th century sometimes collected paintings on religious or mythical themes by foreign artists, but at home they rarely commissioned anything other than portraits, landscapes, and marine paintings, although there was in the early 18th century a vogue for grand allegorical decorations in aristocratic houses. The Protestant church, however, did little to encourage painting. In fact, the preponderance of portraits is the most distinctive characteristic of old British collections. Gerard Soest, Jacob Huysmans, and Willem Wissing were also active in England as portrait painters close in style to Lely, whereas Jan Siberechts and Robert Streeter painted "portraits" of English country houses. The most distinguished painters to settle in England during this period were the van de Veldes, from whom the tradition of British marine painting descends, headed by Peter Monamy and Samuel Scott.

English portraiture in the 17th century

The Glorious Revolution of 1688 was followed by a brief

flowering of decorative painting under Sir James Thornhill, which was the closest that Britain ever approached to the developed Baroque style of the Continent. This process was in part due to the influx, following the end of the War of the Spanish Succession, of Italian painters, including the Venetians Giovanni Antonio Pellegrini and Jacopo Amigoni, and French ones, such as Charles de La Fosse. The German-born Kneller succeeded Lely as court portrait painter, but, although his portraits often have a certain liveliness, his rather heavy use of studio assistants resulted in a tendency to monotony.

Hogarth

The 18th century. Thornhill's son-in-law William Hogarth was, despite his chauvinism and virulently anti-French sentiments, heavily influenced by the continental Rococo style. Early in his career he succeeded in breaking away from the straitjacket of portraiture, and his moralizing paintings are superb evocations of life in the England of George I and George II (Plate 17). His rich, creamy paint handling and brilliant characterization of textures have a freshness and vitality unequaled in the work of any of his contemporaries. He invented a new form of secular narrative painting that imparts a moral. These paintings were often tragicomedies, although dependent upon no texts, and Hogarth's series of such works were always intended to be engraved for a large public as well as seen in a private picture gallery (just as plays were intended to be performed as well as read).

Despite Hogarth's considerable knowledge of and borrowings from continental old masters, he remained in the last analysis English through and through. This, however, was not the case with all the next generation of painters; and the Scottish-born Allan Ramsay studied in Rome and Naples in 1736–38 before settling in London in 1739. Until the return of Joshua Reynolds from Italy in 1752, Ramsay held undisputed sway as the most successful portrait painter in London; and to him must be given the credit for the initial marriage of the Italian "grand style" to English portraiture. Ramsay visited Italy again in 1755–57, and on his return his portraits took on a new delicacy and elegance and a silvery tonality. Reynolds possessed great ambitions and a more profound acquaintance with the old masters than any of his contemporaries. His colouring and handling can be compared with Rembrandt, Rubens, and Veronese, and his poses are indebted to the sculpture of antiquity and to Michelangelo. The *Discourses* that he delivered to the Royal Academy (founded in 1768 with Reynolds as its first president) are the most impressive statement in English of the central ideas of European art theory from the time of Leon Battista Alberti's treatise. Reynolds' own painting gained a genuine heroic power and elevated grace from his frustrated ambition to be a history painter, although for that very reason he occasionally tumbled into bathos.

Wilson's Italianate landscapes

The third major British painter of the period to study in Italy was a Welshman, Richard Wilson, who worked there from 1750 to about 1757 before settling in London. His landscape style was formed on Claude, Gaspard Dughet, and Cuyp; but the clear golden lighting of his Italian landscapes carries the conviction of an artist saturated with the Mediterranean tradition. A cooler clarity and classical simplicity pervade his northern landscapes; and, despite the uneven quality of his work, Wilson was the first British painter to lift the pure landscape above mere decorative painting and topography.

Thomas Gainsborough was in every way the antithesis to Reynolds. Trained entirely in England, he had no wish to visit Italy. Instead of the "grand style," his tastes in portraiture lay in the delicate flickering brushwork and evanescent qualities of the Rococo (Plate 17). He preferred landscape painting to portraiture, and the strong Dutch influence in his earliest works later gave way to spontaneous landscapes composed from models.

In the 1760s Francis Cotes was the most important fashionable London portrait painter after Reynolds and Gainsborough, a position succeeded to by George Romney, who, on returning to London from Italy in 1775, took over Cotes's studio. Romney's portraits deteriorated sadly in quality during the 1780s when the young Sir Thomas Lawrence began to make his mark.

Throughout the 18th century, portraiture remained the most important genre of British painting, despite the efforts of Reynolds and Gainsborough in their "fancy pictures." Even the taste for large-scale scenes illustrating Shakespeare and other themes—which were commissioned toward the end of the century from James Barry, James Northcote, and Edward Penny, among others—never spread far beyond a few patrons. Sporting and animal painting, however, took on an entirely new dimension in the work of George Stubbs. Joseph Wright of Derby was active outside London and, apart from his romantic portraits, is important for his series of paintings of scientific and industrial subjects with strong light effects. Johann Zoffany was born in Germany but moved to Britain about 1761 and became a founder-member of the Royal Academy, specializing in elaborate group portraits and theatrical scenes.

During the second half of the 18th century the evolution of British oil painting was to a great extent paralleled by the extraordinary flowering in watercolours. The early topographical drawings of Paul Sandby gave way to the delicate linear drawings of Francis Towne, with their patches of colour resembling maps, and, at the close of the century, to the atmospheric unity of the landscapes of John Robert Cozens.

Colonial Americas. *North America.* Painting in the Dutch and English colonies of North America reflected generally the portrait styles of the mother countries, though with a note of provinciality. In the late 17th and early 18th centuries the Dutch colony of New Amsterdam (New York) had painters whose names today are forgotten. Their work lives on, however, and is signified by names such as the Master of the De Peyster Boy. Gustavus Hesselius, Swedish born, was painting in Maryland, and Jeremiah Theüs, a Swiss, was at work in South Carolina. Peter Pelham and John Smibert arrived from England and in the second quarter of the 18th century were painting portraits in Boston, Mass. These two self-taught itinerant artists were succeeded by John Wollaston and Joseph Blackburn. Robert Feke, a native American painter, realized his forms more solidly and with greater originality than his predecessors had. Another native American, John Singleton Copley, worked in Boston until 1774, when he went to live permanently in England, and was responsible for the finest painting produced in the American colonies. Benjamin West, another important native figure in the history of American painting, was born in Pennsylvania but settled in London in 1763, where he became the second president of the Royal Academy. Although domiciled in London, he helped to mold the styles of two generations of American painters.

Central and South America. Baroque painting in Central and South America is basically an extension of that of Spain and Portugal, and even the best rarely rises to the general standard of the European schools. Important paintings and sculptures tended to be imported from Europe, and Zurbarán was particularly active in producing works for export, while local productions were more or less heavily influenced by the Indian traditions.

Central Europe. In central Europe the Mannerist tradition remained dominant until the Thirty Years' War (1618–48), particularly in Bohemia and Bavaria, where Italian influence was perhaps strongest.

Influence of Rubens

The Rubensian Baroque became dominant after mid-century, and here the lead was taken by Silesia and Bohemia. Michael Willmann, originally from Königsberg (modern Kaliningrad) on the southeastern Baltic coast, developed a highly charged, emotional Baroque style, based on Rubens, at Lubiąż (modern Dorf Leubus, northwest of Wrocław) from 1661 to 1700 and at Prague after 1700. In Karel Škréta Šotnovský, Bohemia possessed a painter of European stature; his sombre portraits and religious scenes are filled with a deeply serious mystical fervour. The frescoes by Johann Michael Rottmayr in the castle of Vranov in Moravia (1695) and in Breslau (now Wrocław; 1704–06) constitute a prelude to the great development of Baroque painting in the Habsburg domains. There the vigorous and extremely colourful frescoes are closely integrated with the architecture. The

vast majority of the best central European Baroque painting outside portraiture is monumental in scale, and the concept of the *Gesamtkunstwerk* ("total work of art")—where painting, sculpture, and architecture are combined together into a single, unified, and harmonious ensemble—is of overwhelming importance.

Painting in Austria flourished, and Franz Anton Maulbertsch is arguably the greatest painter of the 18th century in central Europe. The vast majority of his brilliant fresco cycles are located in relatively inaccessible areas of Bohemia, Moravia, and northern Hungary. But the mystical intensity of his religious scenes and the joyous abandon of his secular subjects form a triumphant closing chapter to 18th-century central European painting. Maulbertsch's last frescoes at Strahov, Prague (1794), reveal, nevertheless, the impact of the Neoclassicism that descended in the last decades on all Austrian painters, including Troger's pupil Martin Knoller. But Austrian monumental painting remained fully Baroque in the hands of Daniel Gran, Paul Troger, and Bartholomäus Altomonte; and it was not until the latter part of the century that the Rococo made its impact.

Independent development of Bohemian painting

During the first four decades of the 18th century, Bohemian Baroque painting developed almost independently of Vienna, where the Habsburg rulers of Bohemia had their capital. The impetuous work of Jan Petr Brandl and the powerful realism of the portraitist Jan Kupecký, who worked in Rome, Venice, Vienna, and Nürnberg, always remained Bohemian in spirit. The frescoes of Wenzel Lorenz Reiner, however, show more Italian influence. One of the few important Baroque frescoes of the second half of the century is that by Jan Lucaš Kracker in St. Nicholas, Malá Strana ("Lesser Quarter"), Prague. The influence of Bohemian Baroque painting is frequently underestimated. Apart from Vienna and the surrounding area, it was dominant in Silesia and strong later in the century in Franconia.

After the death of Cosmas Damian Asam in 1739, Johann Baptist Zimmermann became the most important fresco painter in the Munich area; his lyrical handling of pale colours is typical of the Rococo period. Christian Wink continued to paint in the same style until the close of the century. In Georg Desmarées the court at Munich gained a painter in whose Rococo portraits there is more than a hint of decadence.

The centre of south German painting had by the late 1730s shifted from Munich to Augsburg in Swabia, where Johann Georg Bermüller became the director of the Academy in 1730; but his frescoes, as well as those of Franz Joseph Spiegler and Gottfried Bernhard Goetz, are perhaps more representative of the Late Baroque than the Rococo. The frescoes of Matthäus Günther, who became director of the Augsburg Academy in 1762, show a steady evolution from his early Baroque compositions, through the much lighter asymmetrical Rococo compositions, to the strongly sculptural quality of his late works, which reveal the onset of Neoclassicism.

In Franconia and the middle Rhineland the most important painters were Johann Zick and Carlo Carlone. Zick's frescoes at Würzburg (1749) had not been entirely successful, and in 1750 he was supplanted by Tiepolo; but at Bruchsal he produced one of the most brilliant series of Rococo frescoes in Germany (now destroyed). His son Januarius began painting in the Rococo style but under the influence of Anton Raphael Mengs produced some late frescoes that were strongly classical.

Prussian Rococo

The French tastes of Frederick I of Prussia at Berlin led him in 1710 to summon Antoine Pesne to court, where Pesne continued for the remainder of his life to paint in an entirely French Rococo style. The homely intimacy of the paintings of Daniel Chodowiecki, however, have a sensitivity and refinement more comparable to Chardin's.

Saxony under Augustus III produced few painters of real importance except Mengs, who rapidly turned from the Rococo to the Neoclassicism propounded by the influential art historian and classical archaeologist Johann Winckelmann.

Poland. King Władysław IV Vasa (reigned 1632–48) assembled an important collection of Italian and Flemish Baroque paintings, but these promising developments were cut short by the destruction of the Swedish Wars in the middle of the 17th century. Under John III Sobieski (reigned 1674–96), a cultivated man, there was a considerable revival, and, although two of the painters active in Poland—Claude Callot and Michelangelo Palloni—were foreign-born and foreign-trained, native talent flowered with the work of Jerzy Eleuter Szymonowicz-Siemiginowski and Jan Tretko. In 1697 the crowns of Poland and Saxony were united under Augustus II, and he and his son Augustus III ruled over Poland until 1763. During this period, Polish painting formed part of the Saxon tradition, but during the reign of the last king of Poland, Stanisław II August Poniatowski (reigned 1764–95), Warsaw quickly became a centre of European importance. Although inclined to Neoclassicism in architecture, Stanisław's taste in painting was more conservative. Accordingly it is the late Rococo portraits of Marcello Bacciarelli that are particularly important. A nephew and pupil of Canaletto, Bernardo Bellotto, settled in Warsaw in 1767 and executed for Stanisław the great series of 26 views of the city that were intended to hang in the Royal Castle. (P.C.-B./Ed.)

Russia. The Baroque in Russia was imported from western Europe and outside court circles made little impact. Indeed the traditional production of icons for the Orthodox church by artists of the Novgorod and Moscow schools continued throughout the Baroque period. Nevertheless the foundation of St. Petersburg (1703) by Peter I the Great marked the beginning of the substitution of Western influence for Byzantine, an important change. During Peter's reign foreign painters began to go to Russia in increasing numbers; conversely, groups of young Russians were sent to Italy, France, Holland, and England to study painting. Western influence determined the character of Russian painting for more than two centuries.

Influx of foreign painters

The art of Peter's age shows almost no trace of Byzantine influence. Only in iconography did the old style persist for some time. Early in the 18th century, religious painting began to give way to secular painting, and the church prohibition of sculpture became ineffective. Dmitry Levitsky stands out as the only important Russian painter of the 18th century to work in the Western style.

Further westernizing occurred under the empress Elizabeth (reigned 1741–62), who had French tastes. A great number of vast and luxurious Rococo-style palaces were built, and painting was primarily concerned with their interior decoration—ceilings and walls. The work was carried on chiefly by Italians and Frenchmen.

In 1757 the Academy of Fine Arts was founded in St. Petersburg, and foreign artists—mostly French—were invited to direct the new school. These trained some remarkable native portraitists, such as Ivan Argunov, Anton Losenko, and Fyodor Rokotov. Their works reflected the ceremonial character of Elizabeth's tastes and showed little evidence of native Russian sensibility. (A.Vo./Ed.)

Scandinavia. In the 17th century, Scandinavian painting derived from traditions of the Low Countries and northern Germany. The works of art carried off as loot from Prague by Swedish soldiers during the Thirty Years' War might conceivably have broadened the outlook of Swedes at home, but the best of them were taken to Rome by Queen Christina when she abdicated in 1654. A generation later, under the influence of the fashionable Venetian woman pastelist Rosalba Carriera, a school of Rococo portraitists flourished in Scandinavia. One such portraitist was Carl Gustav Pilo, who, though trained in Stockholm, executed many frankly Venetian portraits during his years as court painter in Copenhagen. Another was Lorentz Pasch the Younger, who trained under Pilo in Copenhagen, although he subsequently worked mainly in Sweden. Other painters of Swedish origin were Alexander Roslin, who worked throughout Europe, and Georg Desmarées, who settled in Bavaria. The Scandinavian Rococo has a distinctive flavour that is also detectable in the work of two important miniaturists of the period, Niclas Lafrensen and Cornelius Höyer. At the close of the century the paintings of Jens Juel in Denmark bridge the transition from Rococo to Neoclassicism.

Neoclassical and Romantic

NEOCLASSICISM

Neoclassicism was a widespread and influential movement in painting and the other visual arts that began in the 1760s, reached its height in the 1780s and '90s, and lasted until the 1840s and '50s. In painting it generally took the form of an emphasis on austere linear design in the depiction of classical themes and subject matter, using archaeologically correct settings and costumes.

Neoclassicism arose partly as a reaction against the sensuous and frivolously decorative Rococo style that had dominated European art from the 1720s on. But an even more profound stimulus was the new and more scientific interest in classical antiquity that arose in the 18th century. Neoclassicism was given great impetus by new archaeological discoveries, particularly the exploration and excavation of the buried Roman cities of Herculaneum and Pompeii (the excavations of which began in 1738 and 1748, respectively). And from the second decade of the 18th century on, a number of influential publications by Bernard de Montfaucon, Giovanni Battista Piranesi, the Comte de Caylus, and Robert Wood provided engraved views of Roman monuments and other antiquities and further quickened interest in the classical past. The new understanding distilled from these discoveries and publications in turn enabled European scholars for the first time to discern separate and distinct chronological periods in Greco-Roman art, and this new sense of a plurality of ancient styles replaced the older, unqualified veneration of Roman art and encouraged a dawning interest in purely Greek antiquities. The German scholar Johann Joachim Winckelmann's writings and sophisticated theorizings were especially influential in this regard. Winckelmann saw in Greek sculpture "a noble simplicity and quiet grandeur" and called for artists to imitate Greek art. He claimed that in doing so such artists would obtain idealized depictions of natural forms that had been stripped of all transitory and individualistic aspects, and their images would thus attain a universal and archetypal significance.

Neoclassicism as manifested in painting was initially not stylistically distinct from the French Rococo and other styles that had preceded it. This was partly because, whereas it was possible for architecture and sculpture to be modeled on prototypes in these media that had actually survived from classical antiquity, those few classical paintings that had survived were minor or merely ornamental works—until, that is, the discoveries made at Herculaneum and Pompeii. The earliest Neoclassical painters were Joseph-Marie Vien, Anton Raphael Mengs, Pompeo Batoni, Angelica Kauffmann, and Gavin Hamilton; these artists were active during the 1750s, '60s, and '70s. Each of these painters, though they may have used poses and figural arrangements from ancient sculptures and vase paintings, was strongly influenced by preceding stylistic trends. An important early Neoclassical work such as Mengs's "Parnassus" (1761; Villa Albani, Rome) owes much of its inspiration to 17th-century classicism and to Raphael for both the poses of its figures and its general composition. Many of the early paintings of the Neoclassical artist Benjamin West derive their compositions from works by Nicolas Poussin, and Kauffmann's sentimental subjects dressed in antique garb are basically Rococo in their softened, decorative prettiness. Mengs's close association with Winckelmann led to his being influenced by the ideal beauty that the latter so ardently expounded, but the church and palace ceilings decorated by Mengs owe more to existing Italian Baroque traditions than to anything Greek or Roman.

A more rigorously Neoclassical painting style arose in France in the 1780s under the leadership of Jacques-Louis David. He and his contemporary Jean-François-Pierre Peyron were interested in narrative painting rather than the ideal grace that fascinated Mengs. Just before and during the French Revolution, these and other painters adopted stirring moral subject matter from Roman history and celebrated the values of simplicity, austerity, heroism, and stoic virtue that were traditionally associated with the Roman Republic, thus drawing parallels between that time and the contemporary struggle for liberty in France. David's history paintings of the "Oath of the Horatii" (1784; Louvre, Paris) and "Lictors Bringing to Brutus the Bodies of His Sons" (1789; Louvre) display a gravity and decorum deriving from classical tragedy, a certain rhetorical quality of gesture, and patterns of drapery influenced by ancient sculpture. To some extent these elements were anticipated by British and American artists such as Hamilton and West, but in David's works the dramatic confrontations of the figures are starker and in clearer profile on the same plane, the setting is more monumental, and the diagonal compositional movements, large groupings of figures, and turbulent draperies of the Baroque have been almost entirely repudiated. This style was ruthlessly austere and uncompromising, and it is not surprising that it came to be associated with the French Revolution (in which David actively participated).

Neoclassicism as generally manifested in European painting by the 1790s emphasized the qualities of outline and linear design over those of colour, atmosphere, and effects of light. Widely disseminated engravings of classical sculptures and Greek vase paintings helped determine this bias, which is clearly seen in the outline illustrations made by the British sculptor John Flaxman in the 1790s for editions of the works of Homer, Aeschylus, and Dante. These illustrations are notable for their drastic and powerful simplification of the human body, their denial of pictorial space, and their minimal stage setting. This austere linearity when depicting the human form was adopted by many other British figural artists, including the Swiss-born Henry Fuseli and William Blake, among others.

Neoclassical painters attached great importance to depicting the costumes, settings, and details of their classical subject matter with as much historical accuracy as possible. This worked well enough when illustrating an incident found in the pages of Homer, but it raised the question of whether a modern hero or famous person should be portrayed in classical or contemporary dress. This issue was never satisfactorily resolved, except perhaps in David's brilliantly evocative portraits of sitters wearing the then-fashionable antique garb, as in his "Portrait of Madame Récamier" (1800; Louvre).

Classical history and mythology provided a large part of the subject matter of Neoclassical works. The poetry of Homer, Virgil, and Ovid, the plays of Aeschylus, Sophocles, and Euripides, and history recorded by Pliny, Plutarch, Tacitus, and Livy provided the bulk of classical sources, but the most important single source was Homer. To this general literary emphasis was added a growing interest in medieval sources, such as the pseudo-Celtic poetry of Ossian, as well as incidents from medieval history, the works of Dante, and an admiration for medieval art itself in the persons of Giotto, Fra Angelico, and others. Indeed, the Neoclassicists differed strikingly from their academic predecessors in their admiration of Gothic and Quattrocento art in general, and they contributed notably to the positive reevaluation of such art.

Finally, it should be noted that Neoclassicism coexisted throughout much of its later development with the seemingly obverse and opposite tendency of Romanticism. But far from being distinct and separate, these two styles intermingled with each other in complex ways; many ostensibly Neoclassical paintings show Romantic tendencies, and vice versa. This contradictory situation is strikingly evident in the works of the last great Neoclassical painter, Jean-Auguste-Dominique Ingres, who painted sensuous Romantic female nudes while also turning out precisely linear and rather lifeless historical paintings in the approved Neoclassical mode.

Britain. Hamilton—Scottish painter, archaeologist, and dealer—spent most of his working life in Rome, and his paintings include two series of large and influential canvases of Homeric subjects. West and the Swiss-born Kauffmann were the most consistent exhibitors of history pieces in London during the 1760s. James Barry and Fuseli also were important. Blake, poet and painter, was a Neoclassicist to some extent.

France. As well as being a painter, Vien was a friend of the archaeologist Caylus and a director of the French

Academy in Rome. This generation also included Jean-Baptiste Greuze, who painted a few classical history subjects as well as the scenes from contemporary life for which he is best known; Jean-Jacque Lagrenée the Elder, like Vien a director of the French Academy in Rome; and Nicolas-Guy Brenet.

The outstanding and most influential of all French Neoclassicists and one of the major artists in Europe was Vien's pupil Jacques-Louis David. David's early works are essentially Rococo, and his late works also revert to early 18th-century types; his fame as a Neoclassicist rests on paintings of the 1780s and '90s. After winning the Prix de Rome of the French Academy in 1774 (important in the history of French painting because it awarded a stay in Rome, where winners studied Italian paintings firsthand), he was in that city in 1775–81, returning there in 1784 to paint "Oath of the Horatii." David's contemporaries, or near-contemporaries, included Jean-Germain Drouais, whose history paintings almost equaled David's own in severity and intensity.

The slightly younger generation of painters included Jean-Baptiste Regnault, Louis-Léopold Boilly, and Louis Gauffier. They were followed by a more important group that included Pierre-Paul Prud'hon. Prud'hon blended in his paintings a mild classicism and the lyrical mood and soft lights of Correggio; he was patronized by the empresses Josephine and Marie-Louise. Baron Pierre-Narcisse Guérin painted in a style close to the Neoclassicism of David, although he was not one of his pupils.

David's pupils

Of David's pupils, three became well-known and one became very famous. Baron François-Pascal-Simon Gérard had a high reputation as a portraitist under both Napoleon and Louis XVIII. Antoine-Jean Gros executed many large Napoleonic canvases and after David's death was the leading Neoclassicist in France. Anne-Louis Girodet de Roucy, known as Girodet-Trioson, won a Prix de Rome but stopped painting after 1812 when he inherited a fortune and turned to writing. The famous pupil was Ingres, who was important as a Neoclassicist in his subject paintings but not in his portraits.

Germany and Austria. Mengs was born in Aussig in Bohemia (modern Ústí nad Labem, Czech Republic) in 1728, the son of the court painter there. He was himself appointed Dresden court painter in 1745. In 1755 he met Winckelmann, and subsequently he became a prominent figure in Roman Neoclassical circles. Mengs is important both as a painter and as a theorist. Apart from him, Germany's and Austria's main contribution to Neoclassicism was theoretical, not practical, however. The early Neoclassicists included Cristoph Unterberger; Anton von Maron, who married Mengs's sister; and Friedrich Heinrich Füger. After Unterberger, the most interesting painter was Johann Heinrich Wilhelm Tischbein, who executed both portraits and subject pieces. He was a director of the art academy in Naples and supervised the publication of engravings of the Greek vases in the collection of Sir William Hamilton, the British ambassador to Naples, who was a notable connoisseur.

The German painter Asmus Jacob Carstens worked in Berlin and was a professor at the Berlin Academy. Members of his artistic circle included the painters Karl Ludwig Fernow, Eberhard Wächter, Joseph Anton Koch (who was the most outstanding of this German group), and Gottlieb Schick.

Italy. One of the earliest Neoclassicists and one of the foremost painters of his generation in Italy was Batoni. His style blends Rococo with Neoclassical elements, and his work includes classical subject pieces as well as portraits in contemporary dress, the sitter posing with antique statues and urns and sometimes amid ruins. The painter Domenico Corvi was influenced by both Batoni and Mengs and was important as the teacher of three of the leading Neoclassicists of the next generation: Giuseppe Cades, Gaspare Landi, and Vincenzo Camuccini. These artists worked mostly in Rome, the first two making reputations as portraitists, Landi especially being noted for good contemporary groups.

Rome was indeed the city where the principal Italian painters of this period were most active. One such was Felice Giani, whose many decorations include Napoleonic palaces there and elsewhere in Italy (especially Faenza) and in France.

Major Italian Neoclassicists outside Rome

Important painters outside Rome include Andrea Appiani the Elder in Milan, who became Napoleon's official painter and executed some of the best frescoes in northern Italy. He was also a fine portraitist. One of his pupils was Giuseppe Bossi. Another leading Lombard painter was Giovanni Battista dell'Era, whose encaustic paintings were bought by Catherine the Great and others. Other good examples of Neoclassical decorative schemes outside Rome are in Florence (Pitti Palace) by the Florentine Luigi Sabatelli and by Pietro Benvenuti, who was born at Arezzo, and in Venice (Palazzo Reale) by Giuseppe Borsato, who was born in that city and was both painter and architect. Another painter of the time, though only given to a mildly Neoclassical style, was Domenico Pellegrini, born near Bassano, who traveled widely. The principal Neoclassicists in the south were the Sicilians Giuseppe Velasco, who did important frescoes in palaces in Palermo, and Giuseppe Errante.

Other countries. The main Danish painter who produced original Neoclassical works was Nicolai Abraham Abildgaard. Other Danish painters include Abildgaard's and David's pupil Christoffer Wilhelm Eckersberg. David was very influential in Brussels, where he retired late in life. The paintings of his Belgian pupil François-Joseph Navez, for example, are pure French Neoclassicism. The two main Neoclassical artists in The Netherlands were Humbert de Superville and Jan Willem Pieneman. The principal Neoclassicist in Spain was José de Madrazo y Agudo. (D.I./Ed.)

ROMANTICISM

Romanticism is a term loosely used to designate numerous and diverse changes in the arts during a period of more than 100 years (roughly, 1760–1870), changes that were in reaction against Neoclassicism (but not necessarily the classicism of Greece and Rome) or against what is variously called the Age of Reason, the Augustan Age, the Enlightenment, or 18th-century materialism. In the sense of a personal temperament Romanticism had always existed, but in the sense of an aesthetic period it signified works of art whose prime impulse and effect derived from individual rather than collective reactions. Romanticism can generally be said to have emphasized the personal, the subjective, the irrational, the imaginative, the spontaneous, the emotional, and even the visionary and transcendental in works of art. The Romantic movement first developed in northern Europe with a rejection of technical standards based on the classical ideal that perfection should be attained in art.

It was writers and poets who gave initial expression to Romantic ideas; painters, while subject to similar feelings, acquired fundamental inspiration from the literature of the period. There was an increasing awareness generally of the way the various arts interacted. The Frenchman Eugène Delacroix and the German Philipp Otto Runge explored the implications of musical analogies for painting, and everywhere writers, artists, and composers could be found in close association.

Romantic critics agreed that experience of profound inner emotion was the mainspring of creation and appreciation of art. Received ideas, and especially aesthetic values sanctioned by the authority of official institutions, were distrusted, and the individual was pitted against society. The artist asserted the right to evolve his own criteria of beauty and in so doing encouraged a new concept of artistic genius. The genius whom the Romantics celebrated was one who refused to conform, who remained defiantly independent of society, and whose chief virtues were novelty and sincerity. This sometimes led to bizarre and extravagant projects in which the intention to shock, excite, and involve struck a melodramatic, almost hysterical note that failed to convince by its very lack of restraint.

As in the literature of the period, tragic themes predominated in Romantic painting, and interest turned sharply from classical history and mythology to medieval subjects, although an interest in the primitive was sometimes com-

mon to both. The fascination with the Middle Ages combined with strong nationalist tendencies, disposing artists to a concern with the history and folklore of their own countries. At the same time they often sought themes or styles that were distant in place as well as time. Accounts of foreign travel and the literary works of Dante, Shakespeare, Byron, Goethe, Sir Walter Scott, and the supposed Celtic bard Ossian greatly influenced painters. Study of medieval culture imbued some painters with a Christian ideal of simplicity and moral integrity.

A salient feature of Romantic sensibility was awareness of the beauties of the natural world. Artists identified their personal feelings with nature's changing aspects. An almost reverential affection, animated by the belief that the divine mind was immanent in nature, engendered at times a Christian or theistic naturalism. The artist was seen as the interpreter of hidden mysteries, to which end imaginative insight must combine with absolute fidelity and sincerity. In Britain and Germany especially, the moral implications inherent in the appreciation of natural or artistic beauty tended to outweigh aesthetic considerations. Interest in transitory phenomena led painters to devote themselves to an accurate study of light and atmosphere and their effects on the landscape. Concern to preserve the spontaneity of the immediate impression brought about a revolution in painterly technique, with the rapid notation of the sketch carried into the final conception. Whether emphasizing expressive or purely visual considerations, the landscape paintings of the period display dazzling colour.

Curiosity about the external world and a spirit of what might be called scientific inquiry led many painters to explore the minutiae of nature. Technological advance also excited artistic interest, though painting was affected less than architecture and the decorative arts; and the humanitarian sympathy and generosity so vital to the Romantic spirit gradually effected a reconciliation between art and life. The political and social upheavals of the 19th century involved many painters in revolutionary movements and stimulated a solicitude toward the helpless and downtrodden that found most passionate and powerful expression in the works executed during and immediately after the Revolutions of 1848.

British early Romantic painters in Rome

Britain. In the late 1760s and '70s a circle of British painters in Rome had already begun to find academic precepts inadequate. James Barry, the brothers John and Alexander Runciman, John Brown, George Romney, and the Swiss-born Henry Fuseli favoured themes—whether literary, historical, or purely imaginary—determined by a taste for the pathetic, bizarre, and extravagantly heroic. Mutually influential and highly eclectic, they combined, especially in their drawings, the linear tensions of Italian Mannerism with bold contrasts of light and shade. Though never in Rome, John Hamilton Mortimer had much in common with this group, for all were participants in a move to found a national school of narrative painting. Fuseli's affiliations with the German Romantic *Sturm und Drang* writers predisposed him, like Flaxman, toward the "primitive" heroic stories of Homer and Dante. Flaxman himself, in the two-dimensional linear abstraction of his drawings, a two-dimensionality implying rejection of Renaissance perspective and seen for instance in the expressive purity of "Penelope's Dream" (1792–93), had important repercussions throughout Europe.

Blake

William Blake absorbed and outstripped the Fuseli circle, evolving new images for a unique private cosmology, rejecting oils in favour of tempera and watercolour, and depicting, as in "Pity" (1795; Tate Gallery, London), a shadowless world of soaring, supernatural beings (Plate 19). His passionate rejection of rationalism and materialism, his scorn for both Sir Joshua Reynolds and the Dutch Naturalists, stemmed from a conviction that "poetic genius" could alone perceive the infinite, so essential to the artist since "painting, as well as poetry and music, exists and exults in immortal thoughts." The spiritual, symbolical expression of Blake's complex sympathies, his ability to recognize God in a single blade of grass, inspired Samuel Palmer, who, with his friend Edward Calvert, extracted from nature a visionary world of exquisite, though short-lived, intensity.

Empiricism and acceptance of the irrational, however, were not mutually exclusive, and each profoundly affected attitudes toward nature. Susceptible to the ideas of Blake and other radical theorists and animated by a growing spirit of inquiry into natural phenomena, painters slowly abandoned the picturesque desire to compose and became willing to be moved, awestruck, and terrified by nature unadorned. Early artists of the sublime, such as Alexander Cozens or Francis Towne, worked largely in watercolours and solved the problem of scale by abstraction—use of broad areas of colour to suggest the vast scope of natural forces—an approach developed by Thomas Girtin and John Sell Cotman.

By the early 19th century, the watercolourist John Varley was echoing current practice when he told his pupils John Linnell, William Mulready, and William Henry Hunt: "Go to nature for everything." But already two outstanding British landscape painters, John Constable and J.M.W. Turner, were going still further. Both men, while admiring the classical landscapes of Claude Lorrain and Poussin, believed that personal feeling was the mainspring of artistic activity and felt an almost mystical sympathy for the natural world. They made atmosphere almost palpable and painted everything from clouds to lichens with astonishing technical diversity. Constable considered himself before all else a "natural" painter and sought, in his own words, to capture "light—dews—breezes—bloom—and freshness" with scientific precision and deepest affection. For Constable, light clarified and enlivened, and his nostalgia for the Suffolk countryside is personal and explicit (Plate 20). With Turner, light increasingly diffused the objects illuminated, and only a more literary expression satisfied his concept of the sublime, drawing him to mountain grandeur, raging seas, storms, and conflagrations (Plate 21). The technical innovations of these two men were better understood in France than in Britain; even John Ruskin's passionate defense of Turner, with its emphasis on absolute fidelity to nature, helped deflect Turner's and Constable's successors onto a very different course.

Constable considered a "natural" painter

George Stubbs's anatomical studies and accurate delineations of animals were echoed a generation later by Thomas Bewick's bird studies, themselves harbingers of the drawings of Edwin Landseer and Ruskin's closely observed renderings of naturalistic detail. Stubbs's empathy for the animal world reemerged in the work of James Ward, together with an exultation in the power of nature, shared by Philip James de Loutherbourg. Demand for information about distant places partially superseded the taste for picturesque European scenes, and following William Hodges, who accompanied Captain James Cook's second voyage (1772–75), such painters as Richard Parkes Bonington, Samuel Prout, John Frederick Lewis, and Edward Lear traveled widely, recording scenes of historic or exotic interest.

In portraiture an interest in extremes of mood found most eloquent expression in the work of Sir Thomas Lawrence, who combined in portraits such as those of Richard Payne Knight (1794; Whitworth Art Gallery, Manchester) and Pope Pius VII (1819; Royal Collection, Windsor Castle) brilliant freedom of handling, at times approaching exhibitionism, with dramatic expression and setting, at times almost melodramatic.

History painting

History painting, too, was transformed: Bonington's "Henri III and the English Ambassador" (1827–28; Wallace Collection, London), while testifying to a sustained delight in the medieval world, already betrays commensurate interest in period detail and the finer points of human insight. The authentic, domestic treatment of biblical themes at the hands of William Dyce and the Pre-Raphaelites (see below) contrasts sharply with the earlier apocalyptic fantasies of John Martin and Francis Danby. Inspired by David Wilkie's mellow, unassuming representation of country life subject matter, William Mulready turned to contemporary scenes of daily life, adopting the brilliant palette that distinguished British painting for the next half-century. The high Victorian Age saw much narrative painting, a genre that was practiced with accurate and sympathetic observation, from the panoramic activ-

ity of William Powell Frith's "Derby Day" (1858; Tate Gallery) to such intimate glimpses of reality as "The Travelling Companions" (1862; City Museum and Art Gallery, Birmingham), by Augustus Egg. Painting as a vehicle for social or moral comment was provided by Sir Luke Fildes and Frank Holl, in whose work a tendency to sentimentality is redeemed by a genuine regard for the sufferings of the poor. In the 1850s the Pre-Raphaelites gave expression to the painting of contemporary life with such memorable images as "The Blind Girl" (1856; City Museum and Art Gallery, Birmingham), by John Everett Millais, or "The Stonebreaker" (1857–58; Walker Art Gallery, Liverpool), by John Brett.

The Pre-Raphaelite movement

The Pre-Raphaelite movement, echoing that of the Nazarenes (a group of religiously minded painters who sought to revive medieval workshop practices; see below), reiterated many earlier Romantic ideals. Literary inspiration and a passion for the Middle Ages were tempered for the Pre-Raphaelites by a moral outlook that recoiled from sophistication and virtuosity and demanded rigorous studies from natural life. These painters handled literary, historical, biblical, and contemporary themes with the same sincerity and fidelity that yielded the sparkling precision of Pre-Raphaelite landscape. Their earnest pursuit of truth, whether in depicting painful social realities or concentrating on the foreground blades of grass in a landscape, entailed a denial of many orthodox artistic pleasures. Together with Ford Madox Brown, the Pre-Raphaelites sustained the devotion to colour and light in painting that underlies the finest endeavours of English Romanticism.

Germany. In Germany also there was a reaction against classicism and the academies, and, as elsewhere, it involved all aspects of the arts. Again, as elsewhere, theory preceded practice: *Herzensergiessungen eines kunstliebenden Klosterbruders* ("Effusions of an Art-Loving Monk"), by Wilhelm Heinrich Wackenroder, had an immediate and widespread influence upon its publication in 1797. Wackenroder advocated a Christian art closely related to the art of the early German masters and provided the artist with a new role as interpreter of divine inspiration through his own feelings.

The painter Philipp Otto Runge had been reared on 17th-century German mysticism, and he proved susceptible to the ideas of writers such as Wackenroder when introduced to them in Dresden at the very end of the 18th century. In Dresden he formed a close association with the leading German landscape painter Caspar David Friedrich (Plate 21). Like Friedrich he was fascinated by the potential symbolic and allegorical power of landscape, which he used as a vehicle for religious expression. His vision of nature was pantheistic (as was Friedrich's), and in his portraits his aim was to capture the soul of the individual as part of the universal soul of nature. "The Artist's Parents and Children" (1806; Hamburger Kunsthalle, Hamburg) reflects not only his constant search for truth but also his admiration for the early German masters, through whose work he was made aware of the expressive power of line and colour. His interest in the German past, including folklore and fairy tales, was reflected in a bizarre fairylike quality in much of his work (*e.g.,* "Night," 1803), and it was this quality that was taken up and popularized by his two most important followers, Moritz von Schwind and Adrian Ludwig Richter, in whose hand the intensity of the first generation declined into popular genre paintings (usually small pictures depicting everyday life, as opposed to some idealized existence) and the comfortable Romanticism of the Biedermeier period (1815–48).

Friedrich was a deeply religious man whose vision demanded complete subjection to the spirit of God in nature; in suggesting through landscape the eternal presence of the Creator, he intended to induce in the beholder a state of religious awe. Among his pupils was Carl Gustav Carus, a physician, philosopher, and self-taught painter whose chief contribution was as a theorist; *Neun Briefe über Landschaftsmalerei* (1831; "Nine Letters on Landscape Painting") elucidates and expands the ideas of Friedrich, adding Carus' own more-scientific approach to natural phenomena. Other important painters influenced by Friedrich were Ernst Ferdinand Oehme, a landscape painter, and Georg Friedrich Kersting, who captured in his stark interiors something of the master's atmosphere of silent worship. However, two other pupils of Friedrich subsequently abandoned tragic landscapes; one, the Norwegian Johan Christian Dahl, reverted to naturalism; the other, Karl Blechen, joined the Romantic realists.

Whereas Runge, Friedrich, and their followers interpreted Wackenroder in a highly personal way, others were inspired to communal activity. A number of young painters in Vienna founded in 1809 a group they called the Guild of St. Luke. The founding members were Johann Friedrich Overbeck (their leader), Franz Pforr, Joseph Wintergerst, Joseph Sutter, and Georg Ludwig Vogel. In 1810 they moved to Rome, where they were soon joined by Peter von Cornelius, Julius Schnorr von Carolsfeld, Friedrich Olivier, the brothers Philipp and Johannes Veit, Wilhelm von Schadow, Johann Evangelist Scheffer von Leonhartshoff, and Josef von Führich. Their semimonastic existence occasioned the nickname Nazarenes.

The Guild of St. Luke

In general, their highest aspirations—toward monumental history painting—produced the least successful results, and they came closest to realizing their intentions on a small scale in highly finished watercolours and drawings, as in Overbeck's "The Raising of Jairus' Daughter" (1814). Only Joseph Anton Koch and Cornelius, who were both older and more experienced, achieved great vigour in their history paintings, combining medievalizing tendencies with the powerful classicism of Carstens (see above *Neoclassicism: Germany and Austria*), as seen in Cornelius' "The Recognition of Joseph by His Brethren" (1815–16; National Gallery, Berlin). Even Overbeck, an articulate leader and a lucid draftsman, could not escape, in his "Joseph Being Sold by His Brethren" (1816–17; National Gallery, Berlin), the self-conscious naïveté common to many of the Nazarenes. This naïveté is also noticeable in Pforr's "The Entry of the Emperor Rudolf of Habsburg into Basel in 1273" (*c.* 1809; Städelsches Kunstinstitut, Frankfurt am Main) and Schnorr's "The Procession of the Three Magi" (1819; Museum of Fine Art, Leipzig). Alfred Rethel, a late arrival, however, manages to avoid such an effect in his haunting "King David with His Harp" (*c.* 1831; Museum of Art, Düsseldorf). Not long afterward there was a move toward the more dramatic, though no less nostalgic, approach of von Schadow and his pupil Karl Friedrich Lessing.

Portraiture required less self-consciousness than history painting, and there are a number of highly sensitive portraits, mainly of their friends, by Overbeck, Schnorr, Scheffer von Leonardshoff, and Carl Philipp Fohr ("Portrait of Wilhelm von Schadow" [1818; Museum of the Palatinate, Heidelberg]). The Nazarenes' greatest contribution, however, was to landscape painting: inspired by the heroic landscapes of Koch (*e.g.,* "Bernese Oberland" [1816; Gallery of Modern Paintings, Staatliche Kunstsammlungen Dresden]), by the German "primitives," and by their own concept of truth to nature, they renounced the conventional Italianate solution and turned instead to the countryside around them and to memories of Germany and German painting. As the movement gathered momentum, the possibilities for development expanded, and the Nazarene landscape was valuable to later painters of the Biedermeier period and to painters of naturalistic landscape, Romantic realism, and secular historical subjects.

France. The French Revolution greatly stimulated interest in the depiction of contemporary events, although richly documented and highly detailed paintings of topical patriotic events were being painted in London by West and John Singleton Copley even before the Revolution. Encouraged by David's example, however, painters in France sought to represent authentically the crucial moments of their own time. Napoleon I enthusiastically endorsed this awareness of modern heroism and demanded pictorial celebration of the glorious achievements of the empire. David recorded the ceremonies of the imperial court with scrupulous precision. Napoleon's potent hold on the artistic imagination is well illustrated by Gros's "Napoleon Visiting the Pesthouse at Jaffa" (1804; Louvre), where he

Depiction of contemporary events



"Aurora," ceiling fresco by Guido Reni, 1613–14. In the Casino Rospigliosi, Rome.

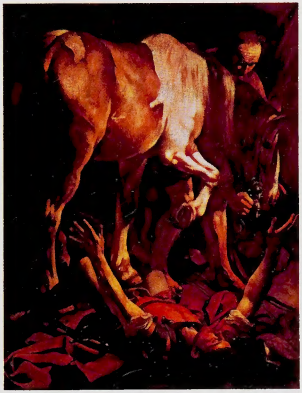


Baroque painting in Italy

"View of the Lagoon," oil painting by Francesco Guardi (1712–93). In the Museo di Castelvecchio, Verona, Italy. 33 × 51 cm.



"The Triumph of the Name of Jesus," ceiling fresco by Giovanni Battista Gaulli (Baciccia), 1674–79. In the Gesù, Rome.



"The Conversion of St. Paul," oil painting by Caravaggio (1573–1610). In Sta. Maria del Popolo, Rome. 2.3 × 1.8 m.

Plate 13: SCALA/Art Resource—EB Inc.



"Young Woman with a Water Jug," oil painting by Jan Vermeer, *c.* 1658–60. In the Metropolitan Museum of Art, New York City. 46 × 41 cm.



"Still Life," oil on wood by Willem Claesz Heda, 1634. In the Museum Boymans-van Beuningen, Rotterdam. 43 × 57 cm.



"The Rape of the Daughters of Leucippus," oil painting by Peter Paul Rubens, 1617. In the Alte Pinakothek, Munich. 2.2 × 2.1 m.



"The Smokers," oil on wood by Adriaen Brouwer, *c.* 1630. In the Metropolitan Museum of Art, New York City. 46 × 38 cm.



"Waterfall," oil painting by Jacob van Ruisdael, *c.* 1670? In the Uffizi, Florence. 52.3 × 61.7 cm.

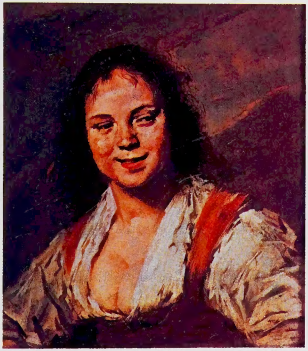
Plate 14: By courtesy of (top left) the Museum Boymans-van Beuningen, Rotterdam, (top right, bottom left) the Metropolitan Museum of Art, New York City, (top right) gift of Henry G. Marquand, 1889, (bottom left) the Michael Friedsam Collection, 1931, (centre right) the Alte Pinakothek, Munich, (bottom right) SCALA/Art Resource, New York City



"Calling of St. Matthew," oil painting by Hendrik Terbrugghen, *c.* 1617. In the Centraal Museum, Utrecht, The Netherlands. 1 × 1.4 m.



"Portrait of Charles I Hunting," oil painting by Sir Anthony Van Dyck, 1635. In the Louvre, Paris. 2.8 × 2.2 m.



"Gypsy Girl," oil on wood by Frans Hals, *c.* 1628–30. In the Louvre, Paris. 57.8 × 52.1 cm.

Baroque painting in the Low Countries

"Jacob Blessing the Sons of Joseph," oil painting by Rembrandt, 1656. In the Staatliche Kunstsammlungen, Kassel, West Germany. 1.8 × 2.1 m.



Plate 15 By courtesy of (top left) the Centraal Museum, Utrecht, Neth.; (bottom) the Staatliche Kunstsammlungen, Kassel, W Ger.; photographs, (top right) Giraudon/Art Resource—EB Inc.; (centre left) Fernand Bourges—LIFE Magazine © Time Inc.



"La Partie Quarrée," oil painting by Antoine Watteau, 1713–14. In the collections of the Fine Arts Museums of San Francisco. 50 × 65.2 cm.



"St. John on Patmos," oil painting by Nicolas Poussin, 1645–50. In the Art Institute of Chicago. 1 × 1.4 m.

"The Lamentation over St. Sebastian," oil painting by Georges de La Tour, 1645. In the Staatliche Museen Preussischer Kulturbesitz, West Berlin. 1.6 × 1.3 m.



"Cupid a Captive," oil painting by François Boucher, 1754. In the Wallace Collection, London. 1.7 m × 85 cm.

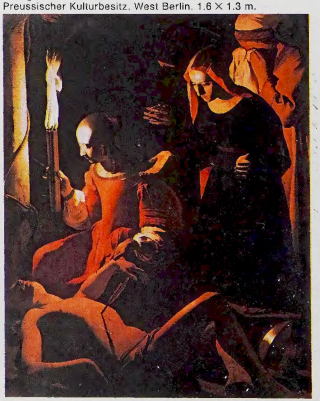


Plate 16. By courtesy of (centre right) the Art Institute of Chicago; (bottom right) the Staatliche Museen Preussischer Kulturbesitz, West Berlin, by permission of (left) the trustees of the Wallace Collection, London; photograph, (top) The Fine Arts Museums of San Francisco, Mildred Anna Williams Collection



"The Embarkation of the Queen of Sheba," oil painting by Claude Lorrain, 1648. In the National Gallery, London. 1.5 × 1.9 m.



"Mrs. Sheridan," oil painting by Thomas Gainsborough, *c.* 1785. In the National Gallery of Art, Washington, D.C. 2.2 × 1.5 m.



Baroque painting in France and Britain

"The Countess's Morning Levée," oil painting, part of the "Marriage à la Mode" series, by William Hogarth, 1743–45. In the National Gallery, London. 69 × 90 cm.





"Family of Country People," oil painting by Louis Le Nain, *c.* 1640. In the Louvre, Paris. 1.1 × 1.6 m.

"The Grace," oil painting by J.-B.-S. Chardin, 1740. In the Louvre, Paris. 48 × 40 cm.

Plate 17: By courtesy of (top left, centre) the trustees of the National Gallery, London, (top right) the National Gallery of Art, Washington, D.C., Andrew Mellon Collection; photographs, (bottom left, bottom right) Telarc—Giraudon/Art Resource—EB Inc.



"The Holy House of Nazareth," oil painting by Francisco de Zurbarán, *c.* 1630. In the Cleveland Museum of Art. 1.7 × 2.2 m.

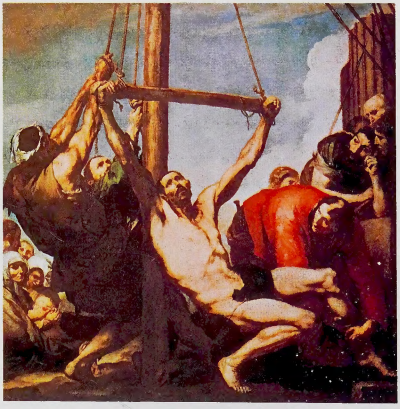


"Infanta Margarita," oil painting by Diego Velázquez, 1660. In the Prado, Madrid. 2.2 × 1.5 m.

Baroque painting in Spain



"Boys Eating Grapes and Melon," oil painting by Bartolomé Esteban Murillo, *c.* 1650. In the Alte Pinakothek, Munich. 1.5 × 1 m.

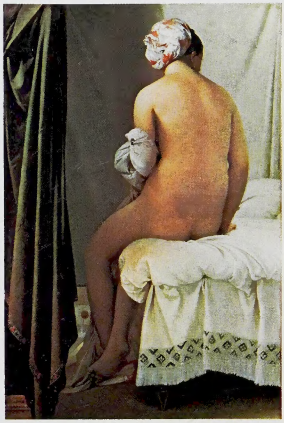


"The Martyrdom of St. Bartholomew," oil painting by José de Ribera, 1630/39. In the Prado, Madrid. 2.4 × 2.4 m.

Plate 18 By courtesy of (top left) the Cleveland Museum of Art, purchase, Leonard C. Hanna Jr. Bequest, (bottom left) the Alte Pinakothek, Munich; photographs, (top right, bottom right) Archivo Mas, Barcelona, (bottom left) Joachim Blauel



"Pity," colour print finished in pen and watercolour by William Blake, 1795. In the Tate Gallery, London. 44 × 55 cm.



"Valpinçon Bather," oil painting by Jean-Auguste-Dominique Ingres, 1808. In the Louvre, Paris. 1.4 m × 99 cm.



"Goethe in the Campagna," oil painting by Johann Heinrich Wilhelm Tischbein, 1787. In the Städelsches Kunstinstitut, Frankfurt am Main, Germany. 1.6 × 2 m.

Neoclassical and early Romantic painting



"Oath of the Horatii," oil painting by Jacques-Louis David, 1784. In the Louvre, Paris. 4.3 × 3.4 m.

Plate 19: By courtesy of (top left) the trustees of the Tate Gallery, London, (centre) the Städelsches Kunstinstitut, Frankfurt am Main, W.Ger., photographs (top left) G. Roberton, A.C. Cooper Ltd., (top right) Giraudon/Art Resource—EB Inc., (centre) Joachim Blauel, (bottom) Marc Garanger



"Boatbuilding near Flatford Mill," oil painting by John Constable, 1815. In the Victoria and Albert Museum, London. 50.8 × 61.6 cm.



"The Third-Class Carriage," oil painting by Honoré Daumier, c. 1862. In the Metropolitan Museum of Art, New York City. 65 × 90 cm.



"The Death of Sardanapalus," oil painting by Eugène Delacroix, 1827. In the Louvre, Paris. 3.7 × 4.9 m.

"Souvenir de Mortefontaine," oil painting by Camille Corot, 1864. In the Louvre, Paris. 64 × 90 cm.



Plate 20: By courtesy of (top) the Board of Trustees of the Victoria and Albert Museum, London (centre right) the Metropolitan Museum of Art, New York City, bequest of Mrs. H.O. Havemeyer, 1929, the H.O. Havemeyer Collection; photographs, (centre left) SCALA/ Art Resource—EB Inc.; (bottom) Lauros—Giraudon/Art Resource—EB Inc.

Romantic and Realist painting



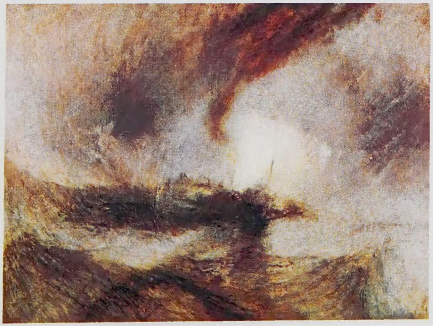
"The Raft of the Medusa," oil painting by Théodore Géricault, *c.* 1819. In the Louvre, Paris. 5 × 7.2 m.



"Man and Woman Gazing at the Moon," oil painting by Caspar David Friedrich, *c.* 1824. In the Nationalgalerie, West Berlin. 34 × 44 cm.

"The Gross Clinic," oil painting by Thomas Eakins, 1875. In the Jefferson Medical College, Thomas Jefferson University, Philadelphia. 2 × 2.5 m.





"Snow Storm—Steam-Boat off a Harbour's Mouth . . . ," oil painting by J.M.W. Turner, 1842. In the Tate Gallery, London. 92 cm × 1.1 m.

Plate 21: By courtesy of (centre left) the Staatliche Museen Preussischer Kulturbesitz Nationalgalerie, West Berlin, (centre right) the Jefferson Medical College, Thomas Jefferson University, Philadelphia, (bottom) the trustees of the Tate Gallery, London, photographs, (top) Cliche Musees Nationaux, (centre left) Marburg/Art Resource—ER Inc., (centre right) A.J. Wyatt, (bottom left) G. Roberton, A.C. Cooper Ltd.

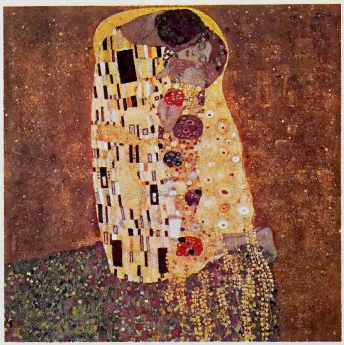
Impressionism, Postimpressionism, and Art Nouveau



"The Starry Night," oil painting by Vincent van Gogh, 1889. In the Museum of Modern Art, New York City. 74 × 92 cm.

"Mont Sainte-Victoire," oil painting by Paul Cézanne, 1904. In the Philadelphia Museum of Art. 72 × 93 cm.

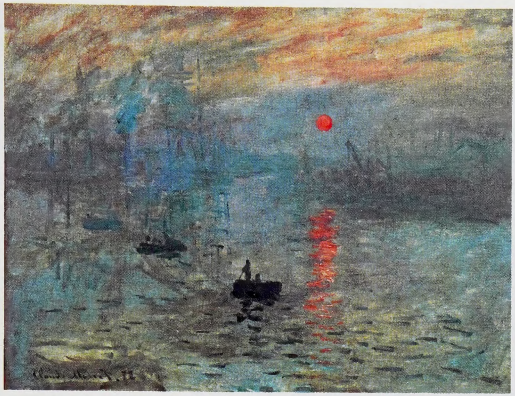




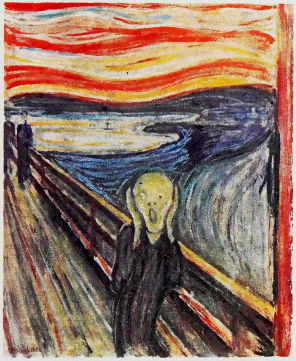
"The Kiss," oil painting by Gustav Klimt, 1908. In the Österreichische Galerie, Vienna. 1.8 × 1.8 m.

"Sunday Afternoon on the Island of La Grande Jatte," by Georges Seurat, 1884–86. In the Art Institute of Chicago. 2.1 × 3.1 m.





"Impression: Sunrise," oil painting by Claude Monet, 1872. In the Marmottan Museum, Paris. 50 × 67 cm.



"The Cry," oil and tempera on wood by Edvard Munch, 1893. In the Nasjonalgalleriet, Oslo. 91 × 74 cm.



"Le Moulin de la galette," oil painting by Pierre-Auguste Renoir, 1876. In the Louvre, Paris. 1.3 × 1.8 m.

"Te Rerioa" ("The Dream"), oil painting by Paul Gauguin, 1897. In the Courtauld Institute Galleries, London. 96 cm × 1.3 m.

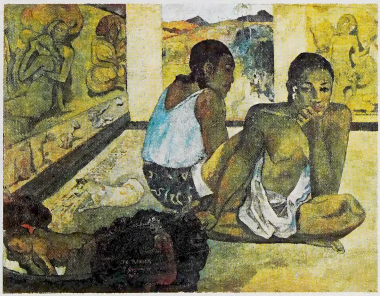


Plate 22: By courtesy of (top) the Museum of Modern Art, New York, acquired through the Lillie P. Bliss Bequest, (centre left) the Philadelphia Museum of Art, George W. Elkins Collection, (centre right) Rudolf Zinnpel, Breitenfurt, Austria, (bottom) the Art Institute of Chicago; photographs, (centre left) A.J. Wyatt, (centre right) Österreichische Galerie, Vienna—Erwin Meyer

Plate 23: By courtesy of (top) the Musee Marmottan, Paris, permission S.P.A.D.E.M. 1972 by French Reproduction Rights, Inc., (centre left) the Nasjonalgalleriet, Oslo, (bottom) the Courtauld Institute Galleries, London; photographs, (top) Photo Routhier, Paris, (centre left) O. Vaering, (centre right) Marc Garanger

"Violin and Palette," oil painting by Georges Braque, 1910. In the Solomon R. Guggenheim Museum, New York City. 93 × 43 cm.





"Composition," oil painting by Joan Miró, 1933. In the Wadsworth Atheneum, Hartford, Connecticut. 1.3 × 1.6 m.

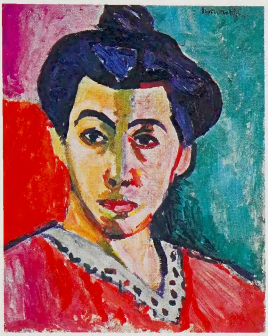
"Improvisation No. 30," oil painting by Wassily Kandinsky, 1913. In the Art Institute of Chicago. 1.1 × 1.1 m.



"Departure," oil on canvas triptych by Max Beckmann, 1932–33. In the Museum of Modern Art, New York City. Centre panel 2.2 × 1.1 m, each side panel 2.2 × 1 m.

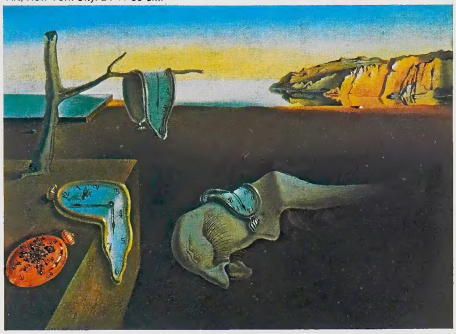


Plate 24: By courtesy of (top left) the Solomon R. Guggenheim Museum, New York, (top right) the Wadsworth Atheneum, Hartford, Connecticut, the Ella Gallup Sumner and Mary Catlin Sumner Collection, (centre right) the Art Institute of Chicago, Arthur Jerome Eddy Memorial Collection, (bottom) the Museum of Modern Art, New York

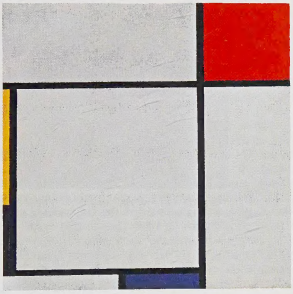


"Portrait of Madame Matisse," oil painting by Henri Matisse, 1905. In the Statens Museum for Kunst, Copenhagen. 40.5 × 32.5 cm.

"The Persistence of Memory," oil painting by Salvador Dali, 1931. In the Museum of Modern Art, New York City. 24 × 33 cm.



"Composition," oil painting by Piet Mondrian, 1929. In the Solomon R. Guggenheim Museum, New York City. 45.5 × 45.5 cm.



Painting 1900–45



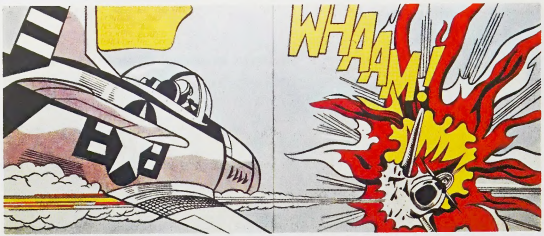
"I and the Village," oil painting by Marc Chagall, 1911. In the Museum of Modern Art, New York City. 1.9 × 1.5 m.



"The Snake-Charmer," oil painting by Henri Rousseau, 1907. In the Louvre, Paris. 1.7 × 1.8 m.

Plate 25: By courtesy of (top left) the Royal Museum of Fine Arts, Copenhagen, with permission of Mme Marguerite Duthuit, permission S.P.A.D.E.M. 1973 by French Reproduction Rights, Inc., (top right, centre right) the Museum of Modern Art, New York, (centre right) Mrs. Simon Guggenheim Fund, (centre left) the Solomon R. Guggenheim Museum, New York, gift, Katherine S. Dreier Estate; photograph, (bottom left) Marc Garanger

Contemporary painting



"Whaam," oil on two canvas panels by Roy Lichtenstein, 1963.
In the Tate Gallery, London. 1.7 × 4.1 m.



"Linda," acrylic on canvas by Chuck Close, 1975–76.
In the Pace Gallery, New York City. 2.7 × 2.1 m.



"Orange and Yellow," oil painting by Mark Rothko, 1956. In the Albright-Knox Art Gallery, Buffalo, New York. 2.3 × 1.8 m.

"Autumn Rhythm," oil painting by Jackson Pollock, 1950. In the Metropolitan Museum of Art, New York City. 2.7 × 5.3 m.



Plate 26: By courtesy of (top) the trustees of the Tate Gallery, London, (centre left) the Pace Gallery, New York City, (centre right) the Albright-Knox Art Gallery, Buffalo, gift of Seymour H. Knox, (bottom) the Metropolitan Museum of Art, New York City, George A. Hearn Fund, 1957.

is endowed with godlike authority and the humanitarian sensibility of the true Romantic hero. At the same time, other artists—such as Gérard, Girodet-Trioson, and Ingres—readily responded to the Emperor's admiration for the stories of Ossian. After the fall of Napoleon few were disposed to depict contemporary subjects. Théodore Géricault was something of an exception, but he was separated from his immediate predecessors both by temperament and by the sincerity of his approach. Individual suffering rather than collective drama is vividly portrayed in "The Raft of the Medusa" (*c.* 1819; Louvre; Plate 21). This, Géricault's masterpiece, echoes in its strenuous forms the school of Caravaggio in the 17th century. His studies of the poor, aged, and insane are realistically observed and have a sympathetic intensity unmatched before the generation of Honoré Daumier and Gustave Courbet.

The paintings of Delacroix frequently disrupted the salons of the 1820s and '30s with their tumultuous colour and emotive energy. To many young men after 1815, France appeared to settle into a bourgeois respectability that implicitly disparaged the exhilarating years of the republic and the empire. In consequence, the art of the period often seems melancholic and introverted, the discontent expressing itself in historical and exotic themes or in a passionate concern with the humble and rejected members of society. Delacroix has justly been acclaimed the leader of the Romantic school in France. His fertile imagination, embracing a novel range of literary and historical themes and fastening with a characteristic sense of the sadness of life on moments of death, defeat, and suffering, together with his prodigious technical resources exemplify Romanticism in its most obvious aspects. His vigorous handling of paint and expert use of colour values for both description and expression were important for the later development of French painting. "The Massacre at Chios" (1824; Louvre) transposes contemporary events into a realm of tragic fiction soon established unrestrainedly with such melodramatic works as "The Death of Sardanapalus" (1827; Louvre), a riot of brilliant colour and ebullient forms (Plate 20).

Exoticism of setting

Delacroix's Moroccan paintings released a flood of North African subjects, although, in the hands of lesser artists—such as Eugène Fromentin, Ary Scheffer, and Eugène Devéria—the treatment is less effective. Alexandre-Gabriel Decamps, whose small canvases have a delicate, jewellike quality, provided the most refreshing variations on the theme. But Delacroix was not the first to handle Oriental subjects; Ingres had already done so with a reticence that belies the sensuous delight in "Valpinçon Bather" (1808; Louvre; Plate 19) and in "La Grande Odalisque" (1814; Louvre). Early in his career Ingres made notable contributions to the historical genre with episodes from medieval French history painted in a style of linear purity that parallels the methods of Flaxman and Blake in Britain and the Nazarenes in Germany. Under the spell of Raphael he returned to the academic fold, but his portraits always retained that trenchant simplicity and lucid insight that make him such a memorable exponent of lyric realism. The career of Ingres and in a converse sense that of Paul Delaroche well illustrate the imprudence of too readily distinguishing between academic and Romantic artists. Delaroche, perhaps the most popular representative of the Romantic school, specialized in highly charged narratives with royal and child characters, of which "The Children of Edward" (*c.* 1830; Louvre) is a typical example, being executed with a flatness that lacks either linear or colouristic inspiration. In comparison, the work of Théodore Chassériau is animated by powerful emotional overtones reminiscent of Delacroix. "The Cossack Girl Finding the Body of Mazeppa" (1851; Museum of Fine Art, Strasbourg) shows a similarly expressive use of paint, together with poignant imagery, both characteristic of his regrettably slender oeuvre. At the end of the century, Gustave Moreau and Odilon Redon transformed these features, along with others in Louis Boulanger's work, into whimsical, haunting fantasies that delighted the Symbolist poets.

In the 1830s and '40s it was Honoré Daumier, more than any other artist, who portrayed relatively lowly members of society, expressing in numerous drawings and paintings their patient resignation (Plate 20). In contrast, his truly excoriating depiction of the weaknesses and vices of the privileged classes, particularly officialdom, often displeased authority, which had long identified Romanticism with liberalism—and with good reason. A strain of poetic realism in the 1840s, essentially Romantic in approach, gathered sudden momentum with the Revolution and short-lived republic of 1848. Jean-François Millet and Gustave Courbet depicted peasant life, investing it with a certain timeless quality. Courbet's "Stone-Breakers" (1849; destroyed during World War II) and Millet's harrowing "Quarriers" (*c.* 1847; Toledo Museum of Art, Ohio) powerfully express their creators' concern for the poor. Courbet created a sombre monument to his own village in "Burial at Ornans" (1849; Louvre), and Millet succeeded in conferring an epic grandeur on scenes of rural life.

Strain of poetic realism

A new approach to the familiar and unsophisticated occurs in the landscape painting of the 1830s and '40s; for, although French Romanticism produced no Turner, it did give rise to the Barbizon school, a group of naturalist painters who were particularly active in the forest of Fontainebleau. In this period the charm of the spontaneous sketch as opposed to the finished study was recognized: painters readily set up their easels in the open air and scrutinized the scene before them. A direct approach to nature and an interest in transitory moments, especially the changing effects of light, were features common to Romantic landscape painters throughout Europe and the United States. Paul Huet, a friend of Delacroix and Bonington and a painter closely associated with the Romantic school, represented dramatic, stormy scenes of solitude; yet, though scarcely a naturalist, he was deeply impressed by the works of Constable, several of which he copied and which inspired him to adopt a broken style of brushwork with dabs of bright pigment. The changed attitude to landscape is aptly expressed in the words of Théodore Rousseau, the most controversial representative of the new school: "Our art can only attain pathos through sincerity." Rousseau attempted to render nature as he found it, though his melancholic temperament is inevitably reflected in the desolate panoramas and gloomy sunsets in which he expressed an almost pantheistic feeling for the natural world. At the same time, his close attention to detail and painstaking accuracy in the delineation of plants and grasses betray the scientific concern shared by many Romantic artists. A similar penetration informed his studies of light, and both he and Charles-François Daubigny repeated virtually the same subjects under different weather conditions in order to capture the ephemeral effects of light and atmosphere. The freedom and freshness of Constable's handling is echoed in Daubigny's flickering treatment of sunset and light over water. A particularly poetic insight into nature was that of Narcisse-Virgile Diaz de La Peña and Constant Troyon. The work of Camille Corot, despite the restrained classicism of his style, is similarly enlivened by an instinctive feeling for naturalistic landscape (Plate 20). For, while they laid the foundation for the painterly revolution of the Impressionists, the Barbizon painters always retained the generous appreciation of natural beauty and emotional involvement with their subject that everywhere distinguish the Romantic temperament.

Barbizon school

United States. American Romantic painters were largely influenced by trends in late 18th-century Europe, especially Britain, but the absence of an indigenous artistic tradition permitted a much more intuitive development. At the same time, their work, like that of the early French Romantics, is closely associated with the new spirit fostered by a national revolution. The American Revolution, by reinforcing the democratic ideal, inspired a unique brand of Romantic realism that was a strong force in American painting from the late 18th century onward and that anticipated the emergence in Europe by a whole generation. Benjamin West, in addition to his contribution to Neoclassicism, developed a style of narrative painting with dramatic subjects taken from contemporary life; while he painted his most significant work in Britain, it was on American rather than English artists that it made the most impact. John Trumbull undertook a series of 12 scenes

Influence of the American Revolution on Romantic realism

from the American Revolution, in which careful studies of the principal participants were incorporated into colourful, baroque compositions. At their best, these works, for example "Sortie from Gibraltar" (1789; Museum of Fine Arts, Boston), carry great conviction, even if they tend to be somewhat theatrical. In 1784 one of the most candid portraitists of the period, Charles Willson Peale, completed a similarly ambitious project in his paintings of the leading figures of the Revolution. A more limited enthusiasm for precise naturalistic study informs the work of Alexander Wilson, whose devoted love of birds emerges in the freshness and simplicity of the plates to his *American Ornithology* (9 volumes; 1808–14). His achievement has been overshadowed by his greater successor, John James Audubon, who combined scientific precision with a delight in his specimens that transforms his watercolour drawings of birds into works of rare and delicate beauty.

Development of an American style of portraiture

At the beginning of the Romantic period, artists were still influenced by British painting, but this influence grew less and less perceptible as the 19th century progressed. For instance, the portrait of "Colonel Thomas Handasyd Perkins" (1831–32; Boston Athenaeum), by Thomas Sully, the leading exponent of a new portraiture supposedly expressive of mood, has touches of Sir Thomas Lawrence in the delicately brushed surface, strong contrasts of light and dark, and exquisite elegance of pose. But, though Samuel F.B. Morse, Samuel Waldo, William Page, and others also practiced an emotive style, portraits of the 19th century increasingly tended to endorse the native tradition of solid characterization.

The career of the landscape painter Washington Allston reflects the development of American painting in his lifetime. Absorbed by German and English Romantic poetry, he began on a note of high drama, moving in cosmopolitan artistic circles in Rome and producing a number of early landscapes that seem to have played a part in winning the friendship of the English Romantic poet Samuel Taylor Coleridge. At this point, what was obviously an impetuous and brooding strain in Allston's temperament found expression by depicting nature in the darker, more destructive moods dear to Turner. "The Deluge" (1804; Metropolitan Museum of Art, New York City) is a typical macabre invention, with bodies in a raging tempest swept ashore to where wolves and serpents lurk. On his return to the United States, however, his work assumed a quieter, more pensive aspect. "The Flight of Florimell" (1819; Detroit Institute of Arts) illustrates this later style.

An uncomplicated love for their own natural scenery emerges in the work of a succession of landscape painters who frequently strike a contemplative, lyrical note. Thomas Cole reverently recorded scenes in the valley of the Hudson River that echo the loneliness and mystery of the North American forests. With his generous humanitarian sympathies, Asher B. Durand gave a serene and artless account of nature. His feeling for space and finely diffused light renders "Kindred Spirits" (1849; New York Public Library) a touching tribute to the friendship of Cole with the American Romantic poet William Cullen Bryant. An interest in light and atmosphere was shared by George Loring Brown, FitzHugh Lane, Frederic Edwin Church, and George Harvey; all followed Durand and painted in the open. Simplicity and reticence distinguish the landscapes of Thomas Doughty, who concentrated on painting the Hudson River valley as he knew and loved it. The details of country life that fill the stories of Washington Irving are portrayed with affection by William Sidney Mount, who in "Eel Spearing at Setanket" (1845; New York State Historical Association, New York City) transcends the merely anecdotal. George Caleb Bingham approached the life of the frontier without the passionate concern that motivated many contemporary French artists. Solemn and severe in style and glowing with colour, his "Fur Traders Descending the Missouri" (*c.* 1845; Metropolitan Museum of Art) captures the silence and solitary grandeur of frontier life. The wildness of the frontier caught the imagination of many 19th-century artists: George Catlin, Seth Eastman, John M. Stanley, Alfred Jacob Miller, and Karl Bodmer all discovered a picturesque drama and excitement in Indian life. The Romantic period witnessed the emergence of a truly national school of painting in the United States, where events and scenery provided a constant source of stimulation for artists content to distill their own poetry from the world around them. (S.E.B./Ed.)

Painters of the frontier

Russia. Napoleon's invasion of Russia (1812) had far-reaching consequences. It marked the revival of national consciousness and the beginning of a widespread cult of Russian separateness from Europe, thus precipitating the long controversy between "Westerners" and "Slavophiles" that ran through so much of Russian 19th-century literature and thought. At the same time, Russia shared in the Romanticism—cultivated by France and Germany—that gripped Europe during the era of the Napoleonic Wars. This is reflected in the paintings of Orest Kiprensky and Vasily Tropinin. The most notable contribution to the Romantic spirit, however, was made by Karl Pavlovich Bryullov, with his monumental painting "The Last Days of Pompeii" (1830–33; State Russian Museum, St. Petersburg). A completely different trend appears in the work of Aleksandr Ivanov, the first Russian painter to express religious emotions in a western European manner. Other outstanding artists of that period were Aleksey Venetsianov and Pavel Fedotov, the forerunners of Realist painting in Russia.

The second half of the 19th century saw the maturing of Realism in Russia. A sympathetic attitude toward the hard life of the people is reflected in the works of most of the painters and sculptors of that time. The new trend in art had as its basis the populist revolutionary ferment prevalent toward the end of the 1850s and the beginning of the 1860s, much of it inspired by the writers Nikolay Dobrolyubov and Nikolay Chernyshevsky. Chernyshevsky's dissertation *Esteticheskiye otnosheniya iskusstva k deystvitelnosti* (1855; "The Aesthetic Relations of Art to Reality"), the main thesis of which was that art must not only reflect reality but also explain and judge it, provided a starting point for contemporary artists.

From the last third of the 19th century onward, the history of Russian art is the history of a series of school struggles: the Slavophiles against the Westerners; the Academy against the Peredvizhniki ("Wanderers"); and later the joint effort of the last two against a new movement, born in the 1890s and directed by the art review *Mir Iskusstva* ("The World of Art").

The Peredvizhniki society

The Peredvizhniki was a society formed in 1870 by a group of essentially Romantic artists who, however, regarded themselves as Realists. They seceded from the Academy in 1863 in protest against alien dogmatic formulas and the constricting programs of the Academy's annual competitions. Most prominent among the Peredvizhniki were Ivan Kramskoy, Ilya Repin, Vasily Ivanovich Surikov, Vasily Perov, and Vasily Vereshchagin. The society attached far more importance to the moral and literary aspects of art than to aesthetics. Its artistic creed was realism, national feeling, and social consciousness. Art was to be placed at the service of humanitarian and social ideals; it was to be brought to the people. Accordingly the society organized mobile (*peredvizheniye*) exhibitions—hence the name. The influence of the Peredvizhniki spread throughout Russia and was dominant for nearly 30 years, but by the end of the century it had greatly declined. (A.Vo./Ed.)

Modern

The term modern art has come to denote the innovating and even revolutionary developments in Western painting and the other visual arts since the second half of the 19th century. It embraces a wide variety of movements, styles, theories, and attitudes, the modernity of which resides in a common tendency to repudiate past conventions and precedents in subject matter, mode of depiction, and painting technique alike. Not all the painting of this period has made such a departure; representational work, for example, has continued to appear, particularly in connection with official exhibiting societies. Nevertheless, the idea that some current types of painting are more properly of their time than are others, and for that reason are more interesting or important, applies with particular force to the painting of the last 150 years.

By the mid-19th century, painting was no longer basically in service to either the church or the court but rather was patronized by the upper and middle classes of an increasingly materialistic and secularized Western society. This society was undergoing rapid change because of the growth of science and technology, industrialization, urbanization, and the fundamental questioning of received religious dogmas. Painters were thus confronted with the need to reject traditional, historical, or academic forms and conventions in an effort to create an art that would better reflect the changed social, material, and intellectual conditions of emerging modern life. Another important, if indirect, stimulus to change was the development from the early 19th century on of photography and other photomechanical techniques, which freed (or deprived) painting and drawing of their hitherto cardinal roles as the only available means of accurately depicting the visual world. These manually executed arts were thus no longer obliged to serve as the means of recording and disseminating information as they once had been and were eventually freed to explore aesthetically the basic visual elements of line, colour, tone, and composition in a nonrepresentational context. Indeed, an important trend in modern painting has been that of abstraction—*i.e.*, painting in which little or no attempt is made to accurately depict the appearance or form of objects in the realm of nature or the existing physical world. The door of the objective world was thus closed, but the inner world of the imagination offered seemingly infinite possibilities for exploration, as did the manipulation of pigments on a flat surface for their purely intrinsic visual or aesthetic appeal.

Abstraction in painting

The beginnings of modern painting cannot be clearly demarcated, but it is generally agreed that it started in mid-19th-century France. The paintings of Gustav Courbet, Edouard Manet, and the Impressionists represent a deepening rejection of the prevailing academic traditions of Neoclassicism and Romanticism and a quest for a more truthful naturalistic representation of the visual world. These painters' Postimpressionist successors—notably Paul Cézanne, Vincent van Gogh, Edgar Degas, and Paul Gauguin—can be viewed as more clearly modern in their repudiation of traditional subject matter and techniques and in their assumption of a more subjective and personal vision. From about the 1890s a succession of varied styles and movements arose that are the core of modern painting and are also one of the high points of the history of the Western visual arts in general. These modern movements include Neo-Impressionism, Symbolism, the Nabis, Fauvism, Cubism, Expressionism, Futurism, the Ashcan School, Suprematism, Constructivism, Orphism, Metaphysical painting, de Stijl, Purism, Dada, Surrealism, Social Realism, Abstract Expressionism, Pop Art, Op Art, Minimalism, and Neo-Expressionism. (F.W.W.-S./Ed.)

ORIGINS IN THE 19TH CENTURY

As long ago as 1846 the qualities proper to a specifically modern art were discussed by the French writer Charles Baudelaire in an essay on the French Salon. He argued that colour would be foremost among these modern qualities (a prediction that subsequent events confirmed), but he still saw the new art in the context of the Romantic movement. Subsequent modernity came to be seen as necessitating not only a new style but also contemporary subject matter, and in 1863 Baudelaire praised the draftsman Constantin Guys as "le peintre de la vie moderne" ("the painter of modern life"). In 1862, with Baudelaire's support, the French painter Édouard Manet brought together a subject from contemporary social life and an unconventional style in "Concert in the Tuileries Gardens" (National Gallery, London). This painting, though rather isolated in his work of the time, was influential in establishing a new outlook. Another literary figure whose critical writings were influential was the French novelist Émile Zola, though Zola had limited sympathy for what he called the "new manner in painting" of Manet; nevertheless he contributed from 1866 onward to the emergence of the Impressionist group. The first appearance of the phrase "modern art" in the relatively permanent form of a book title was in 1883, when it was used by the French writer Joris-Karl Huysmans, a friend of Zola's, to describe the theme of various reviews of painters' work he had collected. Other books on the subject followed, such as the Anglo-Irish novelist George Moore's *Modern Painting* (1893). It was about this time that the term avant-garde was introduced by the critic Théodore Duret, who used it of certain young painters. From then on, modernity was to be a recurrent concern of artists and critics. Public acceptance of the new standpoint was slow, however. The first museums dedicated specifically to modern art grew out of the fervour of individual collectors—for example, the Folkwang Museum at Essen, Ger., and the Kröller-Müller State Museum at Otterlo, Neth., both largely consisting of collections built up before 1914. The Museum of Modern Art in New York City, the outstanding public collection in the field, was founded in 1929, and the Western capital that lacks a museum explicitly devoted to modern art is rare.

Birth of modern art museums

The conflict between the new forces and the established academic tradition in France came into the open in 1863. The jury of the official Salon, which had long exercised great despotism in matters to do with painting, rejected more than 4,000 canvases—an unusually high figure. The resulting outcry prompted the emperor Napoleon III to order that the rejected works, if the painters agreed, be shown in a special exhibition known as the Salon des Refusés. The exhibition included works by Manet; Johan Barthold Jongkind, an older Dutch painter who was working in a tonal and summary style from nature; Camille Pissarro and Paul Cézanne, who had met two years before at the Académie Suisse; Armand Guillaumin; James McNeill Whistler; and others. One of the greatest scandals was caused by Manet's painting "The Luncheon on the Grass" (Louvre, Paris), which was considered an affront to decency as well as taste. The younger painters became aware of their common aims. Claude Monet, whose landscape style had been influenced from the outset by the atmospheric sketches of the Channel coast of Eugène Boudin, as well as by Jongkind (whom he described to Boudin as "quite mad"), had met Pierre-Auguste Renoir, Alfred Sisley, and Jean-Frédéric Bazille studying in the studio of Charles Gleyre. Abandoning academic study, they worked together outdoors in the forest of Fontainebleau, where contacts with the Barbizon painters Narcisse-Virgile Diaz de La Peña and Charles-François Daubigny strengthened their direction.

The implicit acceptance of the visual scene on which the new style was based owed something to the example of Courbet, who influenced Renoir in particular in the next few years. The plein air ("open-air") paintings of the Barbizon painters also had an effect, but the suggestion of an art based on the notation of pure colour was suggested by several sources. The example of Eugène Delacroix had a deep significance for the 19th century in France, and the reliance on separate, undisguised touches of the brush in the form that became characteristic of Impressionism is perhaps first apparent in sketches of the sea at Dieppe painted by Delacroix in 1852. The economy of Manet's touch in the 1860s was affected by Spanish and Dutch examples as well as by Delacroix, but his seascapes and racecourse pictures of 1864 are also important. The full Impressionistic style did not develop until the end of the 1860s.

Beginnings of Impressionism

Though the figurative aims of Impressionism can be regarded as the conclusion of 19th-century Realism, the method, which made no attempt to hide even the most basic means of preparing a finished painting, was an original one. Brushstrokes did not pretend to be anything but dashes of paint, thus conveying their coloured message without any disguise or effect at individual illusion. It was in this respect and in the all-embracing unity of colour and handling that resulted, rather than its realism, that Impressionism founded modern painting. Other developments in the 1860s had no immediate sequels in Impressionism. The presentation of some of Manet's figures, such as "The Fifer" (Louvre) of 1866, as vignettes or decorative designs shading into virtually blank backgrounds was a radical departure from the coherent pictorial construction of Western tradition since the Renaissance; it is the first

sign of the form built outward from a central nucleus without reference to the classic frame that has appeared repeatedly in modern art. Honoré Daumier is supposed to have said that "The Fifer" reduced painting "to faces on playing cards," and in 1865 Courbet compared Manet's "Olympia" (1863; Louvre) to "the Queen of Spades after a bath." The possibility of making an image out of the bare, almost heraldic juxtaposition of flat colours was neglected while the complex notation of Impressionism held sway, but it came to be regarded with interest as Impressionism receded. Other unconventional principles of design—suggested equally by Japanese prints, such as those that Manet placed in the background of his portrait of Zola (Louvre) in 1868, and by the chance arrangements of photography—appeared in the work of Edgar Degas, who sympathized with the aims of the new group, associating himself with them in seven of their eight exhibitions, which he largely helped to organize.

Other qualities that Baudelaire in 1846 had specified as the qualities of modern art—spirituality and aspiration toward the infinite—evolved quite apart from Impressionism. The visionary implications of Romantic painting were explored by Gustave Moreau, whose elaborate biblical and mythological scenes, weighed down with sumptuous detail, gave colour an imaginative and symbolic richness. His example had a special value to the next generation. The imagination of Pierre Puvis de Chavannes was of the opposite order, preserving the large-scale clarity of mural painting, a policy that made him appreciated when a reaction against Impressionism set in.

Another possibility of Romanticism was pursued in isolation by the Marseille painter Adolphe Monticelli. The richness of his colour is thought to have contributed something crucial to Cézanne's development. The counterpart of Moreau in Britain was Sir Edward Burne-Jones. The intricate and perverse linear formulations that he developed from the Pre-Raphaelites greatly influenced the international Symbolist style of the last decades of the century.

Early developments in Britain and Germany

The influence of the trend in the direction of the modern in France, together with its controversial element, was introduced to Britain by Whistler, whose concern was narrowly aesthetic rather than analytic. The harmonies he developed were close to being monochromatic; his use of Spanish and Japanese elements had little of the radical originality of Manet and Degas. His influence dominated and also limited the development of avant-garde painting in Britain for many years. John Singer Sargent, like Whistler an American who came to live in Britain, popularized a less-discriminating version of the Impressionistic style.

In Germany a Romantic strain coexisted with a Realistic style that remained unaffected by the most advanced French painting. Anselm Feuerbach, one of the Romantics, was influenced by Delacroix. In 1855 he went to Italy where the effect of the 16th century came to predominate in his work. The landscapes of Hans von Marées were also essentially Romantic. He had visited France but spent most of his working life in Italy; the frescoes he executed in Naples echo Puvis de Chavannes in their style. Realism found exponents in Wilhelm Leibl and Hans Thoma. In Italy the reaction against the academies was centred in Florence, where a group known as the Macchiaioli (from *macchia,* "patch") worked from 1855, producing landscapes, genre paintings, and Romantic costume pieces executed in the highly visible brushstrokes that gave the group its name.

The beginning of modern art in the United States

In the United States, Thomas Eakins developed a broadly handled, powerful Realist style that became almost Expressionistic in his later years (Plate 21). He had visited Paris in 1866, and the influence of Manet can be detected in his paintings. His interest in anatomy and perspective gave him a role analogous to that of Degas. The early development of Winslow Homer, who was in France a year later, ran parallel to Monet's style in the 1860s. The work of Albert Pinkham Ryder was, by contrast, introverted and visionary. He was among the artists who adapted the Romantic vocabulary to the symbolic purposes of modern art.

In France in the mid-1860s Monet produced a series of large-scale open-air conversation pieces in which elements derived from Courbet and Manet were fused with a wholly original expression of dappled light in solid paint. The approach of Pissarro, who had arrived in Paris from the West Indies in 1855, was more delicate; influenced by Camille Corot as well as Courbet, he recorded pure landscape motives in a limited range of tones, though with a natural lyricism of feeling. The starting point of Cézanne was, by contrast, vigorous to the point of violence. In 1866 he evolved a style in which paint was applied in thick dabs with a palette knife; this combined a handling (a technical term in painting meaning the individual's manipulation of materials in the execution of a work; it has been likened to a person's signature in handwriting) derived from Courbet with the gray tonality of Manet; its rough-hewn crudity has a consistency that was essentially new. His alternative style in the 1860s, with curling brushstrokes related to Daumier, is equally virile and was often applied to subjects of violent eroticism. The unbridled force of Cézanne's early work gave the first sign of qualities that were to become characteristic of modern painting. Though exceptional, it was not unique; in Italy during the 1860s the Russian painter of historical and scriptural themes, Nikolay Nikolayevich Ge, produced sketches with loose, expressive brushwork sometimes resembling Cézanne's.

The prelude to French Impressionism

Impressionism. The first steps toward a systematic Impressionist style were taken in France in Monet's coast scenes from 1866 onward, notably the "Terrace" (1866; Metropolitan Museum of Art, New York City), in which he chose a subject that allowed use of a full palette of primary colour. The decisive development took place in 1869, when Monet and Renoir painted together at the resort of La Grenouillère on the Seine River. The resulting pictures suggest that Monet contributed the pattern of separate brushstrokes, the light tonality, and the brilliance of colour; Renoir the overall iridescence, feathery lightness of touch, and delight in the recreation of ordinary people. Working at Louveciennes from 1869, Pissarro evolved the drier and more flexible handling of crumbly paint that was also to be a common feature of Impressionist painting.

It was in the environs of Paris after the Franco-Prussian War that there developed the fully formed landscape style that remains the most popular achievement of modern painting. An exhibition held in the studio of the photographer Nadar (Gaspard-Félix Tournachon) in 1874 included Monet's picture "Impression: Sunrise" (Marmottan Museum, Paris; Plate 23), and it was this work that, by being disparaged as mere "impressionism," gave a name to an entire movement. The exhibition itself revealed three main trends. The Parisian circle around Monet and Renoir had developed the evanescent and sketchlike style the furthest. The vision of those working near Pissarro in Pontoise and Auvers was in general more solid, being firmly rooted in country scenes. A relatively urbane, genrelike trend was detectable in Degas's picture of Paul Valpinçon and his family at the races called "Carriage at the Races" (1870–73; Museum of Fine Arts, Boston) and Berthe Morisot's "The Cradle" (1873; Louvre). Manet himself was absent, hoping for academic success; his "Gare Saint-Lazare" (1873; National Gallery of Art, Washington, D.C.), influenced by the Impressionist palette, was accepted at the Salon. Modeling himself on Pissarro, Cézanne sublimated the turbulent emotions of his earlier work in pictures that were studied directly and closely from nature; he followed the method for the rest of his life.

The first Impressionist exhibition

The experiment of an independent exhibition was repeated in 1876, though with fewer participants. Monet now began to make studies of the Gare Saint-Lazare. Renoir used effects of dappled light and shadow to explore genre subjects such as "Le Moulin de la galette" (1876; Louvre; Plate 23). In 1877 only 18 artists exhibited. The major painters began to go their separate ways, particularly as there were disputes about whether to continue with the independent exhibitions. Cézanne, who did not exhibit with the Impressionists again, was perhaps the first to realize that a critical stage had been reached. For the first time, a style had been based on the openly individual character of a technique rather than on the form of a particular subject or the way it was formulated. A style that

admits to painting as being only a matter of paint raises in a peculiarly acute form the question of how far the qualities of art are intrinsic. Impressionism in the 1870s was inseparable from heightened visual experience of a sensuously satisfying world. But the blocklike shapes in Cézanne's pictures, such as the portrait of his patron Victor Chocquet (*c.* 1877; Columbus Museum of Art, Ohio), suggest that for him the relationship between the colour patches on his canvas was equally important. In the years that followed, he systematized his technique into patterns of parallel brushstrokes that gave a new significance to the pictorial surface. An unassuming series of still lifes and self-portraits by Cézanne were painted in 1879–80, and these, when they became known, profoundly impressed the younger generation, who reckoned them to be as monumental as the great art of the past, yet in a subtly different way that was inherent in the actual manner of painting.

The style of the 1870s was formless from a traditional standpoint, and at the beginning of the next decade Renoir decided that he had gone to the limit with Impressionism and "did not know either how to paint or draw." Following a trip to Italy, he set about acquiring a wiry, linear style that was the direct opposite of his relaxed, freely brushed manner of earlier years.

The appearance of a new generation posed a fresh challenge. Georges Seurat was moving away from the empirical standpoint of Impressionism toward a technique (Pointillism) and a form that were increasingly deliberately designed. Paul Gauguin, taking his starting point from Cézanne's style of about 1880, passed from a capricious personal type of Impressionism to a greater use of symbols. He exhibited with the Impressionists from 1880 onward, but it was soon evident that group shows could no longer accommodate the growing diversity. In 1884, after the Salon jury had been particularly harsh, the Société des Artistes Indépendants was formed. The last Impressionist group show was held in 1886. Only Monet and Armand Guillaumin, to whose efforts the group owed much of its eventual recognition, were now in the strict sense Impressionists. Monet, who had exhibited only once since 1879, continued to build on the original foundation of the style, the rendering of visual impression through colour in paintings that studied a single motif in varying lights. For him the formlessness and the homogeneity of Impressionism were its ultimate virtues. In his last series of "Water Lilies," painted between 1906 and 1926, the shimmering of light eventually lost its last descriptive content, and only the colour and curling movement of his brush carried a general all-pervading reference to the visual world. Renoir's later work was equally expansive; his sympathetic vision of humanity revealed its own inherent breadth and grandeur.

The last Impressionist exhibition

Impressionism, in one aspect, continued the main direction of 19th-century painting, and after 1880 the movement was an international one, taking on independent national characteristics. Russia produced an exponent in Isaak Ilich Levitan, and Scotland one in William MacTaggart. In Italy Telemarco Signorini and in the United States such painters as Childe Hassam developed modified forms of the style. In France, and to some extent in Germany with Max Liebermann, Impressionism provided a basis for the styles that followed.

Symbolism. During the decades before 1900, the Symbolists were the avant-garde, and one of quite a new kind, influencing not only the arts but also the thought and spirit of the epoch. Maurice Denis, their theoretician, enunciated in 1890 the most famous of their artistic principles:

> Remember that a picture—before being a war-horse, a nude or an anecdote of some sort—is essentially a flat surface covered with colours assembled in a certain order.

Such ideas inspired a group of young painters, among whom was Denis himself, to call themselves Nabis (from the Hebrew word for "prophet"). They were in revolt against the faithfulness to nature of Impressionism; in addition, largely because they were in close touch with Symbolist writers, they regarded choice of subject as important. They included Paul Ranson, who gave the style a decorative and linear inflection; Pierre Bonnard; and Édouard Vuillard.

The Nabis

Other than the Nabis, one of the chief Symbolists was Odilon Redon, who moved from the same starting point as the Impressionists—the landscape style of the Barbizon school—but in precisely the opposite direction. Redon's visionary charcoal drawings (which he called his black pictures) led to successive series of lithographs that explored the evocative, irrational, and fantastic orders of creation that Impressionism excluded. Redon later wrote:

> Nothing in art can be done by will alone. Everything is done by docile submission to the coming of the unconscious . . . for every act of creation, the unconscious sets us a different problem.

Redon established one of the characteristic standpoints of modern art, and his influence on the younger Symbolists was profound. In 1888 Gauguin, already affected by a trip to Martinique, settled at Pont-Aven in Brittany. The influential style he developed there was based on the juxtaposition of flat areas of colours enclosed by black contours, the total effect suggesting cloisonné enamel (a technique in which metal strips differentiate the colour areas of the design, thereby creating an outline effect), hence the name Cloisonnisme used to describe this style. The spirit in which Gauguin rendered Breton scenes was mystical. He wrote:

Cloisonnisme

> Do not copy nature too much. Art is an abstraction; derive this abstraction from nature while dreaming before it, but think more of creating than of the actual result.

At Pont-Aven, Gauguin was joined by Émile Bernard and Louis Anquetin, who had lately begun to work in a similar way. Paul Sérusier painted under Gauguin's direction a little sketch entitled "Bois d'amour" that appeared more independent of appearances and bolder in its synthesis of pattern than anything that had been seen before; it became known in Paris as "The Talisman." The liberation of Synthetism, as the new style was called, indeed worked like a charm, and after the Café Volpini exhibition of 1889 it spread rapidly. The movement was linked with literature and, in particular, with drama; it inspired its own periodical, *La Revue Blanche,* and Le Théâtre de l'Oeuvre (both founded in Paris in 1891); there were exhibitions twice a year at a Paris gallery, Le Barc de Boutteville, from 1891 to 1897.

The decorative style known as Art Nouveau, or Jugendstil, spread across Europe and the Americas in the 1890s. The pursuit of natural and organic sources for form still further alienated art from the descriptive purpose that had been the basis of figurative style, and an artistic movement without taint of historicism that molded the fine arts, architecture, and craftsmanship in a single, consistent taste recovered the creative unity that had been lost since the early 18th century. In The Netherlands the fin de siècle ("end of the century"; specifically the end of the 19th century, and a phrase that has overtones of a rather precious sophistication and world-weariness) style and sense of purpose appeared in the paintings of Johan Thorn Prikker and Jan Toorop. The Viennese Gustav Klimt made bolder and more arbitrary use of pattern (Plate 22). In Russia the demonic genius of Mikhail Aleksandrovich Vrubel had points of contact with the Art Nouveau style. It even affected Seurat and his circle, who were known as the Neo-Impressionists; the popular imagery of Seurat's later works, such as "The Circus" (1890–91; Louvre), was expressed in sinuous rhythms not far from Art Nouveau, and the Belgian Henry van de Velde passed from Neo-Impressionism by way of fin de siècle decorations that were near abstraction to a place among the founders of 20th-century architecture. A strange and beautiful blend of Symbolism with an alpine clarity of colour close to Neo-Impressionism appeared in compositions such as "The Unnatural Mothers" by the Italian Giovanni Segantini.

The end of the 19th-century tradition. Until Seurat no painter had expressly founded a style on the intrinsic reactions of colour to colour and a codified vocabulary of expressive forms. The consistent granulation of colour in Seurat's work from 1885 onward was specific to the picture, not to the sensation or the subject (Plate 22). The coherent images of space and light that he made out of this granulation ended with him. Seurat's followers, grouped as Neo-Impressionists under the leadership of Paul Signac,

Neo-Impressionists

developed his technique rather than his vision. Seurat's influence was nonetheless widespread and fertile; his system in itself supplied a clarity that painters needed. It was Neo-Impressionism that was in the ascendant when the Dutch painter Vincent van Gogh arrived in Paris in 1886. The emotional travail evident in van Gogh's early work was marvelously lightened in the new aesthetic climate. But in his hands the dashes of pure colour turned and twisted, trading invisible and unstable lines of force. They were woven into rhythmical and convulsive patterns reflecting the mounting intensity of his own feelings. Such patterns converted the Neo-Impressionist style into something quite different—a forerunner of what was to be known as Expressionism (Plate 22). Other painters were less radical in their approach. Pissarro assimilated the Neo-Impressionist method to the vision of the older generation; Henri-Edmond Cross and Maximilien Luce gave it the characteristic economy of the age that followed. Henri Matisse's repeated experiments with it, culminating in his contact with Signac and Cross in 1904, finally converted the pure colour of Impressionism to the special purposes of 20th-century art.

In the meantime, the older Impressionists were producing the broadly conceived works that crowned their artistic achievement and formed, as it seems in retrospect, the great traditional masterpieces of modern art. Degas's lifelong absorption in the human body as a subject led him to produce a series of bathing scenes and drawings from the nude in which the form expanded to an amplitude that filled the picture. Fullness of form was an effect that Renoir also achieved. Cézanne announced a determination "to do Poussin over again from nature" and was reckoned to have fulfilled that aim with his "Great Bathers" and the series of landscapes of Mont Sainte-Victoire (Plate 22). In the pictures of Henri de Toulouse-Lautrec, the style and standpoint derived from Degas, but his graphic work reflected the aims of the Symbolist generation. The most original contribution of Édouard Vuillard lay in the evocative patterning of the little pictures that he painted before 1900. The art of Pierre Bonnard, on the other hand, developed throughout his life. His subjects and his method remained, on the surface, those of the Impressionist tradition, but they were re-created from memory and imagination; Bonnard's pictures have the quality of a cherished private order of experience.

Developments outside France

Developments outside France were not of comparable importance. In Britain in the 1880s, Philip Wilson Steer painted a small group of landscapes with figures that were among the earliest and loveliest examples of the fin de siècle style. The work of Walter Sickert revolved around an idiosyncratic fascination with the actual touch of a brush on canvas. His affinities remained essentially with the tonal Impressionism of the earliest stages of the modern movement rather than with the art of colour that developed from it, though he eventually made the transition in old age. In Germany the artists of the Postimpressionist generation, such as Lovis Corinth and Max Slevogt, working with the peculiar recklessness that is endemic to German painting, laid the technical foundations of Expressionism. Ferdinand Hodler in Switzerland developed a painterly Symbolist style in the 1890s. The Belgian painter James Ensor abandoned Impressionism at the end of the 1880s for a bitter and fantastic style that was a pioneer example of extreme expressive alienation.

The most remarkable painter of the fin de siècle outside France, however, was the Norwegian Edvard Munch. "The Cry" (Nasjonalgalleriet, Oslo; Plate 23), the famous picture in which the rhythms of Art Nouveau were given a hysterical expressive force with hardly a vestige of the Impressionist description of nature, was painted in 1893. For many years before a breakdown interrupted his development in 1907, he worked abroad. He was particularly influential in Germany.

In the United States, Maurice Prendergast transformed Impressionism into pattern. In Russia the fin de siècle styles of Léon Bakst and the Mir Iskusstva ("World of Art") group, as well as a vivid revival of folk decoration, flourished, later becoming known internationally through their connection with the Russian ballet.

THE 20TH CENTURY

By 1903 the impetus of Symbolism was expended and a new and enigmatic mood was forming. The new attitude drew on a vein that was comic, poetic, and fantastic, exploring an irrational quality akin to humour inherent in the creative process itself, as well as on a reserve of ironic detachment. The new painters drew strength from unexpected sources. The work of Henri Rousseau, a former clerk in the Paris municipal customs service who was known as "Le Douanier" accordingly, and who had exhibited at the Indépendants since 1886, attracted attention. The apparent innocence of his pictures gave them a kind of imaginative grandeur that seemed beyond the reach of any art founded on sophistication (Plate 25).

The attraction of naive and "primitive" painting

The art of supposedly primitive peoples had a special appeal in the early years of the 20th century. Gauguin, who had made direct contact with it in his last years, proved prophetic not only in the forms he adopted but in the spirit of his approach (Plate 23). Maurice de Vlaminck and André Derain, who met in 1900, evolved a style together based on crude statements of strong colours. Matisse had been moving more circumspectly in the same direction (Plate 25). The apparent ferocity of the works that the three exhibited in 1905 earned them the nickname of the Fauves ("Wild Beasts"). It appears that Matisse was responsible for introducing Pablo Picasso to African sculpture. Picasso had already shown signs of dissatisfaction with existing canons; his use of fin de siècle styles in his earliest works has a quality close to irony. Primitive art, both African and Iberian, provided him with an austerity and detachment that led after 1906 to a radical metamorphosis of the image and style hitherto habitual in European art. In 1904 Ernst Ludwig Kirchner, at Dresden, discovered the art of the Pacific Islands as well as African art. His first reflection of the primitive spirit was parallel to that of the Fauves and may have depended on them, if only partially.

Concept of art as expression

The idea of art, first and last, as a matter of expression (in contrast to Impressionism) was common to Germany and France in the first decade of the 20th century; it appears in Matisse's *Notes of a Painter*, published in 1908. Matisse, in fact, hardly differentiated expression from decoration; his ideal of art as "something like a good armchair in which to rest" explicitly excluded the distortion and disquiet that earned the style of Kirchner and Die Brücke ("The Bridge") group, which was founded in 1905, the label of Expressionism. The worldly subjects of Kirchner represented only one aspect of the group; the earthy Primitivism of Emil Nolde and the emphatic pictorial rhetoric of Karl Schmidt-Rottluff are more typical. Both Nolde and Max Pechstein (another member of the group) traveled to the Pacific. The development of the Austrian Oskar Kokoschka, who was influenced by members of Die Brücke, spanned the first two-thirds of the 20th century; the tempestuous emotion of his finest pictures places them among the masterpieces of painting of the German-speaking world in his time.

The transformation of painting after 1907 was particularly apparent in works executed in Germany. Wassily Kandinsky had come to Munich from Moscow at the age of 30 in 1896. His earliest mature works were painted in a jewellike, fairy-tale Cloisonniste style. He later told how one evening in his studio he came upon "an indescribably beautiful picture, drenched with an inner glowing . . . of which I saw nothing but forms and colours . . ." (from R.L. Herbert [ed.], *Modern Artists on Art*, 1965). It was one of his own works, standing on its side, so that its content was incomprehensible. Kandinsky's first nonfigurative watercolour was painted in 1910, and in the same year he wrote much of *Concerning the Spiritual in Art*, which converted the aesthetic doctrines of Goethe to the purposes of the new art. The series of "Improvisations" that followed preserved reminiscences of figuration, made illegible by the looseness of the pictorial structure: their diffuse and amorphous consistency had little connection with the main objectives of painting at the time (Plate 24). In the first decade of the 20th century, the idea of painting implied by Postimpressionism and that of a reasoned structure analogous to the structure of nature, if not to

Development of nonfigurative painting

appearances, were far from exhausted. The influence of Kandinsky's "Improvisations" from 1911 onward, though delayed, was nonetheless great and pointed in a direction that abstract painting was to take 40 years later.

The Munich group Der Blaue Reiter ("The Blue Rider"), named after one of Kandinsky's earlier pictures, was formed in 1911 to represent the new tendencies when Kandinsky and Franz Marc withdrew from the heterogeneous Neue Künstlervereinigung ("New Artists' Association"). The group soon became, in its turn, a broadly based assembly of the international avant-garde artists of the day, although the stylizations of Marc himself now appear commonplace. Among the early members of the group, the Russian Alexey von Jawlensky evolved a structured form of Expressionism that culminated in the 1930s in a series of abstractions of a head, but the chief importance of the group was as a stage in the development of the Swiss painter Paul Klee.

Cubism and its consequences. Picasso's Primitivism, joined to the influence of Cézanne's "Great Bathers," culminated in 1907 in the enigmatic and famous picture "Les Demoiselles d'Avignon" (Museum of Modern Art, New York City). Those who saw it were astonished and perplexed, not only by the arbitrary disruption in the right-hand part of the picture of the continuity that had always united an image but also by the defiant unloveliness, which made it plain that the traditional beauties of art, the appeal of the subject, and the credibility of its imitation were now, at any rate to Picasso, finally irrelevant. "What a loss to French art!" a Russian collector said, and Picasso himself was not sure what to think of the picture; it was not reproduced for 15 years or publicly exhibited for 30. Nevertheless, the effect on his associates was profound. Matisse and Georges Braque, who, unlike Picasso, had been experimenting with Fauvism, immediately started painting female nudes of similar stridency. Subsequently, however, Matisse turned back toward relatively traditional forms and the flooding colour that chiefly concerned him. Braque, on the other hand, became more and more closely associated with Picasso, and Cubism, as the new style was labeled by one of Braque's hostile critics in the following year, was the result of their collaboration. In the first phase, lasting into 1909, the focus of their work was the accentuation and disruption of planes (Plate 24). In the next two years Braque went to paint at Cézanne's old sites, and the inspiration of Cézanne's style at this stage is indubitable. In the second phase, from 1910 to 1912, the irrelevance of the subject, in any integral form, became evident. It was no longer necessary to travel in search of a motif; any still life would do as well. The essence of the picture was in the treatment. If Analytical Cubism, as this phase is generally labeled, analyzed anything, it was the nature of the treatment. The great Cubist pictures were meditations on the intrinsic character of the detached Cézannesque facets and contours, out of which the almost-illegible images were built. Indeed, the objects were not so much depicted as denoted by linear signs, a spiral for the scrolled head of a violin or the trademark from a label for a bottle, which were superimposed on the shifting, half-contradictory flux of shapes. The element of paradox is essential; even when it approaches monumental grandeur, Cubism has a quality that eludes solemn exposition. Subtle and elegant geometric puns build up into massive demonstrations of pictorial structure, demonstrations that its complex parallels and conjunctions build nothing so firmly and so memorably as the picture itself. This proof that figurative art creates an independent reality is the central proposition of modern art, and it has had a profound effect not only on painting and sculpture, as well as on the arts of design that depend on them, but also on the intellectual climate of the age.

Early Cubism: 1907–10

Analytical Cubism: 1910–12

The experimental investigation of what reality meant in artistic terms then took a daring turn that was unparalleled since pictorial illusion had been isolated five centuries earlier. The Cubists proceeded to embody real material from the actual world within the picture. They included first stenciled lettering, then pasted paper, and later solid objects; the reality of art as they saw it absorbed them all. This assemblage of material, called collage, led in 1912 to the third phase of the movement, Synthetic Cubism, which continued until 1914. The textured and patterned planes were composed into forms more like pictorial objects in themselves than recognizable figurations. In the later work of Picasso and Braque, it is again possible to construe their pictorial code as referring plainly to the objective world—in the case of Braque, to still life chosen with an appreciation of household things and, with Picasso, to emotive yet enigmatic human subjects as well. The message of Cubism remained the same: meaning had been shown to reside in the structure of the style, the basic geometry implied in the Postimpressionist handling of life. The message spread rapidly.

Synthetic Cubism: 1912–14

The first theoretical work on the movement, *On Cubism,* by the French painters Albert Gleizes and Jean Metzinger, was published in 1912. It was argued that geometric and mathematical principles of general validity could be deduced from the style. An exhibition in the same year represented all Cubism's adherents except the two creators. The exhibition was called the Section d'Or ("Golden Section"), after a mathematical division of a line into two sections with a certain proportion to each other. Among the exhibitors were the Spaniard Juan Gris and the Frenchman Fernand Léger, who in their subsequent work were both concerned with combining the basic scheme of Synthetic Cubism with the renewed sense of a coherent subject. Cubism stimulated parallel tendencies in The Netherlands, Italy, and Russia. In The Netherlands, Piet Mondrian, on the basis of Cézanne and the Dutch painters of the fin de siècle, had reached a very simple, symbolic style analogous to the Dutch landscape. He first saw Cubist paintings in 1910 and moved to Paris two years later. The subsequent resolution of his sense of natural conflict into increasingly bare rectangular designs balancing vertical against horizontal and white against areas of primary colour is one of the achievements of modern art (Plate 25). In 1917 the de Stijl movement formed in The Netherlands around him, with lasting consequences for the architecture, design, and typography of the century.

The Section d'Or group

In Italy in 1909 a program for all the arts was issued by the poet Filippo Marinetti, who called his exercise the Futurist manifesto. He rejected the art of the past and exalted energy, strength, movement, and the power of the modern machine. In painting, his ideas were taken up by Carlo Carrà. Umberto Boccioni, the most talented of the group, pursued its ideas not only in painting but also in sculpture. The most memorable serial images of movement were those of Giacomo Balla; they reveal that, under its vivid fragmentation, the vision of Futurism was not far from the photographic. Its imperative mood and disruptive tactics nonetheless had their effect, finding an echo in Britain in the Vorticist circle around Wyndham Lewis. Lewis' analytical intelligence and the toughness of his artistic temper marked equally his near-abstract early works and the incisive classical portraits he painted later. Among his early associates, David Bomberg developed from the Cubist idiom in 1912–13 images of a striking clarity and force; and William Roberts combined a Cubist formulation with social commentary analogous to that of the 18th-century painter William Hogarth.

Italian Futurists

In Russia, where Western developments were well known, the avant-garde, with its own roots in primitive art, had already evolved a simplified Expressionistic style. Kazimir Malevich produced formalized images of peasants at work that anticipated the later style of Léger. The strip-like and often abstract formulations of Mikhail Larionov and Natalya Gontcharova, to which they gave the name of Rayonism, date from 1911. In 1912 Malevich exhibited his first "Cubo-Futurist" works, in which the figures were reduced to dynamic coloured blocks, and in 1913 he followed these with a black square on a white background. This increasing tendency to abstraction reached its culmination in 1915 with the arrival of what he called Suprematism, in which simple geometric elements provided the whole dynamic force. The Russian movement, complicated by its own politics, was both accelerated and eventually broken by the Revolution, which gave it, for a time, a social function that the avant-garde has hardly achieved elsewhere. The Russian artists dispersed after

1922, however, and their legacy, the tradition of Constructivism, was transmitted to western Europe by El Lissitzky, Antoine Pevsner, and the latter's brother, Naum Gabo.

Orphic Cubism

Prismatic colour, the element in Cézanne that the Cubists had neglected in dismantling his style, was taken up by Robert Delaunay. Delaunay's variety of Cubism was named Orphism, after Orpheus, the poet and musician of ancient Greek myth. The essential discovery of Orphism was proclaimed as a realization that "colour is both form and subject." After an exquisite series of "Windows," Delaunay freed himself from representation and based his designs on the effects of simultaneous colour contrast. These dictated the concentric patterns of his "Discs" and "Cosmic Circular Forms," which occupied him and his wife, Sonia, thenceforward. The Czech František Kupka painted his first totally abstract work at about the same time. Even the simplest of his subsequent works never quite lost the rhythms of the fin de siècle style. Delaunay realized that a new order of painting was beginning, but his immediate influence was strongest abroad. Two American followers, Stanton Macdonald-Wright and Morgan Russell, exhibited as "Synchromists" in 1913. The Munich group Der Blaue Reiter was in touch with Delaunay, who exhibited with them, and the subsequent development of Klee was founded on his conversion to Delaunay's ideal of colour.

Fantasy and the irrational. The identity of a work of art as a thing in itself, independent of representation, was on the way to general recognition when the outbreak of war in 1914 interrupted artistic life throughout most of Europe. The activities of a group of painters, writers, and musicians who sought refuge in Zürich reflected the disorientation and disillusion of the time. Dada, as the movement was called, owed much to the iconoclasm of the Cubists and to the polemical tactics of the Futurists. Nonetheless its attack on art was fundamentally artistic; one wing of the avant-garde has owed allegiance to the Dadaist tradition ever since. As well as the need continually to attack the limits of the fine arts, it was felt important to "épater ["shock"] les bourgeois." The Dadaists enlarged the field open to artists in three ways. They questioned the idea that some subjects were simply not relevant to painting, a question that had been hovering over art for some time, by the simple expedient of arguing that anything and everything was fair game. The repetitive and amorphous trends of Impressionism had in fact already given grounds for such a supposition. The next step was to make a reluctant public accept that any object was a work of art if an artist chose to proclaim it one. In 1914 Marcel Duchamp, the exhibitor of serial images of movement in the Section d'Or, produced a bottle rack bought in a Paris store. Better and more *épatant* still, he submitted a urinal to a New York exhibition under a pseudonym in 1917. Duchamp did not paint again, and this is perhaps the single Dadaist gesture that time has failed to reconcile with art. It was also the Dadaists who posed the question, if art (as Redon had realized) is not within the reach of will, how is it different from chance? Jean Arp made collages and then reliefs from random shapes obtained "according to the laws of chance." Of all modern artists, he examined most closely the side of art akin to humour. Similarly, the Dadaists explored such elements as incongruity and dissociation, a process that led the way to Surrealism. Finally and almost incidentally, they asked, if the presentation of movement is proper to art, why not movement itself? Viking Eggeling and Hans Richter, with animated drawings and film, made the first works in a kinetic tradition that even by the late 1980s showed no sign of abating.

"Metaphysical painting"

The painter who, more than any other, focused on incongruity—a feature that in painting involves the reinstatement of the subject, rather than its treatment, at the centre of art—was Giorgio De Chirico, an Italian born in Greece. De Chirico, rooted in the Mediterranean world, created from 1910 onward unforgettable images of its dereliction (Figure 16). In the immediate postwar years, he pioneered a style of emblematic, half-abstract still life called *pittura metafisica* ("metaphysical painting"), but by 1924, when the Surrealists began to work a similar vein of fantasy, De Chirico had changed; and in later life he disavowed his early achievement. Metaphysical painting had one unexpected sequel, the serene realism of Giorgio Morandi. Meanwhile, Kurt Schwitters in Germany developed the mediums of collage and assemblage in the new spirit. Francis Picabia, who was associated with Duchamp in the United States during the war, joined forces with the Swiss Dadaists in 1918; his contribution was an epigrammatic elegance of style. The German Max Ernst was the most resourceful pictorial technician of the movement and a continually fertile inventor.

The Peggy Guggenheim Collection, Venice, Solomon R. Guggenheim Foundation, New York City; photograph, Arts Council of Great Britain

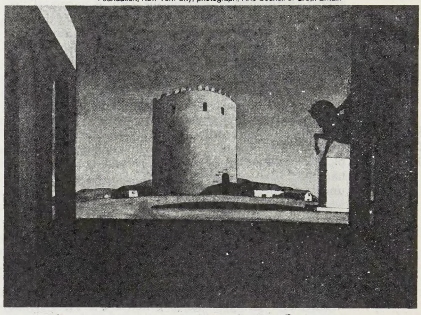


Figure 16: "The Rose Tower," by Giorgio De Chirico, 1913. In the Solomon R. Guggenheim Museum, New York City: The Peggy Guggenheim Collection, Venice. 73 × 100 cm.

It was in 1917 that the term Surrealism was coined, when the poet Guillaume Apollinaire described the style of the ballet *Parade*, for which Picasso had painted the sets, as:

> a sort of sur-realism in which I see a point of departure for a series of manifestations of that New Spirit which . . . promises to transform arts and manners from top to bottom with universal joy.

The Surrealist movement

The manifesto of the Surrealist movement, which was composed by the poet André Breton, did not appear until 1924, however. Surrealism meant different things in different people's hands, but a common feature was absorption in the fantastic and irrational. The questions posed by Dada also preoccupied Surrealists, but for them the problem of the involuntary, fortuitous element in art, for example, was clearly open to psychological solution. The Surrealists demanded "pure psychic automatism"; the automatic drawings that the French artist André Masson made from 1925 onward and, on a more mechanical level, the frottage ("rubbing") devices of Ernst, which added to painting the evocative effect of fortuitously dappled textures, introduced an element that flourished even more fully 20 years later. Another discovery made in the wake of Dada was similarly delayed in its full impact: *Parade* had been the culmination of a series of musical compositions by Erik Satie that were based on ironic quotations of popular material. In the early 1920s the Americans Stuart Davis and Man Ray made paintings out of the designs on commercial packaging, foreshadowing the Pop Art of the 1950s.

The greatest achievement of Surrealist painting, however, was the invention of a new genre: fantastic realism—the prosaic, indeed quasi-photographic, rendering of the forms of fantasy and dream. The invention was the work, after De Chirico, of the Frenchman Yves Tanguy and the Spaniard Salvador Dalí. In the pictorial world of Dalí, everyday things undergo a transformation (Plate 25) that can be almost disturbing; in that of Tanguy, forms are more suggestive than related to actual objects. A different aspect of this dream realism, one that is particularly disturbing, was shown by the Belgian René Magritte.

In the years after 1918, a mood of classical consolidation affected some painters. In Germany a "New Objectivity"



Figure 17: "Guernica," oil on canvas by Pablo Picasso, 1937. In the Prado, Madrid. 3.51 × 7.82 m.

© S.P.A.D.E.M., Paris/V.A.G.A., New York City, 1984; photograph, the Museum of Modern Art, New York City

(Die Neue Sachlichkeit) was imposed on Expressionism; the eventual synthesis appeared in the brutal paintings of Max Beckmann (Plate 24). In France the Italian-born Amedeo Modigliani, affected by the simplicity of the Romanian-born sculptor Constantin Brancusi, arrived at a delicate linear realism, the last of the great Postimpressionist styles.

The Expressionist tradition was developed to an extreme of agonized distortion by Chaim Soutine. Another Russian-born member of the school of Paris, Marc Chagall, who had been influenced both by Cubism and the Russian avant-garde, discovered in the 1920s an individual and inconsequent vein of pictorial fancy (Plate 25). The sombre and devotional art of the Frenchman Georges Rouault bore the marks of his training with Gustave Moreau and as a stained-glass maker. Its crude force had been developed in the context of Fauvism, but the vision was one of refined introspection. The vigour and freedom of Fauvism was developed in the opposite direction in the decorative, extrovert style of another French painter, Raoul Dufy. The classicizing trend of the 1920s had a remarkable sequel in the work of the mural painters of Mexico. One such, Diego Rivera, had learned the formal lessons of Cubism in Paris; José Clemente Orozco was more dependent on the folk art of his country. Their frescoes combined grandeur with a legibility and social awareness rare in modern art.

Mural painters of Mexico

The greatest imaginative achievements between World Wars I and II were, however, again those of Picasso. In the years immediately following World War I he had painted a series of solidly modeled yet oddly ironic figure pictures. Then his mood changed, and in 1925 "The Three Dancers" (Tate Gallery, London) reintroduced an anarchic and convulsive quality. The ambiguities and transformations of his art, both in painting and sculpture, have an emotional character that is entirely his own, but the enlargement of the artistic language greatly influenced others. The metamorphosis of natural shape into abstracted forms that nevertheless curve and bulge with their own life, a metamorphosis initiated by Picasso, became the international style of the early 1930s. The Spaniard Joan Miró gave it his own clarity and gaiety (Plate 24). Biomorphic abstraction, in essence the method of Tanguy, extended the resources of Surrealism, and the Chilean Roberto Matta Echaurren, who began painting in 1938, used it with dramatic effect. A poetic version of the style, rooted in an emotional response to landscape, was evolved in England by Graham Sutherland. In the later 1930s, with "Guernica" (1937; Prado, Madrid; Figure 17) and other pictures, Picasso responded to specific events. Around 1940, two painters in the United States, Arshile Gorky and Willem de Kooning, gave the biomorphic style a new character: relaxed, diffuse, and clear.

The development of abstract painting between the wars was comparatively slow. Klee (in 1921) and Kandinsky (in 1922) gravitated to the Bauhaus, the school in Germany whose work at Weimar and later at Dessau deeply influenced architecture and design as well as basic teaching. Oskar Schlemmer, whose simplified manner paralleled the Italian metaphysical painters, and Lyonel Feininger, an American-born painter working in a style developed from Cubism, were already teaching there. Kandinsky was concerned with refining the geometric ingredient of his work. Klee developed the poetic and fantastic elements of his art with an inconsequent fertility (Figure 18). The systematic purity of the Bauhaus approach survived longest in the work of Josef Albers, who moved to the United States in 1933. In 1940 Mondrian moved to New York City, and his last dynamic pictures reflect the new environment. Mondrian's work was appreciated by only a small circle, although a similar strength of purpose with a delicate re-

Abstract painting between the wars

By courtesy of the Museum of Modern Art, New York City

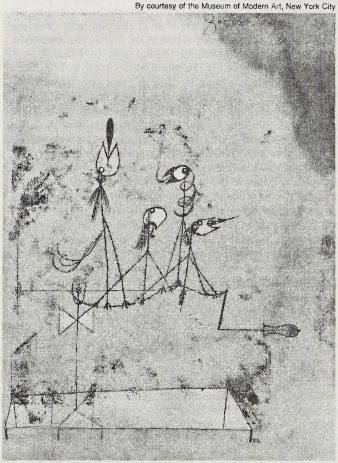


Figure 18: "Twittering Machine," mixed media (watercolour and pen and ink) by Paul Klee, 1922. In the Museum of Modern Art, New York City. 41.3 × 30.5 cm.

sponsiveness to a broader range of forms appeared in the work of the British painter Ben Nicholson. In the 1930s some paintings were executed by artists who formed themselves into groups, such as Abstraction-Création in Paris, Unit One in London, and the Association of American Abstract Artists in New York City. The work of these groups attained wider recognition only after World War II.

After 1945. The postwar work of Braque developed a few basic themes. The space and content of "The Studio" series of five paintings were formulated in vertical phases of varying sombreness; a mysterious bird that featured in this series was a symbol expressive of aspiration. Nicolas de Staël, a friend of Braque who was born in St. Petersburg, reached in 1950 a style in which lozenges of solid paint were built into structures of echo and correspondence. Colour in itself provided the substance, and de Staël's influence was considerable. The painterly and basically traditional vein of abstraction pursued in Paris by such painters as Alfred Manessier remained at root decorative. In Italy, traditional trends in sculpture are reflected in the brilliant accomplished modeling of Giacomo Manzù; Marino Marini, devoting himself almost entirely to the single theme of horse and rider, gave a bald realistic style an oddly apocalyptic force.

The "Cobra" group

The Expressionist tradition was revived in the new spirit by the "Cobra" group of painters from Copenhagen, Brussels, and Amsterdam who came together in Paris in 1948. In the work of Asger Jorn and Karel Appel, the image springs as if by chance from the free extempore play of brushstrokes. Surrealism proved remarkably durable. Among its adherents, the American Joseph Cornell had been evolving from the techniques of collage and assemblage a personal and evocative form of image; the Pole Hans Bellmer and the German Richard Lindner, working in Paris and New York, respectively, explored private and obsessive themes; they were recognized as among the most individual talents of their generation. In general, the most idiosyncratic and anarchic qualities of art were being developed as a new tradition, while geometric abstraction was seen to be the natural basis for the arts that are public and communal in purpose. Victor Pasmore in Britain, for instance, abandoned his earlier Postimpressionist standpoint to start afresh with constructional and graphic symbols deriving from Klee and Mondrian.

The presence of a number of the pioneer Surrealists in the United States during World War II affected later developments there. Surrealism's element of psychic automatism, particularly the spontaneous calligraphy of Masson, was particularly influential. The possibilities had, in fact, been implicit in modern painting for at least two decades; in Paris in the 1920s Jean Fautrier was already basing pictures on spontaneous and informal gestures with paint. In the United States in the 1940s, however, fresh impetus came from the impulsive play of colour in the work of the influential teacher Hans Hofmann. The movement that became known as Abstract Expressionism represented a decisive departure from its European sources, not only because the homogeneous consistency of a painted surface in itself took on a new meaning in the expansive U.S. conditions but at least equally because of the exceptional personality of Jackson Pollock. The style Pollock adopted in 1947 reflected an original involvement in the act of painting that transcended deliberation or control (Plate 26). The influential critic Harold Rosenberg wrote in 1952:

Abstract Expressionism

> At a certain moment the canvas began to appear to one American painter after another as an arena in which to act—rather than as a space in which to reproduce, redesign, analyze or "express" an object, actual or imagined. What was to go on the canvas was not a picture but an event.
>
> (*The Tradition of the New*)

In contrast to Pollock's work, that executed at the time by Willem de Kooning, though equally sweeping and ungovernable, showed a recurrent figurative reference; his series of alarming variations on the theme "Woman" began in 1950. Another Abstract Expressionist, Franz Kline, claimed, in executing his shapes like huge black and white ideograms, to be in some sense depicting figurative images. Rosenberg dubbed the group "action painters." In the course of the 1950s their influence was felt in almost every country. The climate of artistic opinion that spread outward from New York made possible flamboyant gesture paintings such as those of the French-born Georges Mathieu.

"Action painting"

The idea of painting as a homogeneous allover fabric led at the same time to other quite separate developments. Prompted by the primitive and psychotic imagery that he called *l'art brut* ("raw art"), Jean Dubuffet embarked on an extraordinarily resourceful series of experiments in translating the raw material of the world into pictures. The energy that fills the works of the American painter Mark Tobey is by comparison gentle and lyrical and much influenced by East Asian art. Dubuffet's example inspired the abstract "matter" painting that developed in several countries around 1950. At its best, as in the work of the Catalan Antonio Tapies, this style conveys a strong sense of natural substance.

L'art brut

New forms. In painting generally a new directness was strikingly combined with a new simplicity. Beginning at the age of 80, in the five years before his death in 1954, Henri Matisse made a series of large gouaches *découpées* in which the increasingly abstract images were created solely by the juxtaposition of sharply cut patches of brilliant colour. Their influence was widespread and by no means confined to painters, such as the American Ellsworth Kelly, who developed the vibrant interaction of hard-edge colour areas. Even from other starting points, painters were reaching similar conclusions. The very simple yet resonant colour combinations of the New York painter Mark Rothko (Plate 26) or the grand severity of another American, Barnett Newman, furnish examples.

New directness and simplicity

Abstract painting was revealing far wider potentialities than had been apparent between World Wars I and II, but figurative styles showed a new freedom as well. The Swiss Alberto Giacometti, who had worked as a Surrealist, evolved in both sculpture and painting his sensation of the visual impact of figures in space. Francis Bacon in Britain uncovered unexpected and startling connotations in the apparition of a human likeness on canvas.

Painting in the 1960s not only sought originality; it took up a deliberately extreme position that may have seemed almost to pass the bounds of art. Paintings might be extremely large. Alternatively, they might be extreme in some other respect, such as the canvases of the Frenchman Yves Klein, which showed only a plain, arresting blue colour, or the black pictures of the American Ad Reinhardt, with variations so slight as to be hardly perceptible. The element of apparent chance in action painting explained the way the stains of colour in the work of the American painter Morris Louis appeared to flow and soak across the canvas as if of their own accord.

The tradition of Dada and its skepticism regarding what had once been the received definition of art prompted continual experiment with the techniques of assemblage. Robert Rauschenberg in the United States sought to place his subtly calculated "combine paintings" (collections of contrasting objects joined to make an ensemble) in the gap between reality and art, contrasting the significance of paint with the borrowed imagery and objects juxtaposed with it. Jasper Johns, an associate of Rauschenburg's, worked with preexisting designs such as targets and the U.S. flag, giving them an ironic look when subjected to incorporation in his works. In the borrowed imagery and popular quotations, on which much painting was based in the years that followed, the irony was intensified to the point of ceasing to be irony at all. Roy Lichtenstein took strip cartoons and other banal (even banally artistic) imagery as the motifs for pictures (Plate 26). Another American, Claes Oldenburg, began by reconstructing common things out of the random pictorial substance of Abstract Expressionism; his later reconstructions of the rigid furniture of everyday life are tailored out of limp plastic sheeting.

Pop Art and hard-edge painting

There is nothing random about the typical art of the 1960s. On the contrary, it was planned exactly and normally carried out by an efficient, almost mechanical-seeming system. Hard-edge painting developed into a wide range of planar styles having in common only their exploitation of optical reactions and the element of shock that

is the visual concomitant of sharp contrast. The spread of this idiom was particularly influenced by the Hungarian-born Victor Vasarely, who worked in France; its most personal development was in the largely monochromatic work of Bridget Riley in Britain (Figure 19). Again, the initial tendency was to exclude such sensations from the aesthetic canon, but in the event a whole region of visual meaning, void for uncertainty since abstraction began, was reclaimed for painting. Optical art, or Op Art, emphasizes movement, whether potential, actual, or relative, and such effects have been ingeniously investigated by the *Groupe de Recherche d'Art Visuel* ("Group for Visual Research"), founded in Paris in 1960, and the Zero group in Düsseldorf, Ger. In the reliefs of the Venezuelan Jesús Raphael Soto, shifting vision is given a delicate order.

Op Art

By courtesy of the trustees of the Tate Gallery, London

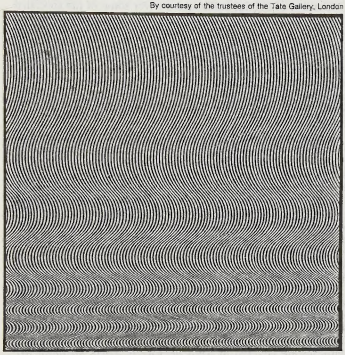


Figure 19: "Fall," by Bridget Riley, 1963. In the Tate Gallery, London. 1.41 × 1.41 m.

Other developments have proved more fertile. In the hands of the American painters Kenneth Noland and Frank Stella, painting discovered new shapes, both within the rectangular canvas and beyond it. The new value given to the painted plane did not benefit painting only. The British painter Richard Smith deployed it in three dimensions in painted constructions that re-create impressions of commercial packaging in terms of the spatial imagination of the arts.

Minimal art

The extreme in this reduction of means and sophistication of aesthetics was perhaps reached when a group of sculptors in the United States and England turned to investigate the possibilities of minimal and primary forms, normally the simplest geometric solids, alone or arranged in baldly repetitive series. Here, it is the spectator (as perhaps in a sense it always is) who brings the interpretation and supplies the art. The proposition had the apparent preposterousness expected of avant-garde art, yet it seemed likely, in its turn, to shed light on problems that are very much older. It is characteristic of sculpture and painting in the 20th century to deal more and more consciously and directly with the ultimate definition of art. The perennial compulsion to reverse previously accepted definitions has threatened ever more directly the recognizable identity of art. At the end of the 1960s the tendency to emphasize the systems and attitudes of art rather than its product led to a move in several countries to deny the validity of the art object. Instead artists prepared written specifications for ideal, imaginary art, the fulfillment of which was superfluous, or self-sufficient programs for performances paradoxically analogous to some aspect of the more familiar artistic acts. Conceptual art has opened the way to activities notable in their defiance of conventional expectations. The designs of earth art or earthworks are fulfilled by moving large amounts of soil, preferably in inaccessible places, perhaps in token of the potency of traditional art to impose its shape on the world. Activities of this order may appear to belong as much to theory as to artistic creation itself; in the 1970s these distinctions, with other familiar cornerstones of artistic thought, were held to have lost their validity. (La.G./Ed.)

In the 1970s, critical and public interest centred on the reductive constructions of Minimalism and the nihilistic questioning of conceptual art. The late 1970s and early 1980s, however, witnessed a resurgence of excitement in painting and a return to figurative representation. A new movement called Neo-Expressionism arose in New York City and in the art capitals of western Europe, especially in West Germany, combining the heavy paint surfaces and dynamic brushstrokes of Abstract Expressionism with the emotional tone of early 20th-century German Expressionism. The new movement's subject matter ranged from basically literal, though self-consciously primitive, treatments of the human figure to a range of imaginary subjects indicative of modern urban life, particularly its glamour, alienation, and menace. A notable characteristic of Neo-Expressionism was the newly prominent role played in its commercial acceptance by gallery owners and art dealers who adroitly publicized the movement's artists. Indeed, Neo-Expressionism's sudden success was an indication of the growing commercialization of the avant-garde and its unhesitating acceptance by wealthy, influential collectors and progressive-minded museum curators. Some critics voiced doubts over what they saw as the reflexive pursuit of artistic novelty under the influence of commercial pressures, and some even asserted that critical and public acclaim had to some extent become divorced from the goals of finding and patronizing painters whose works had lasting artistic significance. (Ed.)

Neo-Expressionism of the 1970s

BIBLIOGRAPHY

General: Among the standard surveys of the history of Western art are HELEN GARDNER, *Gardner's Art Through the Ages*, 8th ed., edited by HORST DE LA CROIX and RICHARD G. TANSEY (1986); H.W. JANSON, *History of Art*, 3rd ed., revised and expanded by ANTHONY F. JANSON (1986); and DAVID M. ROBB and J.J. GARRISON, *Art in the Western World*, 4th ed. (1963). Among the major standard reference works are the *Encyclopedia of World Art*, 16 vol. (1959–83); ULRICH THIEME and FELIX BECKER (eds.), *Allgemeines Lexikon der bildenden Künstler von der Antike bis zur Gegenwart*, 37 vol. (1907–50, reprinted 1970–71); *McGraw-Hill Dictionary of Art*, edited by BERNARD S. MYERS, 5 vol. (1969); HAROLD OSBORNE (ed.), *The Oxford Companion to Art* (1970, reprinted 1984); and the *Praeger Encyclopedia of Art*, 5 vol. (1971). See also ARNOLD HAUSER, *The Social History of Art*, trans. from German, 4 vol. (1962); ELIZABETH G. HOLT (ed.), *A Documentary History of Art*, 3 vol. (1957–66, vol. 1 and 2 reprinted in 1982); PETER MURRAY and LINDA MURRAY, *A Dictionary of Art and Artists*, 4th ed. (1976); and E.H. GOMBRICH, *The Story of Art*, 14th rev. ed. (1984). An extensive bibliography may be found in E. LOUISE LUCAS, *Art Books: A Basic Bibliography on the Fine Arts* (1968). Later reference works include the *Larousse Dictionary of Painters* (1981; originally published in French, 1976); and DAVID PIPER (ed.), *The Random House Library of Painting and Sculpture*, 4 vol. (1981). F. DAVID MARTIN, *Sculpture and Enlivened Space: Aesthetics and History* (1981), examines the comparative importance of painting and sculpture in Western art.

European Metal Age cultures: STUART PIGGOTT, *Ancient Europe, from the Beginnings of Agriculture to Classical Antiquity* (1965); and WALTER TORBRUGGE, *Prehistoric European Art*, (1968; originally published in German, 1968), provide general surveys. There is a vast literature on the Aegean and eastern Mediterranean, including SPYRIDON MARINATOS, *Crete and Mycenae* (1960; originally published in Greek, 1959), photographs by MAX HIRMER; REYNOLD HIGGINS, *Minoan and Mycenaean Art*, rev. ed. (1981); and SPYRIDON MARINATOS, "Life and Art in Prehistoric Thera," *Proceedings of the British Academy*, vol. 57 (1972). SIR ARTHUR EVANS, *The Palace of Minos: A Comparative Account of the Successive Stages of the Early Cretan Civilization as Illustrated by the Discoveries at Knossos*, 4 vol. in 6 (1921–35, reissued 1964), is a classic work but is superseded in part by RICHARD W. HUTCHINSON, *Prehistoric Crete* (1962, reprinted 1968); SINCLAIR HOOD, *The Minoans: The Story of Bronze Age Crete* (1971); and KEITH BRANIGAN, *The Foundations of Palatial Crete* (1970). For Helladic art, the best works are GEORGE E. MYLONAS, *Mycenae and the Mycenaean Age* (1966); and LORD WILLIAM TAYLOUR, *The Mycenaeans*, rev. ed. (1983). For Cyprus, see VASSOS KARAGEORGHIS, *Cyprus* (1969), with a useful bibliography. Among the numerous works on the western Mediterranean are the

following: ANTONIO ARRIBAS, *The Iberians* (1964); L. BERNABÒ BREA, *Sicily Before the Greeks* (rev. ed., 1966; originally published in Italian, 1966); DAVID H. TRUMP, *Central and Southern Italy Before Rome* (1966); DAVID RANDALL-MACIVER, *The Iron Age in Italy: A Study of Those Aspects of the Early Civilization Which Are neither Villanovan nor Etruscan* (1927, reprinted 1974) and *Villanovans and Early Etruscans: A Study of the Early Iron Age in Italy as It Is Seen near Bologna, in Etruria and in Latium* (1924); MARIO MORETTI and GUGLIELMO MAETZKE, *The Art of the Etruscans* (1970; originally published in Italian, 1969); P.J. RIIS, *An Introduction to Etruscan Art* (1953; originally published in Danish, 1948); RAYMOND BLOCH, *The Etruscans* (1958; originally published in French, 1958); MASSIMO PALLOTINO, *Etruscan Painting*, trans. from French (1952), and *Art of the Etruscans*, trans. from French (1955); MARGARET GUIDO, *Sardinia* (1964); and JOHN D. EVANS, *Malta* (1959). For northern European art, see PAUL JACOBSTHAL, *Early Celtic Art*, 2 vol. (1944, reprinted 1970). See also JOHN E. PFEIFFER, *The Creative Explosion: An Inquiry into the Origins of Art and Religion* (1982); ANDRÉ LEROI-GOURHAN, *The Dawn of European Art: An Introduction to Palaeolithic Cave Painting* (1982; originally published in Italian, 1981); and PIERRE AMIET *et al., Art in the Ancient World: A Handbook of Styles and Forms*, trans. from French (1981).

Ancient Greek: Important general works include JOHN D. BEAZLEY and BERNARD ASHMOLE, *Greek Sculpture and Painting to the End of the Hellenistic Period* (1932, reprinted 1966); JOHN BOARDMAN, *Greek Art*, new rev. ed. (1985); JOHN BOARDMAN *et al., Greek Art and Architecture* (1967; originally published in German, 1966), with photographs by MAX HIRMER; RHYS CARPENTER, *The Esthetic Basis of Greek Art of the Fifth and Fourth Centuries B.C.*, rev. ed. (1959); GISELA RICHTER, *A Handbook of Greek Art*, 8th ed. (1983), and *Archaic Greek Art Against Its Historical Background* (1949); GEORGE BOAS (ed.), *The Greek Tradition* (1939), a collection of essays published in connection with an exhibition and accompanied by its catalog, GEORGE BOAS *et al., The Greek Tradition in Painting and the Minor Arts* (1939); BERNHARD SCHWEITZER, *Greek Geometric Art* (1971; originally published in German, 1969); and VINCENT J. BRUNO, *Form and Color in Greek Painting* (1977). Among the standard studies on vase painting are PAOLO E. ARIAS, *A History of 1000 Years of Greek Vase Painting* (1962; originally published in Italian, 1960), with photographs by MAX HIRMER; JOHN D. BEAZLEY, *The Development of Attic Black-Figure* (1951, reprinted 1986), *Potter and Painter in Ancient Athens* (1944), *Attic Black-Figure Vase-Painters* (1956, reprinted 1978), and *Attic Red-Figure Vase-Painters*, 2nd ed., 3 vol. (1963, reissued 1984); ERNST BUSCHOR, *Greek Vase Painting* (1921, reprinted 1971; originally published in German, 1913); JOSEPH V. NOBLE, *The Techniques of Painted Attic Pottery* (1965); and MARTIN ROBERTSON, *Greek Painting* (1959; reissued 1979). See also MANOLIS ANDRONIKOS, *Vergina: The Royal Tombs and the Ancient City* (1984; originally published in Greek, 1984). MARTIN ROBERTSON, *A Shorter History of Greek Art* (1981), is an excellent condensed study of Greek art from the Geometric through the Hellenistic periods.

Roman: General surveys include GEORGE M.A. HANFMANN, *Roman Art: A Modern Survey of the Art of Imperial Rome* (1964, reissued 1975); HEINZ KÄHLER, *The Art of Rome and Her Empire* (1963; originally published in German, 1962); GERMAN HAFNER, *Art of Rome, Etruria, and Magna Graecia*, trans. from German (1969); G.A. MANSUELLI, *The Art of Etruria and Early Rome*, trans. from Italian (1965); JOCELYN M.C. TOYNBEE, *Art in Roman Britain* (1962); and SIR MORTIMER WHEELER, *Roman Art and Architecture* (1964, reissued 1985). More detailed information on painting may be found in AMEDEO MAIURI, *Roman Painting*, trans. from French (1953); and VLADIMIRO DORIGO, *Late Roman Painting* (1971; originally published in Italian, 1966). BERNARD ANDREAE, *The Art of Rome* (1978; originally published in German, 1973), is a comprehensive survey; and OTTO J. BRENDEL, *Prolegomena to the Study of Roman Art* (1979), includes criticism of previous writings on the subject.

Early Christian, Byzantine, Armenian, Georgian, and Coptic: VIKTOR LAZAREV, *Storia della pittura bizantina* (1967; originally published in Russian, 2 vol., 1947–48), is a well-annotated general survey, now seriously out of date but still useful. Good coverage is provided by a combination of two books: BEAT BRENK (ed.), *Spätantike und frühes Christentum* (1977); and WOLFGANG F. VOLBACH and JACQUELINE LAFONTAINE-DOSOGNE, *Byzanz und der christliche Osten* (1968). See also WOLFGANG F. VOLBACH, *Early Christian Art*, with photographs by MAX HIRMER (1962; originally published in German, 1958); and D. TALBOT RICE, *Byzantine Art*, rev. ed. (1968); THOMAS F. MATHEWS, *The Byzantine Churches of Istanbul: A Photographic Survey* (1976); and CYRIL MANGO, *The Art of the Byzantine Empire, 312–1453: Sources and Documents* (1972, reprinted 1986).

A number of books consider the art of the period as a whole and propose various ways of interpreting the material: ANDRÉ GRABAR, *Christian Iconography: A Study of Its Origins* (1968, reprinted 1980; originally published in French, 1957; 2nd rev. French ed., 1984), *The Beginnings of Christian Art, 200–395* (1967; originally published in French, 1966), *Byzantium: From the Death of Theodosius to the Rise of Islam*, trans. from French (1967), and *Byzantine Painting: Historical and Critical Study*, trans. from French (1953, reissued 1979); ERNST KITZINGER, *Byzantine Art in the Making: Main Lines of Stylistic Development in Mediterranean Art, 3rd–7th Century* (1977); CLIVE FOSS and PAUL MAGDALINO, *Rome and Byzantium* (1977); and CYRIL MANGO, *Byzantium, the Empire of New Rome* (1980). HENRY MAGUIRE, *Art and Eloquence in Byzantium* (1981), discusses art as influenced by sermons and other theological writings; alternatively, ROBIN CORMACK, *Writing in Gold: Byzantine Society and Its Icons* (1985), interprets Byzantine art as a major element in the Byzantine outlook.

A number of helpful publications cover some of the specialist areas: KLAUS WESSEL, *Coptic Art* (1965; originally published in German, 1963); GUISEPPE BOVINI, *Ravenna: Its Mosaics and Monuments* (1956, reissued 1970; originally published in Italian, 1956); HEINZ KÄHLER, *Hagia Sophia* (1967; originally published in German, 1967); ANTHONY BRYER and JUDITH HERRIN (eds.), *Iconoclasm* (1977); KURT WEITZMANN, *The Icon: Holy Images—Sixth to Fourteenth Century* (1978); and KURT WEITZMANN *et al., The Icon* (1982; originally published in Italian, 1981).

Early medieval and Romanesque: The most satisfactory survey of the whole period is provided in ANDRÉ GRABAR and CARL NORDENFALK, *Early Medieval Painting from the Fourth to the Eleventh Century*, trans. from French and German (1957), and *Romanesque Painting from the Eleventh to the Thirteenth Century*, trans. from French and German (1958). The most exhaustive study on wall painting is still EDGAR W. ANTHONY, *Romanesque Frescoes* (1951, reprinted 1971). For manuscript illumination, see ALBERT BOECKLER, *Abendländische Miniaturen bis zum Ausgang der romanischen Zeit* (1930); a stimulating discussion by OTTO PÄCHT, *Book Illumination in the Middle Ages: An Introduction* (1986; originally published in German, 1984); and a survey by CHRISTOPHER DE HAMEL, *A History of Illuminated Manuscripts* (1986). An excellent short introduction to the subject is contained in ERNST KITZINGER, *Early Medieval Art, with Illustrations from the British Museum and British Library Collections*, rev. ed. (1983). The 9th through 12th centuries are covered in JOHN BECKWITH, *Early Medieval Art: Carolingian, Ottonian, Romanesque*, rev. ed. (1969, reprinted 1985); and C.R. DODWELL, *Painting in Europe, 800 to 1200* (1971). The most comprehensive modern overview of Dark Age and early medieval painting is CARLO BERTELLI, "Traccia allo studio delle fondazioni medievali dell'arte italiana," in *Storia dell'arte italiana*, part 2, *Dal medioevo al novecento*, vol. 1, *Dal medioevo al quattrocento*, pp. 3–163 (1983). For Rome, see RICHARD KRAUTHEIMER, *Rome, Profile of a City, 312–1308* (1980). The best survey of early Anglo-Saxon art is still T.D. KENDRICK, *Anglo-Saxon Art to A.D. 900* (1938, reprinted 1972). For book illumination in Britain and Ireland, see J.J.G. ALEXANDER, *Insular Manuscripts, 6th to the 9th Century* (1978); CARL NORDENFALK, *Celtic and Anglo-Saxon Painting: Book Illumination in the British Isles, 600–800* (1977); FRANÇOISE HENRY, *Irish Art in the Early Christian Period, to 800 A.D.* (1965), *Irish Art During the Viking Invasions, 800–1020 A.D.* (1967), and *Irish Art in the Romanesque Period, 1020–1170 A.D.* (1970; originally published in French together in a set of 3 vol., 1963–64).

Merovingian illumination is discussed in JEAN HUBERT, JEAN PORCHER, and WOLFGANG F. VOLBACH, *Europe in the Dark Ages* (1969; U.S. title, *Europe of the Invasions;* originally published in French, 1967). There are several good surveys of Carolingian painting: WOLFGANG BRAUNFELS, *Die Welt der Karolinger und ihre Kunst* (1968); JEAN HUBERT, JEAN PORCHER, and WOLFGANG F. VOLBACH, *Carolingian Art* (1970; U.S. title, *The Carolingian Renaissance;* originally published in French, 1968); and FLORENTINE MÜTHERICH and JOACHIM E. GAEHDE, *Carolingian Painting* (1976). Excellent particular studies can be found in WOLFGANG BRAUNFELS and HERMANN SCHNITZLER (eds.), *Karolingische Kunst* (1965). Late Anglo-Saxon art is considered in ELZBIETA TEMPLE, *Anglo-Saxon Manuscripts, 900–1066* (1976); JANET BACKHOUSE, D.H. TURNER, and LESLIE WEBSTER (eds.), *The Golden Age of Anglo-Saxon Art, 966-1066* (1984); and C.R. DODWELL, *Anglo-Saxon Art: A New Perspective* (1982). HANS JANTZEN, *Ottonische Kunst* (1947, reissued 1963), is an excellent review of Ottonian art; and ADOLPH GOLDSCHMIDT, *German Illumination*, vol. 2, *Ottonian Period* (1929, reprinted as part of a 1-vol. edition, 1970; originally published in German, 1928), is still valuable. A brief survey is LOUIS GRODECKI *et al., Le Siècle de l'an mil* (1973). C.R. DODWELL and D.H. TURNER, *Reichenau Reconsidered: A Re-assessment of the Place of Reichenau in Ottonian Art* (1965), offers a stimulating but

contested examination of late 10th-century Ottonian illumination. The Romanesque period is surveyed in FRANÇOIS AVRIL, XAVIER BARRAL I ALTET, and DANIELLE GABORIT-CHOPIN, *Le Temps des Croisades* (1982), and *Les Royaumes d'Occident* (1983); HANNS SWARZENSKI, *Monuments of Romanesque Art: The Art of Church Treasures in North-Western Europe,* 2nd ed. (1967, reissued 1974); OTTO DEMUS and MAX HIRMER, *Romanesque Mural Painting* (1970; originally published in German, 1968); and WALTER CAHN, *Romanesque Bible Illumination* (1982). The best study of French Romanesque illumination is JEAN PORCHER, *Medieval French Miniatures* (1960; U.K. title, *French Miniatures from Illuminated Manuscripts;* originally published in French, 1959).

For England, there is an excellent survey of manuscripts: C.M. KAUFFMANN, *Romanesque Manuscripts, 1066–1190* (1975); and a later discussion in the exhibition catalog, GEORGE ZARNECKI, JANET HOLT, and TRISTRAM HOLLAND (eds.), *English Romanesque Art, 1066–1200* (1984). The standard work on early Spanish illumination, in English, is J. DOMÍNGUEZ BORDONA, *Spanish Illumination* (1930, reissued 1969; originally published in Spanish, 1930); see also JOHN WILLIAMS, *Early Spanish Manuscript Illumination* (1977). Illumination in northwest Germany is surveyed in two exhibition catalogs, *Rhein und Maas, Kunst und Kultur, 800–1400,* 2 vol. (1972–73); and ANTON LEGNER (ed.), *Ornamenta ecclesiae: Kunst und Künstler der Romanik,* 3 vol. (1985). The Gospels of Henry the Lion are discussed in *The Gospels of Henry the Lion, Count of Saxony, Duke of Bavaria* (1983), an auction catalog of Sotheby Parke Bernet Inc.; and HORST FUHRMANN and FLORENTINE MÜTHERICH (eds.), *Das Evangeliar Heinrichs des Löwen und das mittelalterliche Herrscherbild* (1986).The standard work on Salzburg illumination is still the masterly GEORG SWARZENSKI, *Die Salzburger Malerei von den ersten Anfängen bis zur Blütezeit des romanischen Stils: Studien zur Geschichte der deutschen Malerei und Handschriftenkunde des Mittelalters,* 2nd ed., 2 vol. (1969). The best survey of the art of the late 12th century is found in the three volumes of *The Year 1200,* a set published in conjunction with an exhibition: vol. 1, *A Centennial Exhibition at the Metropolitan Museum of Art: A Catalog,* edited by K. HOFFMANN (1970); vol. 2, *A Background Survey,* edited by F. DEUCHLER (1970); and vol. 3, *A Symposium,* texts by FRANÇOIS AVRIL *et al.* (1975). OTTO DEMUS, *Byzantine Art and the West* (1970), discusses the influence of Byzantine art on western Europe throughout the Middle Ages.

Gothic: General social and intellectual studies of Gothic art include JOAN EVANS (ed.), *The Flowering of the Middle Ages,* new ed. (1985); and JOHAN HUIZINGA, *The Waning of the Middle Ages: A Study of the Forms of Life, Thought and Art in France and the Netherlands in the XIVth and XVth Centuries* (1924, reprinted 1985; originally published in Dutch, 1919). See also ANDREW MARTINDALE, *Gothic Art* (1967, reprinted 1985); GEORGE HENDERSON, *Early Medieval* (1972), and *Gothic* (1967); EMILE MALE, *Religious Art in France, the Thirteenth Century: A Study of Medieval Iconography and Its Sources* (1984; originally published in French, 9th rev. ed., 1958), and *Religious Art from the Twelfth to the Eighteenth Century* (1949, reprinted 1970; originally published in French, 1945); TERESA G. FRISCH, *Gothic Art 1140–c. 1450* (1971); and JACQUES DUPONT and CESARE GNUDI, *Gothic Painting* (1954, reissued 1979; originally published in French, 1954).

For a discussion of Italian Gothic painting, see EVE BORSOOK, *The Mural Painters of Tuscany: From Cimabue to Andrea del Sarto,* 2nd rev. ed. (1980); MILLARD MEISS, *Painting in Florence and Siena After the Black Death* (1951, reprinted 1978); JOHN WHITE, *Art and Architecture in Italy, 1250–1400* (1966), and *The Birth and Rebirth of Pictorial Space,* 2nd ed. (1967); and FREDERICK ANTAL, *Florentine Painting and Its Social Background: The Bourgeois Republic Before Cosimo de'Medici's Advent to Power, XIV and Early XV Centuries* (1948, reprinted 1986). CHARLES D. CUTTLER, *Northern Painting from Pucelle to Bruegel: Fourteenth, Fifteenth, and Sixteenth Centuries* (1968), is an introductory work. No adequate monograph on International Gothic art exists, but, for France, see MILLARD MEISS, *French Painting in the Time of Jean de Berry: The Late Fourteenth Century and the Patronage of the Duke,* 2 vol. (1967), *French Painting in the Time of Jean de Berry: The Limbourgs and Their Contemporaries,* 2 vol. (1974), and *French Painting in the Time of Jean de Berry: The Boucicaut Master* (1968).

For late Gothic art, French painting is surveyed in GRETE RING, *A Century of French Painting, 1400–1500* (1949, reprinted 1979). Netherlandish painting is dealt with in MAX J. FRIEDLÄNDER, *Early Netherlandish Painting,* 14 vol. (1967–76; originally published in German, 1924–37); the best monograph on this early period is ERWIN PANOFSKY, *Early Netherlandish Painting, Its Origins and Character,* 2 vol. (1953, reprinted 1971). ANNE SHAVER-GRANDELL, *The Middle Ages* (1982), is an introduction to medieval art intended for the general reader; WALTER OAKESHOTT, *The Two Winchester Bibles* (1981), is a scholarly study of 12th-century painting; RICHARD I. ABRAMS and WARNER A. HUTCHINSON, *An Illustrated Life of Jesus* (1982), includes an analysis of 94 paintings from the collection of the National Gallery of Art in Washington, D.C., with background notes on the artists; and PETER S. BEAGLE, *The Garden of Earthly Delights: Illustrations Taken from the Paintings of Hieronymus Bosch* (1982), is a creative and entertaining introduction to the artist's works.

Renaissance: (*Italy*): General works include CREIGHTON GILBERT, *History of Renaissance Art: Painting, Sculpture, Architecture Throughout Europe* (1973); MICHAEL LEVEY, *Early Renaissance* (1967, reprinted 1979), and *High Renaissance* (1975); and ROBERT KLEIN and HENRI ZERNER, *Italian Art, 1500–1600: Sources and Documents* (1966). A useful introduction to the theorists of the period is provided by SIR ANTHONY BLUNT, *Artistic Theory in Italy, 1450–1600* (1940, reprinted 1982). See also SYDNEY J. FREEDBERG, *Painting of the High Renaissance in Rome and Florence,* new rev. ed., 2 vol. (1985), and *Painting in Italy, 1500 to 1600,* 2nd ed. (1983), a fine survey of the often-complex movements in 16th-century Italian painting; JAMES BECK, *Italian Renaissance Painting* (1981), a historical survey of works of individual masters; and BRUCE COLE, *The Renaissance Artist at Work: From Pisano to Titian* (1983).

(*Northern Renaissance*): OTTO BENESCH, *Art of the Renaissance in Northern Europe: Its Relation to the Contemporary Spiritual and Intellectual Movements,* rev. ed. (1965), and *German Painting, from Dürer to Holbein,* trans. from German (1966); WOLFGANG STECHOW, *Northern Renaissance Art, 1400–1600: Sources and Documents* (1966); and ALBERT CHÂTELET, *Early Dutch Painting: Painting in the Northern Netherlands in the Fifteenth Century* (1981; originally published in French, 1980). (*France*): SIR ANTHONY BLUNT, *Art and Architecture in France, 1500 to 1700,* 4th ed. (1980), is an authoritative survey. (*Spain and Portugal*): GEORGE KUBLER and MARTIN SORIA, *Art and Architecture in Spain and Portugal and Their American Dominions, 1500 to 1800* (1959), is the only scholarly study in English. (*Central Europe and Russia*): G.H. HAMILTON, *The Art and Architecture of Russia,* 3rd ed. (1983), is a survey of all the arts of Russia.

Many studies have been done since the 1960s dealing exclusively with Mannerism. The most coherent view as a whole is JOHN SHEARMAN, *Mannerism* (1967). See also FRANZSEPP WÜRTENBERGER, *Mannerism: The European Style of the Sixteenth Century* (1963; originally published in German, 1962); and GIULIANO BRIGANTI, *Italian Mannerism* (1962; originally published in Italian, 1961).

Baroque and Rococo: The classic study of Baroque art, HEINRICH WÖLFFLIN, *Renaissance and Baroque* (1964, reprinted 1984; originally published in German, 1888), remains an important basic study. MICHAEL KITSON, *The Age of Baroque* (1966), provides an excellent modern summary. JOHN RUPERT MARTIN, *Baroque* (1977), is a fuller survey. GERMAIN BAZIN, *Baroque and Rococo,* trans. from French (1964, reprinted as *Baroque and Rococo Art,* 1974), covers the entire period in less detail, but it has in no way replaced the basic study by FISKE KIMBALL, *The Creation of the Rococo* (1943, reprinted as *The Creation of the Rococo Decorative Style,* 1980). ARNO SCHÖNBERGER and HALLDOR SOEHNER, *The Age of Rococo* (1960; U.S. title, *The Rococo Age: Art and Civilization of the 18th Century;* originally published in German, 1959), has excellent illustrations and detailed notes. Patronage during the period has been analyzed in depth by FRANCIS HASKELL, *Patrons and Painters: A Study in the Relations Between Italian Art and Society in the Age of the Baroque,* rev. ed. (1980). (*Italy*): DENIS MAHON, *Studies in Seicento Art and Theory* (1947, reprinted 1971), is of fundamental importance for an understanding of Baroque art in Italy. RUDOLF WITTKOWER, *Art and Architecture in Italy, 1600 to 1750,* 3rd rev. ed. (1973, reprinted with corrections 1980), surveys the 17th century with clarity and includes a massive bibliography. ELLIS WATERHOUSE, *Italian Baroque Painting,* 2nd ed. (1969), provides an introduction to the principal painters and stylistic movements of the time; a similar role is performed by MICHAEL LEVEY, *Painting in Eighteenth-Century Venice,* 2nd rev. ed. (1980). (*Latin America*): PAL KELEMEN, *Baroque and Rococo in Latin America,* 2nd ed., 2 vol. (1967), is an introduction to the art of the 17th and 18th centuries, with an exhaustive bibliography. (*Flanders*): The only comprehensive introduction to this period available in English is provided by HORST GERSON and E.H. TER KUILE, *Art and Architecture in Belgium, 1600 to 1800* (1960; originally published in German, 1942). (*Holland*): JAKOB ROSENBERG, SEYMOUR SLIVE, and E.H. TER KUILE, *Dutch Art and Architecture, 1600 to 1800,* 3rd ed. (1977), provides an excellent survey of the period and a large bibliography; while WOLFGANG STECHOW, *Dutch Landscape Painting of the Seventeenth Century,* 2nd ed. (1968, reprinted 1981), is a particularly detailed and valuable study of this important facet of Dutch painting. INGVAR BERGSTRÖM, *Dutch Still-Life Painting in the Seventeenth Century* (1956, reprinted 1983; originally published

in Swedish, 1947), provides a survey of this group of paintings. (*France*): Important surveys are WEND GRAF KALNEIN and MICHAEL LEVEY, *Art and Architecture of the Eighteenth Century in France* (1972); and PHILIP CONISBEE, *Painting in Eighteenth-Century France* (1981). (*England*): ELLIS WATERHOUSE, *Painting in Britain, 1530 to 1790,* 4th ed. (1978); and MARGARET WHINNEY and OLIVER MILLAR, *English Art, 1625–1714* (1957), are both well-illustrated and include bibliographies. (*Central Europe*): The best introduction to Baroque art in central Europe, available in English, is undoubtedly EBERHARD HEMPEL, *Baroque Art and Architecture in Central Europe: Germany, Austria, Switzerland, Hungary, Czechoslowakia, Poland,* trans. from German (1965). (*Scandinavia*): Most of the information available on Scandinavian art of the 17th and 18th centuries is to be found in museum and exhibition catalogs devoted to wider subjects, but the monograph by GUNNAR W. LUNDBERG, *Roslin: Liv och verk,* 3 vol. in 2 (1957), is available, together with the relevant sections in TORBEN HOLCK COLDING, *Aspects of Miniature Painting: Its Origins and Development* (1953).

Neoclassicism, Romanticism, and Realism: Important works on the period in general are WALTER FRIEDLAENDER, *David to Delacroix,* trans. from German (1952, reprinted 1980); and FRITZ NOVOTNY, *Painting and Sculpture in Europe, 1780–1880,* 2nd ed. (1971, reissued 1980). Among the many general studies of Neoclassical art are HUGH HONOUR, *Neo-classicism* (1968, reprinted 1977), a sound introduction; ROBERT ROSENBLUM, *Transformations in Late Eighteenth Century Art* (1967), one of the best studies of the period; and DAVID IRWIN, *English Neoclassical Art: Studies in Inspiration and Taste* (1966), a book dealing exclusively with Neoclassical painting and sculpture in Britain. Among the most important works covering Romanticism are MARCEL BRION, *Art of the Romantic Era: Romanticism, Classicism, Realism* (1966; originally published in French, 1963); WERNER HOFMANN, *The Earthly Paradise: Art in the Nineteenth Century* (1961; originally published in German, 1960); FRANCIS D. KLINGENDER, *Art and the Industrial Revolution,* rev. ed., edited by ARTHUR ELTON (1968); EDGAR P. RICHARDSON, *The Way of Western Art, 1776–1914* (1939, reprinted 1969); FREDERICK ANTAL, *Classicism and Romanticism, with Other Studies in Art History* (1966); T.J. CLARK, *The Absolute Bourgeois: Artists and Politics in France, 1848–1851* (1973, reprinted 1982), and *Image of the People: Gustave Courbet and the 1848 Revolution* (1973, reprinted 1982); ALBERT BOIME, *The Academy and French Painting in the Nineteenth Century* (1971); HUGH HONOUR, *Romanticism* (1979); and MICHEL LE BRIS, *Romantics and Romanticism,* trans. from French (1981). LINDA NOCHLIN, *Realism and Tradition in Art, 1848–1900: Sources and Documents* (1966), and *Realism* (1971), are provocative studies. See also BARBARA NOVAK, *American Painting of the Nineteenth Century: Realism, Idealism, and the American Experience,* 2nd ed. (1979).

Modern: Among the numerous surveys of modern art are H.H. ARNASON, *History of Modern Art: Painting, Sculpture, Architecture, Photography,* 3rd rev. ed., updated by DANIEL WHEELER (1986); JOHN CANADAY, *Mainstreams of Modern Art,* 2nd ed. (1981); JEAN CASSOU, EMILE LANGUI, and NIKOLAUS PEVSNER, *Gateway to the Twentieth Century: Art and Culture in a Changing World* (1962); SAM HUNTER, *Modern American Painting and Sculpture* (1959), *Modern French Painting, 1855–1956* (1956), and *American Art of the 20th Century: Painting, Sculpture, Architecture* (1973); SAM HUNTER and JOHN JACOBUS, *Modern Art: Painting, Sculpture, Architecture,* 2nd ed. (1985); MARCEL BRION *et al., Art Since 1945* (1958); and ROBERT L. HERBERT (ed.), *Modern Artists on Art* (1964). See also BEVERLY WHITNEY KEAN, *All the Empty Palaces: The Merchant Patrons of Modern Art in Pre-Revolutionary Russia* (1983), an original study of important developments in the history of European art; and SIEGFRIED WICHMANN, *Japonisme: The Japanese Influence on Western Art in the 19th and 20th Centuries* (1981; originally published in German, 1980), a broad study including treatments of individual artists.

Important works dealing with modern painting include: WERNER HAFTMANN, *Painting in the Twentieth Century,* 2nd ed., 2 vol. (1965; originally published in German, 1965); BERNARD S. MYERS, *Mexican Painting in Our Time* (1956); GUIDO BALLO, *Modern Italian Painting: From Futurism to the Present Day* (1958; originally published in Italian, 1956); ALAN GOWANS, *The Restless Art: A History of Painters and Painting, 1760–1960* (1966); and HERBERT READ, *A Concise History of Modern Painting,* 3rd rev. ed. (1975). See also PAUL VOGT, *Expressionism: German Painting, 1905–1920* (1980; originally published in German), and *Contemporary Painting* (1981; trans. from German), a survey of international painting mostly of the 1950s and 1960s; GEORGE H. ROEDER, JR., *Forum of Uncertainty: Confrontations with Modern Painting in Twentieth-Century American Thought* (1980); JOHN RUSSELL, *The Meanings of Modern Art,* rev. ed. (1981); FRANK H. GOODYEAR, JR., *Contemporary American Realism Since 1960* (1981); and ABRAHAM A. DAVIDSON, *Early American Modernist Painting, 1910–1935* (1981).

(*Pre-Raphaelites*): CHRISTOPHER WOOD, *The Pre-Raphaelites* (1981); and *Pre-Raphaelites and Academics* (1981), a catalog of the exhibition organized in celebration of the publication of Wood's work. (*Impressionism and Postimpressionism*): LINDA NOCHLIN (ed.), *Impressionism and Post-Impressionism, 1874–1904: Sources and Documents* (1966); HORST KELLER, *Watercolors and Drawings of the French Impressionists and Their Parisian Contemporaries* (1982; originally published in German, 1980); JOHN REWALD, *The History of Impressionism,* 4th rev. ed. (1973, reprinted 1980), and *Post-Impressionism, from van Gogh to Gauguin,* 3rd rev. ed. (1978). (*Fauvism*): GEORGES DUTHUIT, *The Fauvist Painters* (1950; originally published in French, 1949). (*German Expressionism*): PETER SELZ, *German Expressionist Painting* (1957, reprinted 1974). (*Cubism*): JOHN GOLDING, *Cubism: A History and an Analysis, 1907–1914,* 2nd ed. (1968); CHRISTOPHER GRAY, *Cubist Aesthetic Theories* (1953); and ROBERT ROSENBLUM, *Cubism and Twentieth-Century Art,* rev. ed. (1966, reissued 1976). (*Futurism*): MARIANNE W. MARTIN, *Futurist Art and Theory, 1909–1915* (1968, reprinted 1978). (*Suprematism and Constructivism*): CAMILLA GRAY, *The Russian Experiment in Art, 1863–1922,* rev. ed., edited by MARIAN BURLEIGH-MOTLEY (1986). (*De Stijl and Neoplasticism*): H.L.C. JAFFÉ, *De Stijl, 1917–1931: The Dutch Contribution to Modern Art* (1956, reprinted 1986); and MILDRED FRIEDMAN (ed.), *De Stijl: Visions of Utopia* (1982), an exhibition catalog. (*Dada and Surrealism*): ROBERT MOTHERWELL (ed.), *The Dada Painters and Poets: An Anthology,* 2nd ed. (1981); HANS RICHTER, *Dada: Art and Anti-Art* (1966, reprinted 1978; originally published in German, 1964); WILLIAM S. RUBIN, *Dada, Surrealism, and Their Heritage* (1968, reprinted 1982); PATRICK WALDBERG, *Surrealism,* trans. from French (1962, reissued 1978); and HERBERT S. GERSHMAN, *The Surrealist Revolution in France* (1969, reprinted 1974). (*Abstract Expressionism*): BARBARA ROSE, *American Art Since 1900,* rev. and expanded ed. (1975). (*Pop Art and Op Art*): MARIO AMAYA, *Pop Art and After* (1965, reprinted 1972; U.K. title, *Pop as Art: A Survey of New Super Realism*); LUCY R. LIPPARD, *Pop Art* (1966); and JOHN RUSSELL and SUZI GABLIK (compilers), *Pop Art Redefined* (1969).

(P.J.C./R.S.Co./Jo.B.M./N.B.P./Ed.)

Pakistan

Pakistan (the Islāmic Republic of Pakistan; Urdu: Islām-ī Jamhūrīya-e Pākistān) is a country in South Asia. It is bounded to the west by Iran, to the north by Afghanistan, to the northeast by China, to the east and southeast by India, and to the south by the Arabian Sea. The territory has an area (excluding the Pakistani-held part of Jammu and Kashmir) of 307,374 square miles (796,095 square kilometres). The capital is Islāmābād.

Pakistan was brought into being at the time of the Partition of British India in 1947 in order to create a separate homeland for India's Muslims in response to the demands of Islāmic nationalists, demands that were articulated by the All India Muslim League under the leadership of Mohammed Ali Jinnah. From independence in 1947 until 1971, Pakistan (both de facto and in law) consisted of two regions—West Pakistan, in the Indus River basin, and East Pakistan, located more than 1,000 miles (1,600 kilometres) away in the Ganges River delta. In response to grave internal political problems, however, an independent state of Bangladesh was proclaimed in East Pakistan in 1971.

Since 1947 the territory of Jammu and Kashmir, along the western Himalayas, has been disputed between Pakistan and India, with each holding sectors. The two countries have gone to war over the territory three times, in 1948–49, 1965, and 1971.

This article is divided into the following sections:

Physical and human geography 393
  The land 393
    Relief and drainage
    Soils
    Climate
    Plant and animal life
    Settlement patterns
  The people 397
    Ethnic composition
    Linguistic composition
    Religions
    Demographic trends
  The economy 399
    Resources
    Agriculture, forestry, and fishing
    Industry
    Finance
    Trade
    Transportation
  Administration and social conditions 401
    Government
    Education
    Health and welfare
  Cultural life 402
History 403
  Background to partition 403
    The Muslim League and Mohammed Ali Jinnah
    Birth of the new state
  Islāmic Republic of Pakistan 403
    Political decline
    Military government
    Civil War
    Bhutto's regime
    Zia ul-Haq's regime
    The administration of Benazir Bhutto
Bibliography 406

Physical and human geography

THE LAND

Relief and drainage. Pakistan is situated at the western end of the Indo-Gangetic Plain, which is bounded to the north by the mountain wall of the Great Himalayan mountain ranges and their offshoots. It is situated in the northwestern part of the southern Asian subcontinent and may be divided into six natural regions—the northern mountains, the submontane plateau, the Indus Plain, the Balochistān Plateau, the western bordering mountains, and the desert areas.

The Himalayan sector

The northern mountains. The western ranges of the Himalayan mountains occupy the entire northern end of Pakistan, extending about 200 miles (320 kilometres) into the country; they are among the youngest mountains on Earth, having attained their present elevation only within the past one million years. Four of the peaks exceed an elevation of 26,000 feet (8,000 metres), and most rise to heights of more than 15,000 feet. These include such towering peaks as Nānga Parbat (26,660 feet [8,126 metres]; located in Jammu and Kashmir) and K2, also called Godwin Austen (28,251 feet [8,611 metres]). Beyond the Karakoram Range in the extreme north lies the Chinese Uighur Autonomous Region of Sinkiang; to the northwest, beyond the Hindu Kush, are the Pamirs—the "Roof of the World"—where only a narrow strip of Afghan territory separates Pakistan from Tajikistan. The Himalayan massif (mountainous mass) that isolates the South Asian subcontinent from Central Asia was pierced in 1970 when a road was completed across the Karakoram Range, linking the town of Gilgit in Pakistan with Kashgar in Sinkiang, China. This road carries considerable commerce between Pakistan and China.

The northern mountain barrier influences the rainfall pattern in Pakistan by intercepting monsoon (rain-bearing) winds from the south. Melting snow from the mountains also feeds rivers, including the Indus, which emerge from the east-west aligned ranges to flow southward.

The population in this inhospitable region is generally sparse, although in a few favoured places it is dense. In most of the tiny settlements of this region, the usual crop is barley; fruit culture, especially of apricots, is of special importance. Timber, mainly pine, is found in some parts, but its occurrence varies with rainfall and altitude. Many slopes have been denuded of cover by excessive timber felling and overgrazing.

The submontane plateau. Lying below the Himalayas, the submontane plateau has four distinct divisions—the Trans-Indus plains, the Potwar Plateau, the Salt Range, and the Sialkot region.

The Trans-Indus plains, west of the Indus, comprise the hill-girt plateaus of the Vale of Peshāwar, Kohāt, and Bannu, which are oases in the arid, scrub-covered landscape of the North-West Frontier Province. Of these, the Vale of Peshāwar is the most fertile. It was once a flourishing centre of Greco-Buddhist culture. Much of the area is covered by gravelly or clayey alluvial detritus. Rainfall is generally only between 10 and 15 inches (250 and 380 millimetres). Most of the cultivated area in the Vale of Peshāwar is irrigated from canals.

The Vale of Peshāwar

Kohāt is less developed than the Vale of Peshāwar. Rainfall is about 16 inches. Only a small percentage of the cultivated area is canal-irrigated. Nor is its groundwater adequately exploited, although the water table is generally high. Much of the area consists of scrub and poor grazing land. The region is much broken by limestone ridges, and the uneven limestone floor is variously filled with lacustrine clays, gravel, or boulders.

In Bannu, soils are sandy or gravelly, except for rich silts that occur in places. About a fourth of the cultivated area is irrigated. Rainfall is low, amounting to about 11 inches. In Kohāt and Bannu, fat-tailed sheep, camels, and donkeys are raised; wool is an important cash crop.



The Potwar Plateau lies at a height of 1,200 to 1,900 feet and covers an area of about 5,000 square miles east of the Indus and west of the Jhelum River, in the Punjab. It is open, undulating country, consisting mainly of sandstone and covered by varying thicknesses of loess (a loamy deposit formed by the wind), which erodes easily. Rainfall is between 15 and 20 inches, being higher in the northwest; the southwest is very arid. The landscape is dissected and eroded by streams that, during the rains, cut into the land and wash away the soil. The streams are generally deep-set and are of little or no use for irrigation. It is generally a poor agricultural area, with an excessive pressure of population upon its resources.

The Salt Range lies at the southern edge of the Potwar Plateau and has an average height of 2,200 feet, the highest peak being at Sakesar (4,992 feet). It is an extremely arid territory that sharply marks the boundary between the submontane region and the Indus Plain to the south. The Salt Range is of interest to geologists, as it contains the most complete geologic sequence in the world, in which rocks from the Cambrian period (570 to 505 million years ago) to the Pleistocene epoch (1,600,000 to 10,000 years ago) are represented.

The Siālkot region is a narrow submontane area in the northeast. Unlike the Potwar Plateau, it is a rich agricultural region. Rainfall varies from 25 to 35 inches (650 to 900 millimetres), and the water table is high, facilitating well (and tube-well) irrigation; the soil is heavy and very fertile. The population distribution is dense, and the land is divided into small farms on which intensive cultivation is practiced.

The Indus Plain. The Indus Plain, which covers an area of about 200,000 square miles, is Pakistan's most prosperous agricultural region. It is an unrelieved featureless plain of fertile alluvium extending for 650 to 700 miles from the rim of the Potwar Plateau southward to the Arabian Sea. The province of Punjab (literally, "Five Waters") is located in its northern zone, which is enclosed by the Indus and its five lower tributaries (the Jhelum, the Chenāb, the Rāvi, the Beās, and the Sutlej), the tributaries converging to their confluence with the Indus at Mithankot. The northern zone is divided into several interfluves (lands situated between streams) called *doābs*. The largest but the poorest of them is the Sindh Sāgar Doāb, situated between the Indus and the Chenāb rivers, which is mostly a desert. The *doābs* that lie to the east of it, however, are the richest agricultural lands in the country. Until the advent of irrigation, at the end of the 19th century, much of the area was a desolate waste, for there is little rainfall. But irrigation has been a mixed blessing; it has also caused waterlogging and salinity in some places. This problem is being addressed by the construction of the Left Bank Outfall Drain, which, when completed, will be a river the size of the Indus, carrying salt water from the plains of Punjab and Sindh to the Arabian Sea.

The *doābs*

In the southern zone of the Indus Plain lies the province of Sindh, which is named after the Indus, or Sindhu, River. There, the size of the Indus—its waters augmented by the contribution of its five lower tributaries—is very great; its average annual discharge at Sukkur, in northern Sindh, is more than five trillion cubic feet (140 billion cubic metres), including a large amount of waterborne silt. The land in Sindh, consisting of alluvial sands and clays, tends to give way before the Indus flood, and the river frequently changes course. Manchhar, a marshy lake west of the Indus, has an area of 14 square miles at low water

MAP INDEX

Political subdivisions

Azad Kashmir .... 34 22 N 73 28 E
Balochistān ..... 28 00 N 66 00 E
Federally Administered Tribal Areas ..... 33 00 N 70 15 E
Islāmābād ..... 33 40 N 73 10 E
North-West Frontier Province ..... 34 30 N 72 00 E
Northern Areas .. 36 00 N 75 00 E
Punjab ..... 29 00 N 72 00 E
Sindh ..... 25 30 N 69 00 E

Cities and towns

Abbottābād ..... 34 09 N 73 13 E
Ahmadpur East .. 29 09 N 71 16 E
Badīn ..... 24 39 N 68 50 E
Bahāwalnagar .... 29 59 N 73 16 E
Bahāwalpur ..... 29 24 N 71 41 E
Bannu ..... 32 59 N 70 36 E
Bela ..... 26 14 N 66 19 E
Bhakkar ..... 31 38 N 71 04 E
Bhalwāl ..... 32 16 N 72 54 E
Bhera ..... 32 29 N 72 55 E
Chakwāl ..... 32 56 N 72 52 E
Chaman ..... 30 55 N 66 27 E
Chārsadda ..... 34 09 N 71 44 E
Chichāwatni .... 30 32 N 72 42 E
Chiniot ..... 31 43 N 72 59 E
Chishtīān Mandi .. 29 48 N 72 52 E
Chitrāl ..... 35 51 N 71 47 E
Dādu ..... 26 44 N 67 47 E
Darya Khān ..... 31 48 N 71 06 E
Daska ..... 32 20 N 74 21 E
Dera Ghāzi Khān .. 30 03 N 70 38 E
Dera Ismāil Khān .. 31 50 N 70 54 E
Dir ..... 35 12 N 71 53 E
Faisalābād (Lyallpur) ..... 31 25 N 73 05 E
Fort Sandeman, see Zhob
Ghotki ..... 28 01 N 69 19 E
Gilgit ..... 35 55 N 74 18 E
Gojra ..... 31 09 N 72 41 E
Gūjar Khān ..... 33 16 N 73 19 E
Gujrānwāla ..... 32 09 N 74 11 E
Gujrāt ..... 32 34 N 74 05 E
Gwādar ..... 25 07 N 62 19 E
Hab Chauki ..... 25 01 N 66 53 E
Hāfizābād ..... 32 04 N 73 41 E
Haripur ..... 33 59 N 72 56 E
Hasan Abdāl .... 33 49 N 72 41 E
Hāsilpur ..... 29 43 N 72 33 E
Hyderābād ..... 25 22 N 68 22 E
Islāmābād ..... 33 42 N 73 10 E
Jacobābād ..... 28 17 N 68 26 E
Jāmpur ..... 29 39 N 70 36 E
Jhang Sadar (Jhang-Maghiāna) ..... 31 16 N 72 19 E
Jhelum ..... 32 56 N 73 44 E
Kahror Pakka .... 29 37 N 71 55 E
Kalābāgh ..... 32 58 N 71 34 E
Kamālia ..... 30 44 N 72 39 E
Kambar ..... 27 36 N 68 00 E
Kāmoke ..... 31 58 N 74 13 E
Kandhkot ..... 28 14 N 69 11 E
Kardchi ..... 24 52 N 67 03 E
Kasūr ..... 31 07 N 74 27 E
Khairpur ..... 29 35 N 72 14 E
Khānewāl ..... 30 18 N 71 56 E
Khānpur ..... 27 00 N 67 42 E
Khushāb ..... 32 18 N 72 21 E
Khuzdar ..... 27 48 N 66 37 E
Kohāt ..... 33 35 N 71 26 E
Kot Addu ..... 30 28 N 70 58 E
Kotri ..... 25 22 N 68 18 E
Kundiān ..... 32 27 N 71 28 E
Lahore ..... 31 35 N 74 18 E
Lārkāna ..... 27 33 N 68 13 E
Leiah ..... 30 58 N 70 56 E
Loralai ..... 30 22 N 68 36 E
Lyallpur, see Faisalābād
Mailsi ..... 29 48 N 72 11 E
Malakand ..... 34 34 N 71 56 E
Mandi Bahāuddin .. 32 35 N 73 30 E
Mānsehra ..... 34 20 N 73 12 E
Mardān ..... 34 12 N 72 02 E
Mīānwāli ..... 32 35 N 71 33 E
Mingāora ..... 34 47 N 72 22 E
Mirpur Khās ..... 25 32 N 69 00 E
Mithankot ..... 28 57 N 70 22 E
Montgomery, see Sāhiwāl
Moro ..... 26 40 N 68 00 E
Multān ..... 30 11 N 71 29 E
Muzaffargarh .... 30 04 N 71 12 E
Nawābshāh ..... 26 15 N 68 25 E
Nowshera ..... 34 01 N 71 59 E
Nushki ..... 29 33 N 66 01 E
Okāra ..... 30 49 N 73 27 E
Ormāra ..... 25 12 N 64 38 E
Pākpattan ..... 30 21 N 73 24 E
Panjgūr ..... 26 58 N 64 06 E
Pasni ..... 25 16 N 63 28 E
Peshāwar ..... 34 01 N 71 33 E
Pīpri ..... 24 52 N 67 23 E
Pishīn ..... 30 35 N 67 00 E
Port Muhammad Bin Qāsim ..... 24 46 N 67 20 E
Quetta ..... 30 12 N 67 00 E
Rahīmyār Khān ... 28 25 N 70 18 E
Rāwalpindi ..... 33 36 N 73 04 E
Rohri ..... 27 41 N 68 54 E
Sādiqābād ..... 28 18 N 70 08 E
Sāhiwāl (Montgomery) ... 30 40 N 73 06 E
Samundri ..... 31 04 N 72 58 E
Sānghar ..... 26 02 N 68 57 E
Sāngla (Sangla Hill) ..... 31 43 N 73 23 E
Sargodha ..... 32 05 N 72 40 E
Shahqadar ..... 31 13 N 71 34 E
Shāhdādkot ..... 27 51 N 67 54 E
Shāhdādpur ..... 25 56 N 68 37 E
Shekhūpura ..... 31 42 N 73 59 E
Shikārpur ..... 27 57 N 68 38 E
Shujābād ..... 29 53 N 71 18 E
Sialkot ..... 32 30 N 74 31 E
Sibi ..... 29 33 N 67 53 E
Sīī ..... 28 37 N 69 19 E
Sukkur ..... 27 42 N 68 52 E
Tando Ādam ..... 25 46 N 68 40 E
Tando Allāhyār ... 25 28 N 68 43 E
Tando Muhammad Khān ..... 25 08 N 68 32 E
Tarbela Dam ..... 34 12 N 72 49 E
Thatta ..... 24 45 N 67 55 E
Toba Tek Singh ... 30 58 N 72 29 E
Turbat ..... 25 59 N 63 04 E
Umarkot ..... 25 22 N 69 44 E
Uthal ..... 25 48 N 66 37 E
Vihāri ..... 30 02 N 72 21 E
Wāh ..... 33 48 N 72 42 E
Wazirābād ..... 32 27 N 74 07 E
Zhob (Fort Sandeman) ..... 31 20 N 69 27 E

Physical features and points of interest

Arabian Sea ..... 24 00 N 64 00 E
Bajaur, *region* .... 34 50 N 71 55 E
Baluchistan, *region* ..... 28 00 N 63 00 E
Bolān Pass ..... 29 45 N 67 35 E
Bugti Hills ..... 28 50 N 69 00 E
Central Brāhui Range ..... 29 20 N 66 55 E
Central Makrān Range ..... 26 40 N 64 30 E
Chāgai Hills ..... 29 30 N 64 15 E
Chenāb, *river* .... 29 23 N 71 02 E
Cholistan Desert .. 28 15 N 70 45 E
Dardistān, *region* .. 35 30 N 74 00 E
Great Indian, see Thar Desert
Gumal Pass ..... 31 56 N 69 20 E
Gwātar Bay ..... 25 04 N 61 36 E
Harappā, *historical site* .... 30 38 N 72 52 E
Himalayas ..... 34 00 N 76 00 E
Hindu Kush Mountains ..... 35 00 N 71 00 E
Indo-Gangetic Plain ..... 30 00 N 76 00 E
Indus, *river* ..... 24 20 N 64 47 E
Indus Valley (Indus Plain) .... 29 00 N 71 00 E
Jhelum, *river* ..... 31 12 N 72 08 E
K2, *peak* ..... 35 53 N 76 30 E
Kābul, *river* ..... 33 55 N 72 14 E
Karakoram Range ..... 34 00 N 78 00 E
Khyber Pass ..... 34 05 N 71 10 E
Kirthar Range .... 27 00 N 67 10 E
Kohistān, *region* .. 35 03 N 72 52 E
Kutch, Rann of, *marsh* ..... 24 05 N 70 10 E
Makran, *region* .. 26 00 N 63 00 E
Manchhar Lake .... 26 25 N 67 39 E
Mangla Dam ..... 33 09 N 73 39 E
Mangla Reservoir .. 33 10 N 73 40 E
Mohenjo-daro, *historical site* .... 27 19 N 68 07 E
Nānga Parbat, *peak* ..... 35 15 N 74 36 E
Nāra Canal ..... 27 40 N 68 51 E
Pamirs, *mountains* ..... 37 30 N 73 00 E
Potwar Plateau .... 33 30 N 73 00 E
Punjab, *region* .... 30 00 N 74 00 E
Rās Koh Range .... 28 43 N 64 55 E
Rāvi, *river* ..... 30 35 N 71 49 E
Salt Range ..... 32 40 N 72 35 E
Sulaimān Range .. 30 30 N 70 10 E
Sutlej, *river* ..... 29 23 N 71 02 E
Swāt, *river* ..... 34 20 N 71 34 E
Tarbela Dam ..... 34 12 N 72 49 E
Taxila, *historical site* ..... 33 45 N 72 00 E
Thal Desert ..... 31 30 N 71 40 E
Thar (Great Indian) Desert ... 27 00 N 71 00 E
Tirāh, *region* ..... 33 50 N 70 45 E
Tirich Mir, Mount .. 36 15 N 71 50 E
Waziristān, *region* ..... 33 00 N 70 00 E

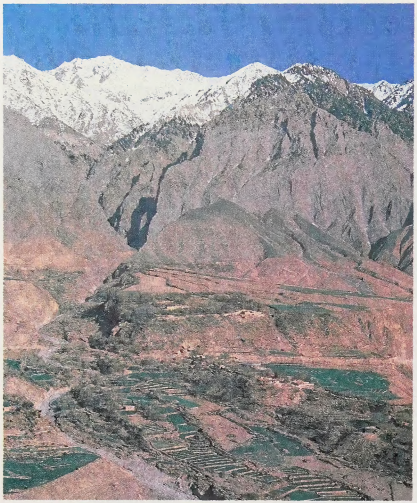
but extends for no less than 200 square miles when full; on such occasions, it is one of the largest freshwater lakes in South Asia. The quality of groundwater in the Indus Plain varies, that in the southern zone (Sindh) being mostly saline and unfit for agricultural use. Extensive areas in both the northern and southern zones of the Indus Plain have been affected by waterlogging and salinity. In the south the Indus delta (in marked contrast to the Ganges delta) is a wild waste. When high tides and Indus floods coincide, the littoral is flooded for 20 miles inland.

The Balochistān Plateau. The Balochistān Plateau extends westward, averaging more than 1,000 feet in elevation, with many ridges running across it from northeast to southwest. It is separated from the Indus Plain by the Sulaimān and Kirthar ranges. It has a remarkable indigenous method of irrigation called the karez, which consists of underground channels and galleries that collect subsoil water at the foot of hills and carry it to the fields and villages. The water is drawn from the channels through shafts that are sunk into the fields at suitable intervals. Because the channels are underground, the loss of water by evaporation is minimized. The technology for the construction of the karez probably came from western China, brought into Balochistān by the Buddhist monks who traveled back and forth over the Karakoram Range. The plateau is an extremely arid country and is the most sparsely populated region in Pakistan. Pastoral activity supplements a primitive form of agriculture. True pastoral nomadism survives in the northwest. Goats and fat-tailed sheep account for the majority of the stock, and much of the local traffic consists of camels and donkeys, although trucks and buses are in use on the new roads.

The western bordering mountains. These mountains run south from the Hindu Kush, in several parallel ranges, outside the path of the monsoons. Elevations in the Hindu Kush exceed 24,000 feet, but peaks are less lofty in the south. Rainfall is low, and there is little vegetation. Occasional heavy rain in the mountains results in destructive torrents, washing away the soil. Cultivation is practiced beside streams and along riverbanks or (where conditions permit) by terracing hillsides. Three minor ranges run south from the Hindu Kush to the Kābul River, south of which lies the Khyber Pass, at the frontier with Afghanistan. Farther south the Sulaimān Range, with peaks averaging 6,000 to 7,000 feet, runs southward for about 300 miles, after which the generally lower Kirthar Range, from 4,000 to 8,000 feet, runs down to the Arabian Sea coast. These hills separate the Indus Plain from Balochistān. The people of these arid hills are tribal. The area in general is very poor; only a minimum of crops are grown in favoured valleys, which are generally malarial. Large numbers of sheep and goats are kept. The mule is the main beast of burden in this broken country.

The Khyber Pass

The desert areas. The desert areas include the steppes of the Sindh Sāgar Doāb (at its centre is the Thal, which has true desert conditions) and Cholistan in the Bahāwalpur region (Punjab), which is known as the Nāra or Registān in the Khairpur region (Sindh) and as the Thar Desert in the Thar Parkar region of southeastern Sindh. All these areas are extensions of the Thar Desert of western India.



Kunar River valley in the Hindu Kush Mountains, North-West Frontier Province, Pakistan.
© Brian A. Vikander/West Light

Soils. Pakistan's soils are classified as pedocals, which comprise a dry soil group with a high content of calcium carbonate and a low content of organic matter; they are characteristic of a land with low and erratic rainfall. The soils of the Indus Plain consist of alluvium, old and new, whereas in the desert areas and most of Balochistān they consist either of windblown deposits known as loess or else of loess mixed with alluvium. The soils of the Thar Desert area are mostly sandy loams with patches of clay and larger areas of pure sand. Some of the virgin soils respond generously to irrigation, and an extensive reclamation scheme has made much progress in the Thal region.

Climate. As Pakistan is located on a great landmass north of the tropic of Cancer (between latitudes 24° and 37° N), it has a continental type of climate characterized by extreme variations of temperature, both seasonally and daily. Very high altitudes modify the climate in the cold, snow-covered northern mountains; temperatures on the Balochistān Plateau are somewhat higher. Along the coastal strip, the climate is modified by sea breezes. In the rest of the country, temperatures reach great heights in the summer; the mean temperature during June is 100° F (38° C) in the plains, where the highest temperatures can exceed 117° F (47° C). In the summer, hot winds called *loo* blow across the plains during the day. Trees shed their leaves to avoid loss of moisture. The dry, hot weather is broken occasionally by dust storms and thunderstorms that temporarily lower the temperature. Evenings are cool; the diurnal variation in temperature may be as much as 20° to 30° F (11° to 17° C). Winters are cold, with minimum mean temperatures of about 40° F (4° C) in January.

The characteristics of different regions are determined by variations in rainfall and irrigation rather than by temperature. Although the country is dominated by monsoon winds, it is extremely arid, except for the southern slopes of the Himalayas and the submontane tract, which have a rainfall from 30 to 35 inches (750 to 900 millimetres). The 20-inch (500-millimetre) precipitation line, which runs northwest from near Lahore, marks off the Potwar Plateau and a part of the Indus Plain in the northeast; these areas receive enough rainfall for dry cropping (farming without irrigation by taking measures to conserve water). South of this region, cultivation was confined mainly to riverine strips until the advent of irrigation.

Plant and animal life. Natural vegetation, except for wooded mountain slopes, is largely limited to tough, wiry grass and stunted bushes with only a few scattered trees, except for a few plantations in forest reservations and ubiquitous orchards of fruit trees. In Balochistān the Salt Range and the western-bordering-mountains vegetation is mostly limited to xerophytic plants (plants adapted for growth under dry conditions). The deserts are virtually devoid of vegetation. The Indus Plain is very fertile when watered. Wheat is the principal winter crop, the main summer crops being cotton and rice. Widespread installation of tube wells (also called driven wells, which are built by driving a casing into the ground), especially in the Punjab and increasingly in Sindh and the North-West Frontier Province, has led to substantial changes in cropping patterns, encouraging substitution of wheat for less valuable drought-resisting crops, such as gram (chick-pea).

Tube wells

Wildlife abounds in the northern mountains and includes brown bear, black Himalayan bear, leopard, the rare snow leopard, Siberian ibex, and varieties of wild sheep. Buffalo rather than cows are kept as dairy cattle, and bullocks are used for plowing, although occasionally camels are also used for that purpose. Donkeys are the traditional beasts of burden. The Manchhar Lake in Sindh has many water birds, including mallards, teals, shovelers, spoonbills, geese, pochards, and wood ducks; desert gazelles are also found in the region. Crocodiles, pythons, and wild boars inhabit the delta area. Jackals, foxes, wild cats, and a variety of rodents and reptiles are found throughout the country. There are no game reserves, and indiscriminate hunting is threatening the survival of several species.

Settlement patterns. *Traditional regions.* The traditional regions of Pakistan, shaped by ecological factors and historical evolution, are reflected in the administrative division of the country into the four provinces of Sindh, Punjab, the North-West Frontier Province, and Balochistān. The encroachment of the Cholistan Desert and of the western bordering mountains narrows down the Indus Plain below Mithankot, thus separating the destinies of Sindh and Punjab, each of which has a distinct language, culture, and history. Although the four provinces of Pakistan are separate, both linguistically and culturally, they were brought together from 1955 to 1970 under an administrative arrangement known as "One Unit" that established a single province of West Pakistan.

In the Punjab, until the advent of irrigation, most of the population was restricted to the *bārānī* ("land dependent on rainfall for cultivation") region—the Potwar Plateau and the upper Indus Plain—which receives more than 20 inches of rainfall. The Canal Colony areas of southern Punjab were formerly large tracts of uncultivated land in the Indus Plain that were irrigated by canals and populated by colonists drawn from every part of the Punjab. They now form the richest agricultural region of the country. Agricultural wealth is concentrated in the area around Lahore (in the *bārānī* region) that has benefited from irrigation, together with the Canal Colony areas and Sindh. These regions contain most of the rural population of Pakistan and produce more than half of its wheat and virtually all of its cotton and rice. Landholdings are larger in the Canal Colony areas of the Punjab and in Sindh.

Land-holding patterns

Elsewhere, in the overpopulated and poor districts of the *bārānī* region, which moreover do not benefit from irrigation, holdings are exceedingly small and fragmented. In these districts there is a great pressure to migrate from the villages to find employment in towns or in the armed forces. These districts were the source of most of the migrants (largely men) who went from Pakistan to the Middle East during the 1970s and '80s to take advantage of the economic boom resulting from the increase in the price of oil. Unlike the richer agricultural areas, this impoverished and overpopulated region maintains a close symbiotic relationship with urban economy and society. Educational levels in these districts are higher than they are in the richer agricultural areas.

Rural settlement. Two-thirds of the rural population of Pakistan lives in nucleated villages or hamlets (*i.e.,* in compact groups of dwellings). Sometimes, as is generally the case in the North-West Frontier Province, the houses are placed in a ring with blank outer walls, so that each complex resembles a protected fortress with a few guarded entrances. Dispersed habitation patterns in the form of isolated single homesteads are rare, occurring only in a few mountainous areas. But it is not uncommon to find numerous satellite hamlets of varying sizes near larger villages; such hamlets are occupied either by a landlord, with his servants and sharecroppers, or else by members of a lineage living together in adjoining houses. The spread of tube wells in the Punjab has increased the tendency for such dispersal, for people often prefer to live near their tube wells in order to guard the valuable machinery. The concept of village, therefore, often tends to be equivalent to that of the *mawzaʿ* (an area of land that, together with a village and its satellite hamlets, forms a unit in land-revenue records). It is difficult to speak of an average size of village, for patterns of habitation are complex. Most groups of dwellings have a minimum of a dozen or a score of houses, and there are usually a few hundred dwellings in each "village." Large villages rarely have a population exceeding 2,500 persons.

Three basic types of village layout are to be found. Most of the older villages are of the "spiderweb" form, having at least one focal point, such as the village mosque, some shops, or a well from which lanes radiate. A few villages follow the contours of hill slopes and other natural features. Finally, in the Canal Colony areas, villages are of a regular rectangular pattern, with a well, a mosque, and a school, as well as the house of the village headman, at the centre and with the houses being arranged in a series of concentric rectangles. Houses are built from available local materials; the vast majority are of mud, a material that is not only cheap and reasonably durable in the dry climate but also provides better insulation from extremes of heat and cold than brick or stone. Houses usually have walled courtyards where animals are tethered and where people sleep in the open in the hot summer.

Urban settlement. The urban population of Pakistan represents about a third of the total. Two cities have a dominating position—Karāchi and Lahore. Since the 1960s, government policy has been directed toward the dispersal of industry, which had become heavily concentrated in Karāchi. As a consequence, urban growth has been more evenly distributed among several cities. Karāchi, the principal port and centre of commerce and industry, was also the national capital until 1959, when the new capital at Islāmābād, 750 miles to the northeast, began to function.

Urban living conditions

Rapid and unplanned urban expansion has been paralleled by a deterioration in living conditions, particularly in the housing conditions of lower income groups. Many urban households are unable to pay rent for the cheapest form of available housing and live in makeshift shacks. Water supply and sewerage systems are inadequate, and in many areas residents have to share communal water taps. Inadequate urban transport is also a major problem.

Karāchi experienced serious ethnic conflict between the *muhajirs* (immigrants who moved to Pakistan from India following partition) and Sindhis and between the Sindhis and Punjabis in the latter part of the 1980s and the early 1990s. Discouraged by civil strife, businesses—both industrial and commercial—began to relocate in Punjab, particularly in and around the city of Lahore. After Karāchi and Lahore, the principal cities are Faisalābād (called Lyallpur until 1979), Rāwalpindi, and Peshāwar. Karāchi is the provincial capital of Sindh, Lahore of Punjab, Peshāwar of the North-West Frontier Province, and Quetta of Balochistān. The capital city of Islāmābād adjoins Rāwalpindi, where the army's general headquarters and the offices of the president of Pakistan are located.

THE PEOPLE

Ethnic composition. Race as such plays little part in defining regional or group identity in Pakistan, and no ideal racial type is accepted by all Pakistanis. The population is a complex mixture of indigenous peoples, many racial types having been introduced by successive waves of migrations from the northwest, as well as by internal migrations across the subcontinent of India. Aryans, Persians, Greeks, Pathans (Pashtuns), and Mughals came from the northwest and spread across the Indo-Gangetic Plain, while the Arabs conquered Sindh. All left their mark on the population and culture of the land. During the long period of Muslim rule, immigrants from the Middle East were brought in and installed as members of the ruling oligarchy. It became prestigious to claim descent from them, and many members of the landed gentry and of upper-class families are either actually or putatively descended from such immigrants. In 1947, when Pakistan and India became independent, there was another massive migration, of a different character, when millions of Muslim refugees were uprooted from different parts of India and settled in Pakistan; an equal number of Hindus were uprooted from Pakistan and driven across to India. This development further complicated the racial mixture of the population of the various regions of Pakistan.

By the early 1990s Pakistan's population was divided into five ethnic groups, defined broadly. The Punjabis constitute the majority, with more than 55 percent of the population; the Sindhis account for another 20 percent, the Pathans and the *mujahirs* for about 10 percent each, and the Balochs for about 5 percent. There are subgroups within each of these five categories. The Arains, Rajputs, and Jats—all Punjabis—regard themselves as ethnically distinct. Some groups overlap the five categories: for instance, there are Punjabi Pathans as well as Hazarvi Pathans. Some smaller groups, such as the Brohis in Sindh and the Seraikis in Punjab, are also ethnically distinct.

Linguistic composition. Pakistan is in general linguistically heterogeneous, and no single language can be said to be common to the whole population. Each of its principal languages has a strong regional focus, although statistics show some languages to be distributed between various provinces because administrative boundaries cut across linguistic regions. The picture is also complicated by the fact that, especially in Sindh, there are substantial numbers of Urdu- and Punjabi-speaking immigrants (*muhajirs*) who moved from India following partition. The languages claimed as mother tongue include Urdu, Punjabi, Sindhi, Pashto, Balochi, and Brahui.

Urdu

Urdu is the youngest of the nation's languages and is not indigenous to Pakistan; it is very similar to Hindi, an official language of India. Although the two languages have a common base, in its literary form Urdu emphasizes words of Persian and Arabic origin, whereas Hindi emphasizes words of Sanskrit origin. Urdu is written in a modified version of the Persian script (written from right to left), whereas Hindi is written in Devānagari script from left to right. Because it is preeminently the language of the educated Muslims of northern India, including the Punjab, Urdu has strong associations with Muslim nationalism; hence the ideological significance of Urdu in Pakistani politics.

The 1956 constitution prescribed the use of English for official purposes for 20 years, and the 1962 constitution made the period indefinite. The 1973 constitution, however, designated a 15-year transition period to Urdu, after which English would no longer be used for official purposes. English is spoken by only a small percentage of the people. Urdu is the mother tongue of only a small percentage of the population of Pakistan; it is taught in the schools along with the regional languages.

Punjabi has its own script, Gurmukhi, but it is mainly used in India. In Pakistan, Punjabi is mainly spoken rather than written; it is also a predominantly rural rather than an urban language. Urdu, rather than Punjabi, is the first language taught in schools in Punjab, so that every educated Punjabi reads and writes Urdu. There was a movement for the promotion of the Punjabi language in the 1980s and '90s, and some Punjabi literature is being published using the Urdu script; among the works published are Punjabi classics that have hitherto been available in Gurmukhi script or preserved in oral tradition.

Sindhi is derived from the Virachada dialect of Prākrit; it has fewer dialects than Punjabi. It is written in a spe-

cial variant of the Arabic script. Most of the educated middle class in Sindh were Hindu, and their departure to India in 1947 had a traumatic effect on Sindhi culture. Vigorous efforts were therefore directed toward a recovery and preservation of the rich Sindhi literary and cultural heritage. Large numbers of Urdu-speaking refugees settled in Sindh, and they now constitute the majority of the population of its larger towns. As a consequence, the movement for the promotion of Sindhi language and culture was sometimes expressed as opposition toward Urdu. The Sindhi population feared that their language and culture would be overrun by the language and culture of the *muhajir* community; this fear led to the language riots of 1972 and to the government's decision to grant special status to the Sindhi language. The rise of militant ethnic politics in the 1980s can be traced to this decision.

Pashto, the language of the Pathans (Pashtuns, or Pakhtuns) of the North-West Frontier Province, has no written literary traditions although it has a rich oral tradition. There are two major dialect patterns within which the various individual dialects may be classified; these are Pakhto, which is the northern (Peshāwar) variety, and the softer Pashto spoken in southern areas. As in the Punjab, Urdu is the language taught in schools, and educated Pathans read and write Urdu. Again, as in the case of Punjabi, there was a movement for developing the written language in Persian script and increasing the usage of Pashto. From 1979 onward, as many as 3.5 million largely Pashto-speaking refugees arrived in Pakistan from Afghanistan. While most were accommodated in refugee camps in and around Peshāwar, a number of them settled in Karāchi, further complicating the city's ethnic and linguistic picture.

The two main spoken languages of Balochistān are Balochi and Brahui. Makrani is an important dialect of Balochi; it is spoken in Makrān, the southern region of Balochistān, bordering on Iran.

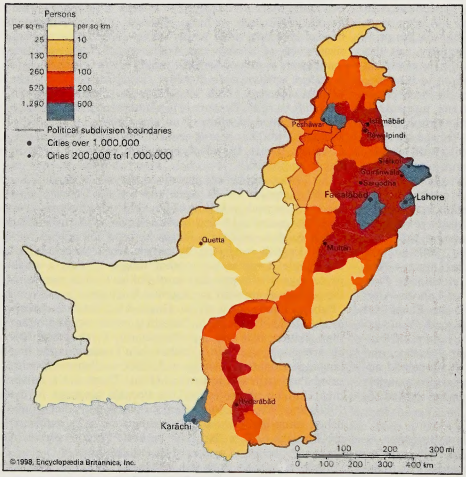
Religions. Almost all the people of Pakistan are Muslims. Most of them belong to the Sunnite sect, the major branch of Islām, with a significant representation among the Shī'ite branch. There is also a very small, though influential, sect called Aḥmadīyah, or Qadianis, which does not regard the prophet Muḥammad as the final prophet, a basic tenet of most Muslim sects. A 1974 constitutional amendment declared the Aḥmadīyah community to be non-Muslims. Since then the community has experienced considerable persecution, particularly during the administration of General Mohammad Zia ul-Haq (1977–88). The majority of Pakistani Sunnites belong to the orthodox Ḥanafī school, which is one of four schools or subsects of Sunnism. Shī'ites are also divided into numerous subsects; among them are Ismā'īlīs (the followers of the Aga Khan), as well as the Twelvers (Ithnā 'Asharīs) and Bohrās, which are prominent communities in commerce and industry. The principal business communities among Sunnites are Gujarati Memons and Chiniotis from Punjab.

Muslim sects

With the exception of some sects, such as Dawoodi Bohrās, there is generally no ordained priesthood among Pakistan's Muslims. Anyone who leads prayers in mosques may be appointed imam. Those who are trained in theology are given the title of mullah or *mawlānā* or, collectively, ulama. There are powerful hereditary networks of "holy men" called pirs, who receive great reverence, as well as gifts in cash or kind, from a multitude of followers. An established pir may pass on his spiritual powers and sanctified authority to one or more of his *murīds* ("disciples"), who may then operate as pirs in their own right. There are also many self-appointed pirs who practice locally without being properly inducted into one of the four Ṣūfī orders. Pirs who occupy high positions in the pir hierarchy wield great power and play an influential role in public affairs.

The number of Hindus in Pakistan was greatly reduced as a consequence of the exodus of refugees to India in 1947. Both Hindus and Christians constitute only a tiny percentage of the population.

Demographic trends. The rate of population growth is extremely high; it increased from less than 2 percent in the decade following independence to slightly more than 3 percent in the 1980s. In 1986 Pakistan's population crossed the 100 million mark. The fertility rate, estimated at about 6.5 children per woman, has not declined, in spite of government-sponsored family planning programs. Internal migration has had a significant impact on Pakistan's demography. The 1947 influx of Muslim refugees from India brought eight million newcomers to what was then West Pakistan. When Pakistan took its first population



Population density of Pakistan.

census in 1951, one out of every four of West Pakistan's residents was a migrant. Minor migrational shifts occurred between India and West Pakistan following the 1965 border hostilities. After the founding of Bangladesh, thousands of non-Bengali Muslims emigrated from Bangladesh to Pakistan and an almost equal number of Bengalis moved to Bangladesh from Pakistan. A second major wave of Muslim immigration began when thousands of refugees began to flee Afghanistan, occupied by the Soviet Union in 1979, to live in camps and settlements along the Pakistani border. By the close of the 1980s, as many as 3.5 million Afghan refugees had arrived in the country. Emigration also has had demographic, as well as economic and social, effects. Millions of single male Pakistanis went to work in the Middle East in the 1970s and '80s, creating a pool of workers that, at its peak, numbered three million.

Afghan refugees

Regional distribution of the population is uneven. The population is dense in the fertile Indus valley, especially in the extreme northeast, around Lahore. By contrast, in the vast region of Balochistān, the population densities are relatively low.

A trend toward urbanization has been accompanied by a faster rate of growth of the larger cities as compared to the smaller ones; this reflects the influence of the location of industry in the pattern of urban growth. By the year 2000 the population of Karāchi, Pakistan's largest city, is expected to exceed 10 million, and Lahore, with its satellite cities, is also likely to have roughly the same number of people.

THE ECONOMY

Pakistan operates a mixed economy in which the state-owned enterprises (including industrial corporations, trading houses, banks, insurance companies, institutions of higher learning, medical schools and hospitals, and transport companies) account for nearly half of the gross national product (GNP). In addition, the state, with the help of an intricate system of industrial licensing and trade regulations, controls new private investments. The state also has at its disposal labour, health, and tax laws to oversee the functioning of the private sector. The balance between the public and private sectors of the economy was altered in favour of the former in 1972–74 as a result of a series of nationalization measures. Until then, and unlike most other developing countries, Pakistan had regarded the private sector as the leading sector of the economy.

The economy, which was primarily agricultural at the time of independence, is now considerably diversified. Agriculture, although still the largest sector, now contributes less than one-fourth of the GNP, while manufacturing provides almost one-fifth. In terms of the structure of its economy, Pakistan resembles the middle-income countries of East and Southeast Asia more than the poor nations of the Indian subcontinent. Economic performance compares favourably with that of many other developing countries; the GNP has increased at an average rate of more than 5 percent a year since independence. At the same time, there has been a relentless increase in population, so that, despite a real growth in the economy, output per capita has risen slowly. By 1990 Pakistan's economy was four times as large as it was at the time of independence in 1947, its population was three and a half times as large, and its per capita income was twice as large. In general, although the GNP per capita is relatively low, Pakistan does not have a high incidence of absolute poverty (the level below which a minimally adequate diet and other essential requirements are not affordable); the proportion of the population living in absolute poverty is considerably smaller than in other South Asian countries. The relative prosperity of the industrialized regions around Karāchi and Lahore contrasts sharply with the poverty of the Punjab's *bārānī* areas, the semiarid Balochistān, and the North-West Frontier Province.

One of the paradoxes of Pakistan's economic situation is that, in spite of a healthy increase in its GNP and in spite of its success in alleviating the worst forms of poverty, it has continued to experience a very low level of social development. The social status of Pakistani women is particularly low. The country has a high rate of infant mortality, losing before they reach one year of age more than 100 children out of every 1,000 born; its maternal mortality rate, at 6 per 1,000 live births, is among the highest in the world; the rate of literacy, with only one of every seven women able to read and write, is very low compared to that of other developing countries.

Social development

A system of medium-term planning was introduced in 1958 with the belated publication of the first five-year plan (1955–60). In the following decades a series of five-year plans were formulated, but these met with varying degrees of success.

During the 1980s a movement toward an "Islāmic economy" was announced by the Pakistani government. This movement involved the purging of economic practices outlawed by Muslim theology, such as *riba* (interest), and the mandatory reinstatement of the *zakat* (an annual tax on several types of personal financial assets that is used to provide aid for the poor) and the *ushr* (the *zakat* on land), which had not been universally adhered to but remained central tenets of Islāmic law. General Zia had promised further Islāmization of the economy, but he died before these steps could be taken. Under Benazir Bhutto, his successor, the Islāmization movement slowed, although the government was obliged to keep on the books most of the legislation enacted during the Zia period.

Taxation accounts for more than three-quarters of government revenue, and government expenditures exceed revenues by a large amount. Income tax rates have been comparatively high, but the tax base has been so small that individual and corporate income tax revenues have remained substantially less than excise, sales, and other indirect taxes. The government has been able to maintain heavy expenditure on development and defense because of the inflow of foreign aid and the remittances sent by Pakistanis working abroad. In the 1970s and '80s external capital inflows were equivalent to as much as one-tenth of the GNP and financed well over half of the total domestic investment. In allowing this dependence on foreign capital to persist, however, the country has accumulated an enormous foreign debt, the financing of which has been a major problem.

The trade union movement dates to the late 19th century, but, because Pakistan's industrial sector (inherited at independence) was so small, organized labour as a proportion of total employment is still in a minority. This has not prevented it from becoming an important political force. Before the 1971 civil war, there were well over 1,000 registered unions, most of them organized within individual establishments. Countrywide unions based on a common craft or industry were very few. Most of the unions were situated in the urban centres and were affiliated to one of three national labour confederations. After the civil war and the emergence of Bangladesh, the number of unions declined to a few hundred, affiliated to one umbrella organization, the Pakistan National Federation of Trade Unions. Because of the high rates of unemployment, employers remained in a strong position, and many of them were able to bypass working agreements and laws. Only the unions in the bigger industries (*e.g.*, cotton textiles) had the necessary coherence to fight back. Labour laws introduced in 1972 met some of the demands (job security, social welfare, pensions) of organized labour but also sought to control political activity by industrial workers. Labour union activity was severely constrained by the military government of 1977–88 but was revived by the administration of Benazir Bhutto.

Trade unions

Resources. *Mineral resources.* The exploration of Pakistan's mineral wealth is far from complete, but more than 20 different types of minerals have been located. Coal mining is one of the country's oldest industries. The quality of the coal is poor, and the mines work below capacity because of the lack of demand. Iron ore deposits are also mostly of poor quality. The most extensive known reserves are situated in the Kālābāgh region in western Punjab. Other low-grade ore reserves have been found in Hazāra in the North-West Frontier Province. Small reserves of high-grade iron ore have been identified in Chitrāl and in the Chilghāzi area (located in northwestern Balochistān), also in the North-West Frontier Province. Deposits of copper

ore, equaling or surpassing the reserves of iron ore, have been located, but most sites remain unexploited. There are enormous reserves of easily exploited limestone that form the basis of a growing cement industry, the largest component of the manufacturing sector. Other minerals that are exploited include chromite (mostly for export), barite (a white, yellow, or colourless mineral resembling marble), celestite (strontium sulfate), antimony, aragonite (a mineral resembling calcite [calcium carbonate]), gypsum, rock salt, and marble. Radioactive minerals have been found in southwestern Punjab.

Pakistan also has small quantities of oil and some very large natural gas fields. The first oil discovery was made in 1915. Pakistan intensified the search for oil and natural gas in the 1980s and was rewarded with the discovery of a number of new oil fields in the Potwar Plateau region and in Sindh. The oil fields near Badīn, in Sindh, are particularly promising. Oil fulfills a substantial portion of Pakistan's energy requirements, and the search for new and richer fields has continued. The largest natural gas deposits are at Sūi (on the border between Balochistān and the Punjab), discovered in 1953. A smaller field, at Māri, in the northwest of Sindh province, was found in 1957. A number of smaller natural gas fields were discovered in the 1980s. A network of gas pipelines links the fields with the main consumption areas: Karāchi, Lahore, Multān, Faisalābād (Lyallpur), and Islāmābād.

Biological resources. The variety of climates and soils has given rise to a wide diversity in biological resources. As Balochistān is mostly desert, only in the small areas of intensive cultivation do crops and orchards thrive. In Sindh and Punjab, where the annual rainfall is also low, most of the vegetation is basically xerophilous, except for the riverine forests along the Indus and its tributaries. Parts of the coastal region have mangrove forests. Regular rain and snow in the Himalayan foothills of the north have given them a variety of vegetation and animal life ranging from the Mediterranean to the Alpine types. There is a fishing industry centred in Karāchi, and part of the lobster and other shellfish catch is exported.

Hydroelectric and other power resources. Although energy production has grown faster than the economy as a whole, it has not kept pace with demand, and as a result there are shortages of fuel and electric power. Great progress, however, has been made in the development of the hydroelectric potential of Pakistan's rivers. A giant hydroelectric plant is in operation at the Mangla Dam on the Jhelum River in Azad Kashmīr (the part of the disputed states of Jammu and Kashmīr under Pakistan's control). Another such source is the giant Tarbela Dam on the Indus River.

The generation, transmission, and distribution of power is the responsibility of the Pakistani Water and Power Development Authority (WAPDA), a public-sector corporation. WAPDA lost its monopoly over generation after Pakistan signed an agreement in 1989 with a consortium of foreign firms to produce power from giant oil-fired plants to be located at Hub, near Karāchi. The majority of Pakistan's energy requirements are now fulfilled by oil, mostly imported, and by natural gas. Nuclear power provides a very small percentage of the total energy used; it is supplied primarily to Karāchi. In 1989 Pakistan concluded agreements with China and France to set up additional nuclear power-generating plants. These reactors are to be located inland on the Indus River to serve the rapidly increasing demand of Punjab.

WAPDA

Agriculture, forestry, and fishing. Agriculture, forestry, and fishing provide employment for at least one-half of the official labour force and a livelihood for an even larger proportion of the population. Land-reform programs implemented in 1959, 1972, and 1977 began to deal with the problems of large-scale, often absentee ownership of land and the excessive fragmentation of small holdings by introducing maximum and minimum area limits. The commercialization of agriculture has also resulted in fairly large-scale transfers of land, concentrating its ownership among middle-class farmers.

The attention given to the agricultural sector in development plans has brought about some radical changes in centuries-old farming techniques. The construction of tube wells for irrigation and salinity control, the use of chemical fertilizers and scientifically selected seeds, and the gradual introduction of farm machinery have all contributed to the notable increase in productivity. Early on, one of the prime objectives of agricultural development programs was self-sufficiency in wheat, which Pakistan achieved in the early 1970s.

Pakistan experienced a "green revolution" during the late 1960s. In this period wheat production increased dramatically, leaving a surplus over domestic consumption that was partly shipped to East Pakistan (Bangladesh) and partly exported. Cotton production also rose. Yields remain low by international standards, but increasing amounts are processed locally; in addition, much of Pakistan's edible oil is produced from cottonseed. Rice is the second major food staple and one of the country's important export crops. By the end of the 1980s, Pakistan had become the third largest exporter of rice in the world, after the United States and Thailand. Large domestic sugar subsidies have been the main cause for the increase in sugarcane production.

Animal husbandry provides important domestic and export products. Apart from the supply of meat and dairy products for local consumption, it includes production of wool for the carpet industry and for export and of hides and skins for the leather industry. The contribution of forestry to national income remains negligible, but that of fisheries has risen.

Industry. Mining and quarrying account for a small percentage of gross domestic product and of total employment. Manufacturing, however, accounts for a healthy proportion. The beginning of the main industrialization effort dates to the cessation of trade between India and Pakistan in 1949 soon after the two countries gained independence. Initially it was based on the processing of domestic agricultural raw materials for the home market and for export. This led to the setting up of cotton textile mills—a development that now accounts for a large part of the total employment in industry. Woolen textiles, sugar, paper, tobacco, and leather industries also provide many jobs for the industrial labour force.

Industrialization

The growing trade deficit in the mid-1950s compelled the government to cut down on imports, which encouraged the establishment of a number of import-substitution industries. At first these industries produced mainly consumer goods, but gradually they came to produce intermediate goods and a range of capital goods, including chemicals, fertilizers, and light engineering products. Nevertheless, Pakistan still has to import a large proportion of the capital equipment and raw materials required by industry. In the 1970s and early 1980s Pakistan set up an integrated iron and steel mill at Pipri, near Karāchi, with the financial and technical assistance of the Soviet Union. A new port, Port Qāsim (officially Port Muhammad Bin Qāsim), was built to bring iron ore and coal for the mill.

Initially Karāchi was the centre of Pakistan's industrialization effort, but in the late 1960s and early 1970s Lahore and the cities around it began to industrialize rapidly. Karāchi's ethnic problems in the late 1980s and early 1990s accelerated this process, and Punjab increasingly became Karāchi's competitor in industrial output.

Finance. Pakistan has a fairly well-developed system of financial services. The State Bank of Pakistan has overall control over the banking sector, which consists of a number of commercial banks and specialist credit institutions. The State Bank acts as banker to the central and provincial governments and administers official monetary and credit policies, including exchange controls. It has sole currency-issuing rights and has custody of the country's gold and foreign-exchange reserves. Pakistan has a number of commercial banks, called scheduled banks, which are subject to strict State Bank requirements as to paid-up capital and reserves. They account for the bulk of total deposits, collected through a network of branch offices. A few specialist financial institutions provide medium- and long-term credit for industrial, agricultural, and housing purposes and include the Pakistan Industrial Credit and Investment Corporation, the Industrial Development

Bank of Pakistan, the Agricultural Development Bank of Pakistan, and the House Building Finance Corporation. A new financial institution, the National Development Finance Corporation, was set up in the 1970s to finance industries managed by the state. The Karāchi and Lahore stock exchanges deal in stocks and shares of registered companies. The Investment Corporation of Pakistan and the National Investment Trust have helped to channel domestic savings into the capital market. As part of the return to an "Islāmic economy," interest-free banking and financing practices have been instituted.

Trade. Although there has been a trend toward increasing exports, overall trade has remained heavily in deficit. Over the years, important changes have taken place in the composition of foreign trade. In particular, while the proportion of total exports from primary commodities, including raw cotton, has fallen, the share of manufactures has greatly increased. But the bulk of the manufactured products coming into the export trade consists of cotton goods, so that Pakistan remains as dependent as ever on its leading cash crop. The other manufactures exported come mostly from industries based on agriculture, such as leather and leather goods and carpets. The shift toward manufactured agricultural exports, which have a higher added-value content than primary commodities, has been encouraged by the government. The trade deficits and shortages of foreign exchange have made it necessary for the government to restrict imports and to provide financial incentives to promote export trade.

Growth of manufactures

Transportation. The dominant role of rail as the principal long-distance carrier has been displaced by the bus and truck. A program of deregulation of the road transport industry was undertaken in 1970; it encouraged the entry of a large number of independent operators into the sector. Motor trucks and tractor-drawn trailers are also displacing the traditional bullock cart for local transport of produce to markets.

The main arterial road, which runs from Karāchi to Peshāwar via Lahore and Rāwalpindi, is 1,080 miles (1,740 kilometres) long. The main rail route runs more than 1,000 miles north from Karāchi to Peshāwar, via Lahore and Rāwalpindi. Another main line branches northwestward from Sukkur to Quetta.

In the early 1990s the limitations of the transportation system emerged as a major constraint on the modernization of the economy, prompting the government to undertake large-scale investments in the highway sector. Private entrepreneurs were also invited to participate on the basis of a "build-operate-transfer" (BOT) approach, which has become popular in developing countries. (In the BOT system, private entrepreneurs build and operate infrastructure facilities such as ports, highways, and power plants and then recover their costs by charging tariffs from the users. Once the investors have recovered their outlay, the facility created is transferred to the government.)

Pakistan International Airlines (PIA), established in 1954, is the sole carrier of internal air traffic. PIA also runs international flights to Europe, the Middle East, Africa, and East Asia, as well as to neighbouring Afghanistan. The principal airports are at Karāchi, Lahore, Rāwalpindi, Quetta, and Peshāwar. Karāchi and Port Qāsim are the principal ports; in addition, a number of small harbours along the Makran Coast handle the small boats that ply between Pakistan and the Persian Gulf states.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. The political system of Pakistan has undergone several far-reaching changes since independence. In 1971 its turbulent politics culminated in the secession of its eastern region (having more than 54 percent of the total population at that time), which established itself as the independent state of Bangladesh. In the aftermath of that event, Pakistan (now reduced to what was previously West Pakistan) faced a number of political and economic problems and uncertainties about its future.

Several conflicts have left their mark on the politics of Pakistan. The first of these, initially obscured by the paraphernalia of a parliamentary form of government but later made manifest by overt seizure of power by men at the head of the military and bureaucratic establishment, was a continuing struggle between political leadership and a military-bureaucratic oligarchy for supremacy and authority in the state; ideologically, this struggle was expressed as a struggle for democracy. The military-bureaucratic oligarchy triumphed for a while and set up three military administrations, in 1958–69, 1969–71, and 1977–88.

Political stresses

A second and quite distinct conflict was a struggle between regional groups. Because it was directed against centralized authority, it merged with the democratic struggle. But its express aims were to secure greater regional representation in the bureaucratic and military establishment, especially in the higher echelons, and to achieve effective decentralization of powers within a federal governmental structure by emphasizing regional autonomy. This struggle manifested itself first in the civil war between East and West Pakistān in 1971–72. It also brought to a virtual standstill the working of the first civilian government after the 1977–88 military administration.

A third conflict concerned the allocation of economic resources and burdens and the distribution of a greater share of the benefits of development among the more deprived regions and strata of the population. This resulted in a number of violent confrontations between the less-privileged segments of society and the state. Some of these confrontations, such as those in 1969 and 1977, led to the fall of constitutional governments and the imposition of martial law.

A fourth conflict was between the landed interests that had dominated Pakistan's political and economic life for much of the country's history and the new urban interests that began to assert themselves in the late 1980s and the 1990s. One manifestation of this was the struggle between Punjab and the federal government in the late 1980s. Under the Islāmic Democratic Alliance, the Punjab government continued to represent the interests of the landed aristocracy, while the government of Benazir Bhutto, with a more liberal bent and a wider support base, espoused the economic and social interests of urban groups and nonpropertied classes. The two governments often clashed in the late 1980s, creating serious economic management problems.

The constitutional framework. The task of framing a constitution was entrusted in 1947 to a Constituent Assembly that was also to function as the interim legislature under the 1935 Government of India Act, which was to be the interim constitution. Pakistan's first constitution was enacted by the Constituent Assembly in 1956. It followed the form of the 1935 act, allowing the president far-reaching powers to suspend federal and provincial parliamentary government. It also included the "parity formula," by which representation in the National Assembly for East and West Pakistan would be decided on a parity, rather than population, basis. (A major factor in the political crisis of 1970–71 was abandonment of the "parity formula" and adoption of representation by population, giving East Pakistan an absolute majority in the National Assembly.)

In 1958 the constitution was abrogated and martial law was instituted. A new constitution, promulgated in 1962, provided for the election of the president and national and provincial assemblies by an electoral college composed of members of local councils. Although a federal form of government was retained, the assemblies had little power, which was, in effect, centralized through the authority of governors acting under the president. In April 1973 Pakistan's third constitution was adopted by the National Assembly; it was suspended in 1977. In March 1981 a Provisional Constitutional Order was promulgated, providing a framework for government under martial law. Four years later a process was initiated for reinstating the constitution of 1973. By October 1985 a newly elected National Assembly had passed an amended constitution, giving extraordinary powers to the president, including the authority to appoint any member of the National Assembly as prime minister. With the end of military rule in 1988 and following elections to the National Assembly held in November of that year, the new president used these powers to appoint a prime minister to form a civilian government under the amended 1973 constitution.

Structure of government

The amended constitution provides for a president as head of state and a prime minister as head of government; both must be Muslims. The president is elected for a term of five years by the National Assembly, the Senate, and the four provincial assemblies. The prime minister is elected by the National Assembly. The president acts on the advice of the prime minister.

The National Assembly has 237 seats. Of these, 217 are filled by direct popular election; 207 are for Muslim candidates and 10 for non-Muslims. The remaining 20 seats are reserved for women who are chosen by the elected members. Members of the National Assembly serve five-year terms. The Senate has 87 members who are chosen by the provincial assemblies for six-year terms. One-third of the senators relinquish their seats every two years.

Political parties. The role of Islām in the political and cultural unification of Pakistan has been controversial. Some factions have argued that Islāmic ideology is the only cement that can bind together its culturally diverse peoples. Opposing factions have argued that the insistence on Islāmic ideology, in opposition to regional demands expressed in secular and cultural idiom, has alienated regional groups and eroded national unity.

The Pakistan People's Party (PPP) was formed in 1968 by Zulfikar Ali Bhutto, working with a number of socialists who wanted Pakistan to disregard the idiom of religion in politics in favour of a program of rapid modernization of the country and the introduction of a socialist economy. The PPP emerged as the majority party in West Pakistan in the elections of 1970 and was invited to form a government after the collapse of the second military administration in 1972. The PPP was suppressed under the military government of 1977–88 but returned to power in 1988.

In 1962 the Muslim League, which had spearheaded the Pakistan movement under Mohammed Ali Jinnah, splintered into two parts, the Pakistan Muslim League and the Council Muslim League. In the elections of 1970 it almost disappeared as a political party, but it was resurrected in 1985 and became the most important component of the Islāmic Democratic Alliance, which took over Punjab's administration in 1988. The Islāmic Assembly, founded in 1941, commands a great deal of support among the urban lower-middle classes. Two other religious parties, the Assembly of Islāmic Clergy and the Assembly of Pakistani Clergy, have strong centres of support, the former in Karāchi and the latter in the rural areas of the North-West Frontier Province. Ethnic interests are served by organizations such as the Muhajir National Movement in Karāchi and Hyderābād, the Sindhi National Front in Sindh, and the Balochistān Students Union in Balochistān.

Local government and administration. Pakistan's four provinces are divided into divisions, districts, and tahsils (subdistricts), which are run by a hierarchy of administrators, such as the divisional commissioner, the deputy commissioner (in the district), and the subdivisional magistrate, subdivisional officer, or tahsildar, at the tahsil level. The key level is that of the district, where the deputy commissioner, although in charge of all branches of government, shares power with the elected chairman of the district council. In addition to the provinces, Pakistan has 11 federally administered tribal areas, which are overseen by agents responsible to the federal government; the Islāmābād Capital Territory; and a number of tribal areas that are administered by the provincial governments.

Justice. There is a formal division between the judiciary and the executive branches of government. The judiciary consists of the Supreme Court, the provincial high courts, and (under their jurisdiction and supervision) district courts that hear civil cases and sessions courts that hear criminal cases. There is also a magistracy that deals with cases brought by the police. The district magistrate (who, as deputy commissioner, also controls the police) hears appeals from magistrates under him; appeals may go from him to the sessions judge. The Supreme Court is a court of record. It has original, appellate, and advisory jurisdictions and is the highest court in the land. At the time of independence, Pakistan inherited legal codes and acts that have remained in force, subject to amendment.

The judicial system also began a reorientation to Islāmic tenets and values designed to make legal redress inexpensive and accessible to all persons. A complete code of Islāmic laws was instituted, and the Federal Shariat Court, a court of Islāmic law (Sharī'ah), was set up in the 1980s.

Islāmization of the legal system

Education. Pakistan's literacy rate is substantially lower than that of many developing nations; only about a fourth of all adults are literate. A significant percentage of those who are literate, however, have not had any formal education. Educational levels for women are much lower than those for men. The share of females in educational levels progressively diminishes above the primary school level.

Education in Pakistan is not compulsory. Since independence Pakistan has increased the number of primary and secondary schools, and the number of students enrolled has risen dramatically. Teacher training has been promoted by the government and by international agencies. Higher education is available at vocational schools, technical schools, and colleges throughout the country. The oldest university is the University of the Punjab, established in 1882, and the largest universities are Allama Iqbal Open University in Islāmābād, the University of Peshāwar, and the University of Karāchi. Universities established during the 20th century include the University of Azad Jammu and Kashmir (1980), the North-West Frontier Province Agricultural University in Peshāwar (1981), the International Islāmic University in Islāmābād (1980), the Aga Khan University in Karāchi (1983), and the Lahore University for Management Sciences (1986). Most university classes are taught in Urdu or English.

Education suffered a major setback in the 1970s as a result of the nationalization of private schools and colleges. The reversal of this policy in the 1980s led to a proliferation of private institutions, particularly in the large cities. In the 1980s the government also began to focus on the Islāmization of the curriculum and the increased use of Urdu as the medium of instruction. The more Westernized segments of the population prefer to send their children to private schools, which continue to offer Western-style education and instruction in English. A number of private schools offer college entrance examinations administered by educational agencies in the United States and the United Kingdom, and many graduates of these schools are educated abroad. The division of the educational system into a private, Westernized section and a state-run Islāmized section has thus caused social tensions and exacerbated the problem of "brain drain," the emigration to the West of many of the better-educated members of the population.

Health and welfare. Although Pakistan has made progress in improving health conditions, a large part of the population does not receive modern medical care. There are insufficient numbers of doctors and nurses, especially in rural areas. Sanitation facilities are also inadequate; only a small percentage of the population has access to safe drinking water and sanitary sewage disposal facilities. Malaria, tuberculosis and other respiratory diseases, and intestinal diseases are among the leading causes of death; drug addiction is an increasingly serious problem.

Pakistan was among the first developing countries to establish a state-funded family planning program. This effort began in the early 1960s when the Family Planning Organization was organized. The program ran into political difficulties in the late 1960s as a result of opposition by fundamentalist Islāmic groups. The regimes of Zulfikar Ali Bhutto, Zia ul-Haq, and Benazir Bhutto gave family planning a relatively low priority. Consequently, Pakistan's total fertility and population growth rates are among the highest in the world.

The *zakat* and *ushr* taxes are used to provide social welfare funds, which go to provincial, division, and district committees for distribution among organizations engaged in social welfare activities or directly to needy persons. *Zakat* funds are also used for scholarships.

The *zakat* and *ushr*

CULTURAL LIFE

Pakistan shares influences that have shaped the cultures of South Asia. There are thus wider regional similarities extending beyond the national boundaries. On the other hand, the specific regional cultures of Pakistan present a

picture of rich diversity. It is difficult, therefore, to speak of a single Pakistani culture.

Family organization is strongly patriarchal, as in most agrarian societies, and most people live in large extended families. A woman's place in society is low, and she is restricted to the performance of domestic chores and to fulfilling the role of a dutiful wife and mother. In wealthy peasant and landowner households and in urban middle-class families, women are kept in seclusion (*pardah*); on the rare occasions on which they set foot outside their houses, they must be veiled. Among poor peasants, women have duties on the farm as well as in the house and do not observe *pardah*. In the Punjab, cotton picking is exclusively a woman's job, and women keep the money thus earned for their own purposes. Houses of those who practice *pardah* have a men's section (*mardānah*) at the front of the house, so that visitors do not disturb the women, who are secluded in the women's section (*zanānah*).

Among the very rich, Western education and modes of living have eliminated *pardah*, but, in general, even among this group, attitudes toward women in society and the family are akin to those of Victorian England. Change is coming most rapidly among the urban middle-income group, inspired by increasing access to the West as well as by the entry of women into the work force. An increasing number of women do not observe *pardah*, and the education of women has been encouraged. Some women have gained distinction in the professions; some of Pakistan's leading politicians, journalists, and teachers are women.

Social organization revolves around kinship rather than caste. *Berādari* (patrilineage) is the most important social institution. A preferable marriage for a man is with his father's brother's daughter, and among many groups marriages are invariably within the *berādari*. The lineage elders constitute a council that adjudicates disputes within the lineage and acts on behalf of the lineage with the outside world—for example, in determining electoral allegiances.

Pakistan's cultural heritage dates back more than 5,000 years, to the epoch of the Indus civilization. But the emphasis on Islāmic ideology has brought about a strong romantic identification with Islāmic culture—not only that of the Indian subcontinent but of the whole of the Islāmic world. *Qawwāli*, a form of devotional singing, is very popular. Poetry is also a popular rather than an esoteric art, and public poetry recitations called *mushā'irahs* are organized like musical concerts. Urdu, Sindhi, and Pashto poets are regional and national heroes. Literature is the richest of all Pakistani art forms; music and, especially, dancing are less-developed arts. The visual arts, too, play little part in popular folk culture. Painting and sculpture, however, have made considerable progress as expressions of an increasingly sophisticated urban culture.

Poetry

The cinema is the most popular form of entertainment. Many feature films are produced each year, mostly in Punjabi and Urdu. The songs and music used in the films have a distinctive character and are often reproduced on phonograph records and broadcast on the radio.

Government-owned radio and television have been used in an attempt to harness folk cultural traditions (especially in song, music, and drama) for political and nonpolitical propaganda purposes. Newspapers, in particular those published in Urdu, Sindhi, and English, have a wide readership. (H.A.A./S.J.B.)

For statistical data on the land and people of Pakistan, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

BACKGROUND TO PARTITION

The concept of a separate Muslim "nation" or "people," *qaum*, is inherent in Islām, but this concept bears no resemblance to a territorial entity. The proposal for a Muslim state in India was first enunciated in 1930 by the poet-philosopher Muḥammad Iqbāl, who suggested that the four northwestern provinces (Sindh, Balochistān, Punjab, and the North-West Frontier Province) should be joined in such a state. It was left to a Cambridge student to coin a name for this grouping: *Pakistan*, from *Punjab*, *Afghania* (North-West Frontier), *Kashmir*, and *Indus-Sind*. Alternatively the name was said to mean "Land of the Pure." (H.R.T.)

The Muslim League and Mohammed Ali Jinnah. The movement among the Muslim population of the India-Pakistan subcontinent that culminated in the creation of Pakistan stemmed from the historical fact that, for more than six centuries before the effective domination of the British in India, Muslim soldiers and administrators had controlled a population in which Hindus were a numerical majority, although mass conversions to Islām in economically backward areas like East Bengal (Bangladesh) produced local Muslim majorities. When the British replaced Muslim domination by their own, the tradition of rule prevented the Muslims from adapting themselves to the new situation as readily as the Hindus; but the failure of the risings of 1857 dashed Muslim hopes of a restoration of their authority. Later, while Hindus were pressing for constitutional reform through the Indian National Congress, the Muslims sought various guarantees to safeguard their minority position and finally founded their own political organization, the All-India Muslim League, at Dacca in 1906. For a more complete discussion of Pakistan under British rule and the history of Pakistan before British domination see INDIA: *History*.

The Muslim League

The gradual clarification of the British intention to grant self-government to India along the lines of British parliamentary democracy aroused Muslim apprehensions regarding ultimate political subjection to the Hindu majority. Mohammed Ali Jinnah, as eager as any Hindu nationalist to bring British rule to an end, was at length driven to the conclusion that the only way to preserve Indian Muslims from complete subordination to the Hindus was to establish a separate Muslim state. By 1940 the demand for Pakistan had been formally endorsed by the Muslim League under his leadership.

British policy, supported by the weight of the Hindu nationalist movement, laboured hard to avoid disrupting the economic and political unity built up during British rule. None of the suggested alternatives to separation of Pakistan commended themselves to Jinnah, whose leadership of the bulk of the community was unchallenged; without his cooperation—of which the price was Pakistan—Indian independence was impracticable. His courage and implacable determination triumphed in the end.

Birth of the new state. Pakistan came into existence as a dominion within the Commonwealth in August 1947, with Jinnah as governor-general and Liaquat Ali Khan as prime minister. With West and East Pakistan separated by more than 1,000 miles of Indian territory and with the major portion of the wealth and resources of the British heritage passing to India, Pakistan's survival seemed to hang in the balance. Of all the well-organized provinces of British India, only the comparatively backward areas of Sindh, Balochistān, and the North-West Frontier came to Pakistan intact. The Punjab and Bengal were divided, and Kashmir became disputed territory. Economically, the situation seemed almost hopeless; the new frontier cut off Pakistani raw materials from the Indian factories, disrupting industry, commerce, and agriculture. The partition and the movement of refugees were accompanied by terrible massacres for which both communities were responsible. India remained openly unfriendly; its economic superiority expressed itself in a virtual blockade. The dispute over Kashmir brought the two countries to the verge of war; and India's command of the headworks controlling the water supplies to Pakistan's eastern canal colonies gave it an additional economic weapon. The resulting friction, by obstructing the process of sharing the assets inherited from the British raj (according to plans previously agreed), further handicapped Pakistan. (L.F.R.W.)

ISLĀMIC REPUBLIC OF PAKISTAN

Mohammed Ali Jinnah died in September 1948, within 13 months of independence. The leaders of the new Pakistan were mainly lawyers with a strong commitment to parliamentary government. They had supported Jinnah in his struggle against the Congress not so much because they desired an Islāmic state but because they had come to re-

Ambivalent attitudes toward Islām

gard the Congress as synonymous with Hindu domination. They had various degrees of personal commitment to Islām. To some it represented an ethic that might (or might not) be the basis of personal behaviour within a modern, democratic state. To others it represented a tradition, the framework within which their forefathers had ruled India. But there were also groups that subscribed to Islām as a total way of life, and these people were said to wish to establish Pakistan as a theocracy (a term they repudiated). The members of the old Constituent Assembly, elected at the end of 1945, assembled at Karāchi, the new capital.

Jinnah's lieutenant, Liaquat Ali Khan, inherited the task of drafting a constitution. Himself a moderate (he had entered politics via a landlord party), he subscribed to the parliamentary, democratic, secular state. But he was conscious that he possessed no local or regional power base. He was a *muhajir* ("refugee") from the United Provinces, the Indian heartland, whereas most of his colleagues and potential rivals drew support from their own people in Punjab or Bengal. Liaquat Ali Khan therefore deemed it necessary to gain the support of the religious spokesmen (the mullahs or, more properly, the ulama). He issued a resolution on the aims and objectives of the constitution, which began, "Sovereignty over the entire universe belongs to Allāh Almighty alone" and went on to emphasize Islāmic values. Hindu members of the old Constituent Assembly protested; Islāmic states had traditionally distinguished between the Muslims, as full citizens, and *dhimmīs*, nonbelievers who were denied certain rights and saddled with certain additional obligations.

Political decline. Liaquat Ali Khan fell to an assassin's bullet in October 1951. Into his place as prime minister stepped Khwaja Nazimuddin, the leading member of the family of the nawab of Dacca. He was a Bengali aristocrat and a man of extreme personal piety. Nazimuddin had followed Jinnah as governor-general under the interim constitution. He was succeeded as governor-general by Ghulam Mohammad, a Punjabi, so that the twin pillars of power represented the two main regional power bases in West Pakistan and East Pakistan.

With Nazimuddin in office, militant Muslims, led by the Ahrars, a puritanical political group, called for the purification of national life. In 1953 they demanded that the Ahmadiyah sect be outlawed from the Islāmic community. Nazimuddin temporized, and rioting and arson enveloped Lahore and other Punjabi towns. The secretary of defense, Colonel Iskander Mirza, pressed the Cabinet into sanctioning martial law in Lahore, and order was restored. Ghulam Mohammad decided that Nazimuddin must go, although he had the support of the Constituent Assembly. The dismissal was effected, and Mohammad Ali Bogra became prime minister.

In March 1954 a general election was held in East Bengal (East Pakistan) to choose a new provincial legislature. The contest was between the official Muslim League and a "United Front" of parties from the extreme right (orthodox religious) to extreme left (quasi-Marxist). There was a landslide defeat for the Muslim League. At the head of the victorious opposition stood two politicians who had previously kept one foot in the Muslim League and the other in the camp of the Congress and regional politics; these were the aged Fazl ul-Haq, with his Workers and Peasants Party, and Hussein Shaheed Suhrawardy, with a new party, the Awami League. This result was a dramatic demonstration of the gulf between West and East Pakistan.

Electoral defeat of the Muslim League (1954)

The Constituent Assembly reflected the new political mood of Pakistan by attempting to curb the powers of the governor-general, Ghulam Mohammad, who retaliated by proclaiming the dissolution of that body. His action was validated by the Supreme Court, with the rider that a new assembly must be convened. This was produced by a system of indirect election. The ministry of Mohammad Ali Bogra was completely reorganized, and three newcomers were introduced as strongmen from outside politics: these were Major General Iskander Mirza, as minister of the interior, General Mohammad Ayub Khan, commander in chief, as minister of national defense, and Chaudhri Mohammad Ali, a senior civil servant, as minister of finance. Mohammad Ali Bogra had little support in the new assembly, and he was replaced by Chaudhri Mohammad Ali.

Ghulam Mohammad, whose health had broken down, was replaced as governor-general in August 1955 by Iskander Mirza. Mirza had no regional power base and little in common with any of the politicians. He insisted that his fellow administrator Chaudhri Mohammad Ali remain prime minister, and the Chaudhri was able to succeed in one objective over which his three predecessors had failed: he induced the politicians to agree to a constitution (February 1956). To create a better balance between the West and East wings, the provinces and parts of West Pakistan were amalgamated into one administrative unit.

The constitution of 1956 embodied the Islāmic provisions of the "aims and objectives" resolution of 1949 and declared Pakistan to be an Islāmic republic. The national parliament was to comprise one house of 300 members, equally representing East and West. Ten seats were reserved for women. The prime minister and Cabinet were to govern according to the will of the parliament, with the president exercising only reserve powers.

The constitution of 1956

Khan Sahib, a former premier of the North-West Frontier Province, was invited by the Muslim League to become the chief minister of the new "one unit" of West Pakistan. Soon after taking office, Khan Sahib was faced with a revolt against his leadership in the Muslim League, but he adroitly turned the tables by forming a new group, the Republican Party, out of dissident Muslim League assemblymen. In the National Assembly also, members adopted the Republican ticket, and Prime Minister Chaudhri Mohammad Ali found himself without a majority. He resigned in September 1956.

Iskander Mirza, then president, was compelled to accept an Awami League government headed by Suhrawardy but dependent on Republican support to retain office. For a time the combination worked, but the flimsy consensus of Pakistani politics soon began to dissolve into factionalism, regionalism, and sectarianism. Khan Sahib found his hold over the West Pakistan legislature slipping, and he asked the president to suspend the constitution. The East Pakistan legislature voted unanimously for autonomy in all matters except foreign affairs, defense, and currency. The country was to hold its first complete general election in 1958, but a dispute over the basis of the constituencies led to Suhrawardy's resignation. His successors proved ineffective, and the legislative process came to a halt.

Military government. President Mirza had made no secret of his dissatisfaction with the working of parliamentary democracy in Pakistan. On Oct. 7, 1958, a presidential proclamation announced that the political parties were abolished, the constitution abrogated, and the country placed under martial law, with General Mohammad Ayub Khan as chief martial-law administrator. Mirza announced that the martial-law period would be brief and that a new constitution would be drafted. On October 27 he swore in his new Cabinet.

General Ayub became prime minister, and three lieutenant generals were named to the Cabinet. The eight civilian members included businessmen and lawyers, one being a young newcomer, Zulfikar Ali Bhutto. On the evening of October 27 the new military ministers called on the president, with contingents of armed soldiers, and informed him that he was to resign. After a short interval, Mirza was exiled to London. A proclamation issued by Ayub announced his assumption of the presidency.

Presidency of Mohammad Ayub Khan

Martial law lasted 44 months. During that time a number of army officers took over vital civil-service posts. A number of politicians were excluded from public life under the Electoral Bodies (Disqualification) Order, or EBDO. A similar purge took place among civil servants.

Ayub sought to create political institutions that would express Islāmic ideals and foster national development. He initiated a plan for "basic democracies," directly elected by the people, as local units of development. Elections took place in January 1960. The Basic Democrats, as they became known, were at once asked to endorse Ayub's presidency and to give him a mandate to frame a constitution. Of the 80,000 Basic Democrats, 75,283 affirmed their support for Mirza in a referendum in February 1960. A constitutional commission was asked to advise on a

suitable form of government. Ayub accepted some of its proposals and substituted some of his own, aiming, he said, for "a blending of democracy with discipline." In the early days of Ayub's regime there were notable reform measures, such as the Muslim Family Laws Ordinance of 1961, restricting polygamy, but later the president found it necessary to make concessions to Muslims in order to bolster his regime.

One feature of the Ayub regime was the quickening pace of economic growth. During the initial phase of independence, the growth rate was less than 3 percent per annum and scarcely moved ahead of the rate of population growth. During the mid-1950s even this rate declined, but from 1960 to 1965 the rate advanced to more than 6 percent per annum. Development was particularly vigorous in the manufacturing sector.

There was considerable imbalance between East and West; during the 1950s East Pakistan was becoming poorer in per capita terms every year, whereas the West was achieving positive growth. A continuing grievance was the contribution made by East Pakistan to foreign exchange by the export of jute and tea, from which it was felt the West reaped more advantage; the West was also the major beneficiary of foreign aid.

The outstanding example of favoured treatment for the West was the great Indus basin scheme for hydroelectric development. Pakistan skillfully negotiated for assistance from the World Bank, the United States, and other friends. In addition to economic aid, Pakistan also received a great deal of military aid from the United States.

The war over Kashmir in 1965 had more far-reaching effects on Pakistan than on India. Ayub received a new mandate in January when he won decisively against a spirited challenge from Fatima Jinnah, the sister of Mohammed Ali Jinnah. In the early days of his presidency, Ayub had moved freely among the rural people, talking to them face-to-face. After the war, he withdrew behind a curtain of dictatorship, becoming a remote figure in a bulletproof limousine. Bhutto, the chief exponent of struggle against India, was relieved of office in 1966. Mujibur Rahman (Sheikh Mujib), who had inherited the leadership of the Awami League, the major force in East Pakistan, was arrested and accused of conspiring with India.

Decline of Ayub

Ayub's autocratic position was suddenly challenged in the autumn of 1968; an unsuccessful attempt on his life was followed by the arrest of Bhutto and other opposition leaders. Ayub summoned a conference of opposition leaders and withdrew the state of emergency under which Pakistan had been governed since 1965, but these concessions failed to conciliate the opposition, and in February 1969 Ayub announced that he would not contest the presidential election due in 1970. Protests and strikes flared everywhere, being especially militant in Bengal. At length, on March 25, 1969, Ayub resigned, handing over responsibility for governing to the commander in chief, General Agha Mohammad Yahya Khan. Once again the country was placed under martial law. Yahya assumed the title of president as well as chief martial-law administrator. He made it clear that his aim was an early general election, which took place in December 1970.

Civil war. The success of the Awami League in East Pakistan surprised even its friends. Sheikh Mujib emerged with a majority at his command among the membership of the new assembly (167 of the 300 total). But what upset all predictions was the victory in West Pakistan of Bhutto's Pakistan People's Party (PPP), which won particularly heavily in Punjab and gained a clear majority (83) of the representation from the West. Yahya's plan provided that when the new assembly met it must produce a constitution within 100 days. Mujib, however, stood out for complete independence for East Pakistan, except for foreign policy, though the East wanted to make its own aid, trade, and defense agreements. Bhutto rejected these terms and refused to bring his party to Dacca to participate in the assembly. On March 1, 1971, President Yahya announced that the National Assembly would be suspended indefinitely. Sheikh Mujib replied by ordering a boycott and general strike throughout East Pakistan. Bowing to the inevitable, Yahya proceeded to Dacca in mid-March to negotiate a compromise that would concede the substance of Mujib's demands while retaining tenuous ties that might still preserve the name of Pakistan. But compromise proved impossible. President Yahya denounced Mujib and his men as traitors and launched a drive to "reoccupy" the East with West Pakistan troops.

Warfare between government troops and supporters of the Awami League broke out in the East in March. Sheikh Mujib and many of his colleagues were arrested, while others escaped to India, proclaiming East Pakistan an independent state under the name Bangladesh ("Bengal Land"). As fighting continued, the number of refugees crossing the border into India grew into the millions. In December 1971 India successfully invaded East Pakistan. The establishment of a Bangladesh government with Mujib as prime minister followed in January 1972.

Establishment of Bangladesh

Bhutto's regime. Accepting responsibility for the defeat and breakup of Pakistan, President Yahya resigned on Dec. 20, 1971, and Bhutto became the undisputed leader of former West Pakistan. Bhutto's declared policy of Islāmic socialism brought few tangible changes, but his populism was undeniably successful. He became increasingly autocratic, however, suppressing criticism, jailing opponents, and using militant methods against the restive Pathans and Balochis. A new constitution was adopted on April 10, 1973, and Bhutto became prime minister.

In January 1977 Bhutto announced that elections would be held within two months, unfolding a national charter of peasant reform. Nine opposition parties hastily patched together the Pakistan National Alliance (PNA) and launched a demand for the Islāmic way of life in Pakistan. The campaign was marked by violence, with opposition candidates complaining of brutal discrimination. The results were a sweeping victory for Bhutto's PPP, although they were denounced as fraudulent by the PNA. Mounting protest soon brought chaos to Karāchi and other major cities, where Bhutto was compelled to call out the army and proclaim martial law. He tried to buy peace by offering concessions to the PNA leaders (most of whom were under arrest), but they would accept nothing short of a new election.

Zia ul-Haq's regime. To avoid total chaos, the chief of staff of the army, General Mohammad Zia ul-Haq, took over as chief administrator of martial law on July 5, 1977. His early efforts to create an acceptable political alternative had only limited success. He announced that elections would be held in 90 days, but it was clear that Bhutto was the only politician of mass appeal. In early September Bhutto was arrested and charged with attempted murder; on March 18, 1978, he was sentenced to death, and, after Supreme Court review, he was hanged on April 4, 1979.

Zia's efforts to create an acceptable political alternative had only limited success. Thirteen months after taking over the martial-law administration, he announced the formation of a civilian Cabinet of administrators, technocrats, and some political leaders drawn from the Muslim League and the religious parties. The PNA was now split, with most elements forming an opposition that demanded early elections, withdrawal of the army from Balochistān, and the introduction of a full Islāmic code of laws. A zealous Muslim, Zia had already imposed Islāmic criminal punishments such as flogging and maiming (these were formally enacted as law in February 1979), but he declined to meet the full opposition demand. On Sept. 16, 1978, he was proclaimed president of Pakistan.

Pakistan became a "frontline state" when the Soviet Union occupied neighbouring Afghanistan in December 1979. A guerrilla war began between Afghan mujahideen (freedom fighters) and the Soviet forces, and millions of refugees fled into Pakistan. The mujahideen used a number of refugee camps and other areas inside Pakistan as bases for their activities. The conflict was further internationalized when the United States channeled massive arms supplies to the mujahideen via Pakistan. This program included renewed U.S. aid to Pakistan of $4.2 billion for the years 1987 to 1992.

Another external pressure was the Islāmic revolution in Iran. Partly in response, Zia extended his own Islāmization program. In addition to Islāmic criminal laws, this

Islāmization

included interest-free banking and other measures in keeping with traditional Islāmic economic practice. A national referendum was held in December 1984 on the Islāmization measures, coupled with an endorsement of Zia's presidency for an additional five years; some 62 percent of those eligible were declared to have voted, with 98 percent of the voters in favour of both propositions. The opposition disputed this result, however, claiming that only 10 to 15 percent of the electorate had participated.

In February 1985 elections for the national and provincial assemblies were held. Political parties were not allowed to participate, but there was a high turnout, despite a boycott by the opposition. Zia chose as prime minister Muhammad Khan Junejo, a Sindhi politician who had previously served in Zia's Cabinet. Martial law was lifted in December 1985. In January 1986 Junejo announced the revival of the Pakistan Muslim League. Soon afterward, Benazir Bhutto, the daughter of Zulfikar Ali Bhutto, sought to re-form the PPP, but her attempts to arouse popular protest met with little success.

In late 1986 Karāchi, Quetta, and Hyderābād were rocked by riots between the *muhajir* majority and Pathans, originally from the North-West Frontier Province and Afghanistan. Ethnic violence continued through the early 1990s and spread to other ethnic groups and other cities in Sindh province.

In May 1988 Zia dissolved the national and provincial assemblies and dismissed the Junejo government, alleging that it was corrupt, weak, and inept. He announced that elections would be held within 90 days, but they were later postponed to November. In June a caretaker government was set up, with Zia acting as head of government.

The administration of Benazir Bhutto. On Aug. 17, 1988, Zia was killed in an airplane crash, together with his leading generals and the U.S. ambassador. The chairman of the Senate, Ghulam Ishaq Khan, a long-standing Zia supporter, took over as acting president.

Elections of 1988

In elections in November, the PPP, led by Benazir Bhutto, won 93 seats; the Islāmic Democratic Alliance (IDC), claiming the mantle of Zia, won 54 seats; and the remaining 58 seats went to independents and candidates from minor parties. Support for Bhutto in Punjab, with 60 percent of the population, was weak, and in subsequent provincial elections the IDC held this key province. PPP candidates became chief ministers of Sindh and the North-West Frontier Province. Bhutto had a mandate, but it was incomplete. In subsequent negotiations conducted by Ishaq Khan, who was elected president in December 1988, she had to make concessions in important areas of policy. Thus, Pakistan's commitment to the Afghan mujahideen continued, and the army retained its premier place in the system. In December 1988 Benazir Bhutto became the first woman to lead an Islāmic state.

Given office without real power, Bhutto responded by projecting her image on the national and international stage without attempting to make fundamental changes at home. Distrusted by the president and the military, she was ousted 20 months later. She was succeeded in November 1990 by a Punjabi industrialist, Nawaz Sharif. However, relations between Sharif and Khan were also tense. Khan dismissed Sharif as prime minister in April 1993, accusing him of mismanagement, corruption, and a "reign of terror." Khan dissolved the National Assembly and promised new elections for July, but the Supreme Court overturned his actions and reinstated Sharif and his government in May. Both Sharif and Khan resigned in July, reportedly because of pressure from the army chief of staff, General Abdul Waheed. In new elections in October the PPP won a majority of seats in the National Assembly, and Bhutto was elected prime minister. (H.R.T./Ed.)

The second Bhutto government and after. The Bhutto government's three-year rule was marked by steadily deteriorating economic conditions and growing ethnic and religious violence, particularly in southern Sindh where clashes between *muhajir*s and Pathans grew into pitched battles that left thousands dead. Violence by Islāmic extremist groups directed against the nation's small Christian community and the heterodox Aḥmadīyah sect continued to rise throughout the country, and in 1995 the government foiled a coup by fundamentalist military officers opposed to the idea of a woman ruling the country.

Bhutto's 1996 crackdown on violence, however, came just as allegations began to surface of financial mismanagement and corruption by her and her family. President Farooq Ahmed Leghari dismissed Bhutto's government in November, and in 1997 elections Pakistanis returned Sharif to office. The new prime minister quashed the president's power to dismiss elected governments and, likewise, abolished the Council for Defense and National Security, thereby earning the resentment of the military.

Sharif's inability to cope with the worsening economy and accusations that he had engaged in corruption far in excess of that alleged of Bhutto alienated many among Pakistan's political elite, especially members of the military who saw the government's failed economic policies as a threat to national security. The prime minister's willingness to respond to India's testing of five nuclear weapons in May 1998—within weeks Pakistan entered the nuclear club by detonating several of its own devices—failed to bring the military to his side, and in October 1999 the army chief of staff, General Pervez Musharraf, suspended the constitution and arrested Sharif on charges of treason.

Under Musharraf tension with India over Kashmir placed the two nuclear powers on the verge of serious armed conflict. Pakistan's support for the U.S. overthrow of the Taliban regime in Afghanistan in 2001 and its aid in tracking down Islāmic militants earned it the gratitude of the United States, and U.S. economic sanctions imposed after Pakistan's nuclear testing were lifted. Despite this gesture, the economy languished, and ethnic and religious violence continued to plague the country. (Ed.)

For later developments in the history of Pakistan, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 935, 936, 968, and 976, and the *Index*.

BIBLIOGRAPHY

Physical and human geography: B.L.C. JOHNSON, *Pakistan* (1979); and PETER R. BLOOD (ed.), *Pakistan: A Country Study*, 6th ed. (1995), are useful introductions to the people and land of Pakistan. Geography is discussed in O.H.K. SPATE and A.T.A. LEARMONTH, *India and Pakistan: A General and Regional Geography*, 3rd ed. rev. (1967); and K.U. KURESHY, *A Geography of Pakistan*, 4th ed. (1977). OMAR NOMAN, *The Political Economy of Pakistan, 1947–85* (1988), describes the impact of politics on the economy. VIQAR AHMED and RASHID AMJAD, *The Management of Pakistan's Economy, 1947–82* (1984), profiles the evolution of a dependent economy. W. CANTWELL SMITH, *Modern Islam in India: A Social Analysis*, 2nd ed. (1946, reissued 1974), is an account of Muslim intellectual and political movements in the creation of Pakistan. SHAHID JAVED BURKI, *Pakistan: A Nation in the Making* (1986), surveys Pakistan's evolution as a nation. ANWAR HUSSAIN SYED, *Pakistan: Islam, Politics, and National Solidarity* (1982), is a thoughtful discussion. LAWRENCE ZIRING, *Pakistan: The Enigma of Political Development* (1980), discusses the quest by Pakistan's people for democracy. G.W. CHOUDHURY, *Constitutional Development in Pakistan*, 2nd ed. rev. and enlarged (1969), and *Democracy in Pakistan* (1963), are oriented toward the "Muslim League" point of view; whereas TARIQ ALI, *Pakistan: Military Rule or People's Power?* (1970), offers a left-wing analysis. ZULFIKAR KHALID MALUKA, *The Myth of Constitutionalism in Pakistan* (1995), discusses Pakistan's failed quest for a constitution. AKBAR S. AHMED, *Pakistan Society: Islam, Ethnicity, and Leadership in South Asia* (1986), is a collection of essays. FREDRIK BARTH, *Political Leadership Among Swat Pathans* (1959, reissued with corrections, 1980), is a classic study in political anthropology. LAWRENCE ZIRING, *Pakistan in the Twentieth Century: A Political History* (1997), is a record of Pakistan's key political figures. ALTAF GAUHAR, *Ayub Khan, Pakistan's First Military Ruler* (1993, reissued 1996), is the most intimate study of the Ayub years. (H.A.A./S.J.B.)

History: KHALID B. SAYEED, *Pakistan: The Formative Phase, 1857–1948*, 2nd ed. (1968), discusses the background. HUGH TINKER, *Experiment with Freedom: India and Pakistan, 1947* (1967), analyzes the pressures that led to partition. LEONARD BINDER, *Religion and Politics in Pakistan* (1961), describes early problems. D.K. PALIT, *The Lightning Campaign: The Indo-Pakistan War, 1971* (1972), is an account by a scholarly Indian general. ANTHONY HYMAN, MUHAMMAD GHAYUR, and NARESH KAUSHKI, *Pakistan: Zia and After* (1989), is perceptive. ALASTAIR LAMB, *Birth of a Tragedy: Kashmir 1947* (1994), charts the origin of the dispute over the region. (H.R.T.)

Palestine

The term Palestine has been associated variously and sometimes controversially with a small region of the eastern Mediterranean. Both the geographic area designated by and the political status of the name have changed over the course of some three millennia. The region, or a part of it, is also known as the Holy Land and is held sacred among Jews, Christians, and Muslims. In the 20th century it has been the object of conflicting claims of Jewish and Arab national movements, and the conflict has led to prolonged violence and in several instances open warfare.

The word Palestine derives from Philistia, the name given by Greek writers to the land of the Philistines, who in the 12th century BC occupied a small pocket of land on the southern coast, between modern Tel Aviv–Yafo and Gaza. The name was revived by the Romans in the 2nd century AD in "Syria Palaestina," designating the southern portion of the province of Syria. After Roman times the name had no official status until after World War I and the end of Ottoman rule, when it was adopted for one of the regions mandated to Great Britain; in addition to an area roughly comprising present-day Israel and the West Bank, the mandate included the territory east of the Jordan River now constituting the Hashemite Kingdom of Jordan.

The name Palestine has long been in popular use as a general term to denote a traditional region, but this usage does not imply precise boundaries. The perception of what constitutes Palestine's eastern boundary has been especially fluid, although the boundary frequently has been perceived as lying east of the Jordan River, extending at times to the edge of the Arabian Desert. In contemporary understanding, however, Palestine is generally defined as a region bounded on the east by the Jordan River, on the north by the border between modern Israel and Lebanon, on the west by the Mediterranean Sea (including the coast of Gaza), and on the south by the Negev, with its southernmost extension reaching the Gulf of Aqaba.

The strategic importance of the area is immense: through it pass the main roads from Egypt to Syria and from the Mediterranean to the hills beyond the Jordan River.

Settlement is closely dependent on water, which is almost never abundant. The rainfall, which arrives in the cool half of the year, decreases in amount in general from north to south and from the coast inland. Perennial rivers are few, and the shortage of water is aggravated by the porous nature of the limestone rocks over much of the country.

For further reading on the political units most closely associated with Palestine, see the *Macropædia* articles EGYPT, ISRAEL, JORDAN, and LEBANON. For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 911, 922, 96/11, and 978, and the *Index*.

This article is divided into the following sections:

Physical and human geography 407
- The land 407
- The people 408

History 408
- The Stone Age and the Copper Age 408
- The Bronze Age 409
  - Early Bronze Age
  - Middle Bronze Age
  - Late Bronze Age
- The Iron Age 410
  - The Israelites in Palestine
  - Assyrian and Babylonian rule
  - The Persian empire
- From Alexander the Great to AD 70 411
  - The Ptolemies
  - The Seleucids
  - The Hasmonean priest-princes
  - The Herodian house and the Roman procurators
- Roman Palestine 415
- From the Arab conquest to AD 1900 415
  - The rise of Islām
  - ʿAbbāsid rule
  - The Fāṭimid dynasty
  - The Crusades
  - Ottoman rule
- From 1900 to 1948 417
  - World War I and after
  - The British mandate
  - The Arab Revolt
  - World War II
  - Post-World War II
  - Civil war in Palestine
- Palestine and the Palestinians (1948–67) 421
  - The partition of Palestine and its aftermath
  - The term "Palestinian"
  - Diverging histories for Palestinian Arabs
  - Emergence of a Palestinian identity
- The Arab-Israeli war of 1967 and its consequences 422
  - The PLO's rise as a revolutionary force
  - The PLO's struggle for Palestinian autonomy
- Negotiations, violence, and incipient self-rule 423
  - The Camp David Accords and the PLO
  - The dispersal of the PLO from Lebanon
  - Years of mounting violence
  - The *intifada*
  - Moving toward self-rule

Bibliography 424

Physical and human geography

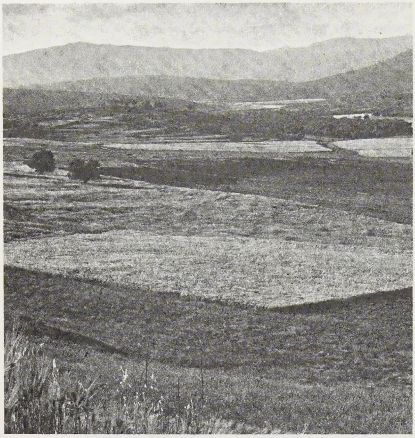
THE LAND

Topography

Of the coastal lowlands the most northerly is the Plain of ʿAkko (Acre), which extends with a breadth of 5 to 9 miles (8 to 14 kilometres) for about 20 miles (32 kilometres) from the Lebanon border in the north to the Carmel promontory, in Israel, in the south, where it is a mere 600 feet (180 metres) wide. Farther southward it opens out rapidly into the Plain of Sharon, about 8 miles wide and extending south to the latitude of Tel Aviv–Yafo. Once covered with marshes, the Sharon plain was reclaimed in the post-Exilic and Hellenistic period and is now a settled area. Fields and fruit groves are laid out between scattered sandstone ridges, on which villages have grown up. South of the spur of low hills that approaches the coast about Yafo, the maritime plain widens into the fertile region of Philistia, a district of orange groves, irrigated orchards, and fields of grain.

The Plain of Esdraelon (ʿEmeq Yizreʾel), formed by subsidence along lines of faults, separates the hills of southern Galilee from the mountains of Samaria. The plain, 16 miles wide at most, narrows to the northwest, where the Qishon River breaks through to the Plain of ʿAkko, and to the southeast, where the river (the modern Ḥarod), which rises at the Well of Harod (ʿEn Ḥarod), has carved the plain into the side of the Jordan Valley. Covered with rich basaltic soils washed down from the Galilean hills, Esdraelon is important both for its fertility and for the great highway it opens from the Mediterranean to the lands across the Jordan. The maritime plain connects with Esdraelon by the pass of Megiddo and several lesser routes between the mountain spurs of Carmel and Gilboaʿ.

The hill country of Galilee is better-watered and more thickly wooded than that of Samaria or Judaea. North of the Plain of Asochis, Upper Galilee, with heights of 4,000 feet, is a scrub-covered limestone plateau and is thinly peopled. To the south, Lower Galilee, with its highest



Farmland and mountain range near Nazareth, in Lower Galilee.
A.L. Goldman—Rapho/Photo Researchers

mountain, Tabor (1,929 feet), is a land of east-west ridges enclosing sheltered vales like that of Nazareth (Naẓerat), with rich basaltic soils.

Samaria

Samaria, the region of the ancient kingdom of Israel, is a hilly district extending from the Plain of Esdraelon to the latitude of Ramla. Its mountains—Carmel, Gilboa', Ebal, and Gerizim—are lower than those of Upper Galilee, while its basins, notably those of the 'Arrābah Plain and Nābulus (Shekhem), are wider and more gently contoured than their equivalents in Judaea. Samaria is easily approached from the coast over the hills of Ephraim and from the Jordan by the Fārī'ah valley. The city of Jerusalem has expanded rapidly along the mountain ridges.

From Ramla in the north to Beersheba in the south, the high plateau of Judaea is a rocky wilderness of limestone, with rare patches of cultivation, as around al-Bīrah and Mount Hebron (al-Khalīl). It is separated from the coastal plain by a longitudinal fosse and a belt of low hills of soft chalky limestone, about 5 to 8 miles wide, known as ha-Shefela. The Judaean plateau falls abruptly to the Jordan Valley, which is approached with difficulty along the wadis Kelt and Mukallik.

The Jordan Valley

The Jordan Valley is a deep rift valley that varies in width from 1.5 to 14 miles. In its northern section the bed of the drained Lake Ḥula and Lake Tiberias (the Sea of Galilee) are blocked by natural dams of basalt. Descending to about 1,310 feet below sea level, the valley is exceedingly dry and hot, and cultivation is restricted to irrigated areas or rare oases, as at Jericho or at 'En Gedi by the shore of the Dead Sea.

The Negev, a desertlike region, is triangular in shape with the apex at the south. It extends from Beersheba in the north, where 8 inches (200 mm) or more of rain falls a year and grain is grown, to the port city of Elat on the Red Sea, in the extremely arid south. It is bounded by the Sinai Peninsula on the west and the Great Rift Valley on the east. (W.C.B./Ed.)

THE PEOPLE

The social geography of Palestine in the 20th century, especially the area west of the Jordan River, has been greatly affected by the dramatic political changes and wars that have brought this small region to the attention of the world. In the early 1990s Jews constituted more than four-fifths of the population west of the Jordan, Muslims about one-seventh, and Christians, Druze, and various other minorities the rest. The Jewish majority is increasingly composed of persons born in Israel itself, although hundreds of thousands of immigrants have arrived since the founding of the State of Israel in 1948. The Arab minority is descended from Arabs who lived in the area during the mandate period and, in many cases, for centuries before that time. The majority of both Jews and Arabs are now urbanized. East of the Jordan the population is overwhelmingly Arab.

According to Jewish nationalists (Zionists), Judaism constitutes a basis for both religious and national (ethnic) identity. Palestinian nationalists usually emphasize that their shared identity as Arabs transcends the religious diversity of their community: thus, both Muslim Arabs, constituting about 14 percent of the Israeli population, and Christian Arabs, about 3 percent, identify themselves in the first instance as Arabs.

The Arab majority resident in the West Bank, west of the Jordan River, and the Gaza Strip, along the Mediterranean coast, and the still larger number of Arab Palestinians living outside the area (many in nearby countries such as Lebanon) have strongly opposed Israeli control and have feared an eventual annexation of the West Bank and Gaza by Israel. Most Jewish Israeli settlers living in the occupied territories would like to have this annexation take place; they think the land ought to be part of Israel proper.

Nationalism as a source of war

Both Zionists and Palestinian Arab nationalists have at various times in recent history claimed rightful possession of the area west of the Jordan River. The rivalry between the two groups and their claims have been major causes of the numerous Arab-Israeli Wars and the continuing crises in the region. Some members of each group still make such sweeping and mutually exclusive claims to complete control of the area, while others are now more willing to seek a peaceful compromise solution. (Ed.)

History

THE STONE AGE AND THE COPPER AGE

The Paleolithic Period (Old Stone Age) in Palestine was first fully examined by the archaeologist Dorothy Garrod in her excavations of caves on the slopes of Mount Carmel in 1929–34. (In this historical survey the term "Palestine" is used as conventional shorthand to refer to the area comprising most of contemporary Israel and parts of Jordan. The eastern boundaries of this area have fluctuated throughout history, and specific definitions are dependent on and will emerge from the historical context.) The finds showed that at this stage Palestine was culturally linked with Europe, and human remains were recovered showing that the inhabitants were of the same group as the Neanderthal inhabitants of Europe. The Mesolithic Period (Middle Stone Age) is best represented by a culture called Natufian, known from excavations at 'Ain Mallāha and Jericho. The Natufians lived in caves, as did their Paleolithic predecessors, but there is a possibility that they were experimenting in agriculture, for the importance to them of the collection of grain is shown by the artistic care that they lavished on the carving of the hafts of their sickles and in the provision of utensils for grinding. During the subsequent Neolithic Period (New Stone Age) humans gradually undertook the domestication of animals, the cultivation of crops, the production of pottery, and the building of towns (*e.g.*, Jericho by 7000 BC).

Natufian culture

Excavations also have provided a picture of events in Palestine in the 5th–4th millennium BC, during which the transition from the Neolithic to the Copper Age took place. It was probably in the 4th millennium that the Ghassulians immigrated to Palestine. Their origin is not known; they are called Ghassulians because the pottery and flints characteristic of their settlements first attracted attention in the excavations of Tulaylāt al-Ghassūl in the Jordan Valley. There was a permanent village site with several successive layers of occupation, and the site probably was associated with reasonably efficient agriculture. The phase can be called Copper Age, since copper axes were found at Tulaylāt al-Ghassūl, and this is confirmed by the finds at sites near Beersheba, with pottery and a

flint industry allied to those of Tulaylāt al-Ghassūl but not identical with them. At Beersheba there was a copper-working industry, which, presumably, imported ore from Sinai, and there was also evidence of an ivory-working industry, both proving the growth of a class of specialist craftsmen. Discoveries near 'En Gedi have revealed a shrine of this period, and basketry, ivory, leather, and hundreds of copper ritual objects were found in the Naḥal Mishmar caves of the Judaean desert.

The region in which the Ghassulian settlements have been found is mainly in the south of Palestine, with an extension up the coastal plain and its fringes. These settlements seem to have died out and disappeared in the last centuries of the 4th millennium, at about the same time as a new population immigrated, probably from the north. Thereafter, the composite elements in Palestine consisted of the indigenous Neolithic-Chalcolithic population, the Ghassulians, and these latest immigrants; in time the peoples were amalgamated into what was to become the sedentary urban population of the Early Bronze Age in the 3rd millennium.

Palestine dating is henceforth linked to Egyptian dating down to the time of the Hebrew monarchy; the interpretation of Egyptian dates in Rolf Krauss's *Sothis- und Monddaten* (1985) is followed in this article.

THE BRONZE AGE

Early Bronze Age. It was in the Early Bronze Age that most of the towns that are known in historic times came into existence. The growth of these towns can be approximately correlated chronologically with the development of the Old Kingdom in Egypt, Early Bronze I corresponding to the late Predynastic Period and Early Bronze II being cross-dated by finds to the time of the 1st dynasty, *c.* 2925 BC. Evidence of the early phases of the Early Bronze Age comes mainly from Megiddo, Jericho, Tell el Far'ah, Bet She'an, Khirbet Kerak, and Ai. All these sites are in northern or central Palestine, and it was there that the Early Bronze Age towns seem to have developed. The towns of southern Palestine, for instance Tell ed Duweir, Tell Beit Mirsim, and Tell el Hesi, seem only to have been established in Early Bronze III. The town dwellers, identified as the original Semitic population, can, for the sake of convenience, be called Canaanites, although the term is not attested before the middle of the 2nd millennium BC.

Early Bronze Age towns

In the course of the 3rd millennium, therefore, walled towns begin to appear throughout Palestine. There is no evidence that the next step of unification under the leadership of a single town took place in Palestine, as it had in Mesopotamia and Egypt; Palestine's towns presumably remained independent city-states, except insofar as Egypt may at times have exercised a loose political control. By about the 23rd century BC, the whole civilization had ceased to be urban. The next phase was pastoral and was influenced by the settlement of nomads probably from east of the Jordan River. Among the nomads, Amorites from the Syrian desert may have predominated. It is not yet fully understood how these events are related to the creation of the Akkadian empire in Mesopotamia under Sargon of Akkad and his grandson Naram-Sin (24th and 23rd centuries BC) and to the latter's destruction of the powerful kingdom of Ebla (modern Tall Mardikh) in neighbouring Syria, nor is the extent of Eblaite and Akkadian hegemony over Palestine in this period known. It does, however, seem reasonable to associate the incursion of nomads from the east with the invasions of Asiatics into Egypt that brought the Old Kingdom to an end. An initial date of 23rd–22nd century BC, depending on the interpretation of the Egyptian evidence, and a final date of the 20th century BC seem probable.

The picture of Palestine at this period is thus unequivocally that of a region occupied by a number of allied tribes; while they had many features in common, there were also many differences. The most significant point is that, with the possible exception of the northern group, they made no contribution at all to town life. The different groups had tribal centres, but they were essentially seminomadic pastoralists. This description fits well that in the Book of Joshua of the Amorites who lived in the hill country, as opposed to the Canaanites who lived in the plains and on the coast—areas favourable to agriculture.

Middle Bronze Age. It was, in fact, the next period—the Middle Bronze Age—that introduced the Canaanite culture as found by the Israelites on their entry into Palestine. The Middle Bronze Age (*c.* 2000–*c.* 1550 BC) provides the background for the beginning of the story of the Old Testament. The archaeological evidence for the period shows new types of pottery, weapons, and burial practices. Once more an urban civilization based on agriculture was established. It is not entirely clear whether the wave of urban development after the 20th century BC was the work of a new immigrant people accustomed to dwell in towns or of the local inhabitants themselves, some of whom may have adopted a sedentary lifestyle and begun, as in Mesopotamia and Syria, to establish dynasties. But where they settled there grew up towns of the widespread Middle Bronze Age civilization of Palestine. This civilization was intimately connected with that of the towns of the Phoenician-Canaanite coast. Extant Egyptian documents provide valuable information about Palestine in the period of the Egyptian 12th dynasty (1938–1756 BC) and argue for significant Egyptian interest and influence in Palestine at this time. (Most notable are the popular literary work known as the *Story of Sinuhe,* detailing the hero's exile in the Palestinian region, and the 20th–19th-century "Execration Texts," inscriptions of names of Egypt's enemies on pottery, which was ceremonially broken to invoke a curse.) The culture introduced at this stage was essentially the same as the culture found by the Israelites who moved into Palestine later, in the 14th and 13th centuries BC.

Egyptian influence in Palestine

A large repertory of new forms in pottery arose, and for the first time in Palestine it was turned entirely on a fast wheel. Comparisons of Palestinian early Middle Bronze pottery forms with metallic and ceramic forms at Byblos, dated by Egyptian contacts, suggest that these forms were brought to Palestine about the 19th century from coastal Syria. Bronze weapons of a distinctive type, paralleled also on the Syrian coast, are found at Megiddo, Jericho, and Tell el 'Ajjul. Town life in Palestine gradually expanded after the mid-19th century BC, but the material culture was essentially a direct development from the preceding stage. Several towns of Middle Bronze Age Palestine were defended by plaster-faced ramparts (clearly discernible at Jericho and many other sites), an imported method of fortification giving evidence of a new and alien influence superimposed upon the existing Canaanite culture. These were probably introduced by the Asiatic Hyksos, possibly related to the Amorites, who secured control of northern Egypt in about 1630. The Hyksos may have included elements of a grouping of people, largely Semitic, called the Habiru, or Hapiru (Egyptian 'Apiru). (The term Habiru, meaning "Outsiders," was applied to nomads, fugitives, bandits, and workers of inferior status; the word is etymologically related to "Hebrew," and the relationship of the Habiru [and aforementioned Hyksos] to the Hebrews has long been debated.) The Habiru appear to have established a military aristocracy in Palestine, bringing to the towns new defenses and new prosperity (as well as many Egyptian cultural elements) without interrupting the basic character of the local culture; this was to survive the destruction of Megiddo, Jericho, and Tell Beit Mirsim that followed the Egyptians' expulsion of the Hyksos into Palestine at the end of the Middle Bronze Age (*c.* 1550).

(K.M.K./G.R.B.)

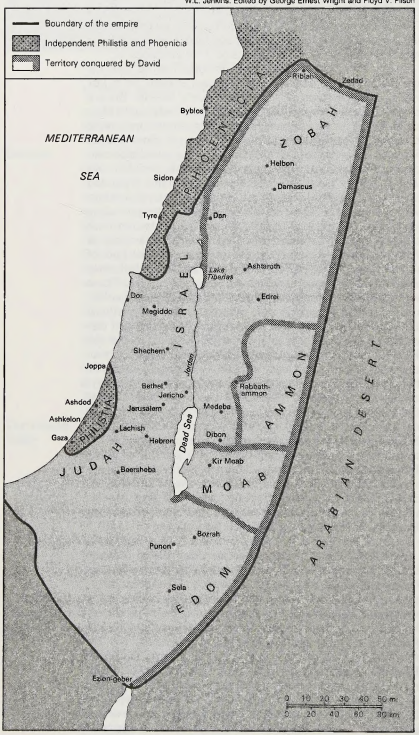
Late Bronze Age. There was no sharp break between the Middle and the Late Bronze ages in Palestine. Shortly before the death of Ahmose I (1514 BC), the first native pharaoh of the New Kingdom, the Egyptian armies began to conquer Palestine, probably completing their task during his successor's reign. Under Queen Hatshepsut (1479–58), Palestine revolted against Egyptian domination, but the rebellion was put down firmly by her successor, Thutmose III, who established a stable administration, maintained through the reigns of his immediate successors. Egyptian administrative documents excavated in both Egypt and Palestine show in considerable detail how the provincial government was organized and even how it operated during the century 1450–1350 BC. Documents show, for

Egyptian conquests of Palestine

example, that the land of Retenu (Syria-Palestine) was divided into three administrative districts, each under an Egyptian governor. The third district (Canaan) included all of Palestine from the Egyptian border to Byblos. This period is often known as the Amarna Age and is vividly illustrated by several hundred cuneiform letters found in Egypt at Tell el Amarna, site of the capital of the "heretic king" Akhenaton. The unusual concern of the pharaohs with the affairs of Palestine was chiefly a result of the fact that control of it was necessary for the defense of Phoenicia and southern Syria, menaced by Mitanni down to about 1375 and by the Hittite empire after that date.

About 1292 BC the increasingly weak rule of the last pharaohs of the 18th dynasty was replaced by the strong arm of the second and third kings of the 19th dynasty, Seti I and Ramses II (1279–13 BC). These kings blunted the southward thrust of the Hittites and consolidated the crumbling Egyptian empire. The exactions of foreign bureaucrats, however, combined with internal decay, had so enfeebled the Canaanite vassal princes of Palestine that it was comparatively easy for the incoming Israelites to occupy most of the hill country east of the Jordan River and in western Palestine in the course of the closing decades of the 13th century BC. Archaeological evidence suggests that the Israelite settlement in Palestine was much more complex and disconnected than the biblical accounts indicate. During a short interlude of anarchy that followed the last weak kings of the 19th dynasty, Egyptian rule was completely extinguished, and the ephemeral victories of Ramses III in the early decades of the 12th century scarcely affected Palestinian history.

Adapted from *The Westminster Historical Atlas to the Bible*, rev. ed., copyright 1945, by the Westminster Press; copyright 1956, by W.L. Jenkins. Edited by George Ernest Wright and Floyd V. Filson



Palestine during the time of David and Solomon.

THE IRON AGE

The Israelites in Palestine. Though the Israelite tribes entered Palestine before the end of the Late Bronze Age, they did not become firmly established in their new home until the early decades of the 12th century. Their number was increased greatly during the settling of Canaan by seminomadic Hebrew tribes already in Palestine, as well as by many settled Canaanites (*e.g.*, the Gibeonites), who joined the invaders against their sedentary neighbours. Excavation has made it clear that the Israelites began building amid the ruins of their precursors and that new settlements sprang up rapidly all through the hill country. Had events followed their normal course, the resurgent Canaanites, who had not been driven from the coastal plain or the Plain of Esdraelon, might have overwhelmed the scattered and unorganized Israelite clans, but this was prevented by the great invasion of the Sea Peoples in the time of Ramses III, in the early decades of the 12th century BC. Among the invaders from the Aegean basin were the Philistines, who were to conquer much of the region within a century and a half after their settlement in the southern coastal plain. (The Philistines have been identified with the so-called Peleset, who were used as garrison troops and mercenaries by Ramses III.) Meanwhile, three other peoples were settling east of the Jordan River: the Edomites in the south, the Moabites east of the Dead Sea, and the Ammonites on the edge of the Syrian Desert east of Gilead. Considered by the Israelites as fellow Hebrews, these peoples had begun to settle down before the Israelite invasion, and they remained polytheists until the end of the Old Testament period.

The invasion of the Sea Peoples

The early Israelites possessed a strong centralizing force in their monotheistic faith, combined with a stern code of ethics, which set them apart from all their neighbours. The Mosaic tradition of the covenant between Yahweh and Israel, made concrete by the tabernacle and its ritual, bound the tribes together in a cultic bond resembling the later Greek amphictyonies. Characteristic of all these organizations was a central sanctuary, surrounded by its worshipers. Straining against this religious bond were disruptive tribal forces held in leash by a loose alliance between the tribes, which was often severed by civil war. But for the constant attacks launched by its neighbours, Israel would perhaps never have attained any political solidarity. As it was, salvation from its foes lay only in union, and, after abortive attempts had been made at one-man rule, Saul became king of all Israel (*c.* 1020 BC).

He defeated the Ammonites and the Philistines but was killed in battle against the latter about 1000 BC and was succeeded by David. King David crushed the Philistines (*c.* 990) and conquered the three Hebrew states east of the Jordan River, after which the intervention of the Aramaeans from Syria forced him to defeat and annex the states of Aram as far north as the borders of Hamath on the Orontes. Farther east he established some sort of control over the nomadic tribes of the Syrian Desert as far as the Euphrates, though it is scarcely probable that Israelite domination was very effective. At home David organized a stable administration based largely on Egyptian models and, according to tradition, carried out a census of the population. He died before he could complete his plans, but they were put into effect by his successor, Solomon.

The reign of King David

The reign of Solomon (mid-10th century) represents the culmination of Israelite political history. Though Solomon gradually lost control over outlying territories conquered by David, he was extraordinarily successful in organizing the economic life of the country. He joined forces with Hiram of Tyre, who was leading the Phoenicians toward the exploitation of Mediterranean trade. Expeditions to Ophir, a region probably in either East Africa or India, brought rich items such as gold, peacocks, and sandalwood to Palestine. At the same time, the Israelite king entered

into trade relations with the Arabs as far south as Sheba (Yemen). These activities would have been impossible but for the development of new principles in shipbuilding and for the recent domestication of the Arabian camel and its use in caravans. Among the king's other undertakings was the construction of a fortress or storehouse at a site near the head of the Gulf of Aqaba. The modern site, Tall al-Khalīfah, may have been the biblical Ezion-geber. Most of the kingdom's wealth was spent in elaborate building operations, which included the Temple and royal palace in Jerusalem, as well as numerous fortified towns. The best-known of these are Megiddo, Hazor, and Gezer. But royal activities on such a vast scale cost more than was produced by foreign trade and the tribute of vassal states, and the Israelites themselves were forced to submit to conscription in royal labour gangs as well as to heavy levies of various kinds. It is not surprising that the people of northern Israel revolted after the great king's death, thus disrupting the united monarchy.

The rump kingdom of Israel lasted for two full centuries, sharing the worship of Yahweh and the Mosaic tradition with its smaller southern neighbour, Judah. After a period of intermittent warfare between Judah and Israel, King Asa of Judah entered into an alliance with the growing kingdom of Damascus, by which the latter attacked northern Israel, thus relieving pressure on Judah. This move cost Israel its territory to the east of the Jordan River and north of the Yarmuk River and ushered in a long series of wars between Israel and Damascus, which did not end until the capture of Damascus by the Assyrians in 732.

The best-known phase of Israelite history is the period during which the great prophets, Elijah and Elisha, flourished, under the Omrides of the 9th century. Omri himself, founder of the dynasty, selected Samaria as his capital and began the construction of elaborate defenses and royal buildings, which have been uncovered by excavations. His son Ahab was alternately hero and villain of the principal stories of the prophets; he became involved in complex international maneuvers, which ended with his ignominious death at Ramoth-Gilead. The dynasty of Omri closed amid torrents of blood (*c.* 841 BC); it was followed by the dynasty of Jehu, which lasted for nearly a century. This was a period of extreme oscillations, from the catastrophic defeat of Israel (*c.* 815 BC) and the destruction of its army by Hazael, king of Damascus, to the triumphs of Jeroboam II (*c.* 786–746 BC). Meanwhile, Judah also oscillated between periods of prosperity and weakness; when it was strong, it controlled Edom and the caravan routes of the south from Midian to the Mediterranean; when it was feeble, it shrank behind its own narrow boundaries. Great kings like Asa, Jehoshaphat, and Uzziah alternated with weak kings.

The dynasty of Omri

In 741/740 the death knell of independence in Syria and Palestine was sounded by the capture of Arpad in northern Syria by the Assyrian king Tiglath-pileser III. Events unfolded with dizzying speed. In 738 Israel and Judah paid tribute to Assyria for the first time in decades; in 733 the Assyrians devastated Gilead and Galilee, turning the entire land into Assyrian provinces except for the territory of two tribes, western Manasseh and Ephraim; in 732 Damascus was captured and Aram ceased to exist as a state; in 725 the siege of Samaria began. Finally, in the first months of 722, Samaria was taken and Israel became politically extinct.

Assyrian and Babylonian rule. Judah was left the sole heir of the glories of David and Solomon. Hezekiah (*c.* 715–*c.* 686), lured by promises of Egyptian aid, attempted to resist Assyria but was defeated and compelled to pay a crushing tribute. It is possible that only the timely intervention of an epidemic that decimated the Assyrian army of Sennacherib saved Judah from total devastation. The eloquent guidance of Isaiah restored the morale of the people, and even the weakness of Hezekiah's son Manasseh did not bring complete ruin. Another strong king, Josiah (*c.* 640–609), arose in time to restore the ebbing fortunes of Judah for a few years, during which much of the ancient territory of united Israel was brought back under the rule of the Davidic dynasty. Assyria was rapidly declining in power, and in 612 its hated capital, Nineveh, was destroyed by the Medes. Josiah's successful rebellion ended when he fell in battle against a more powerful contender for the Assyrian succession, Necho of Egypt.

Meanwhile, the Chaldean kings of Babylonia were rapidly gaining strength. Nabopolassar of Babylon and Cyaxares of Media divided the old Assyrian empire between them, and the former's son, Nebuchadrezzar, gained control of Syria and Palestine in swift campaigns. The defeated Egyptians, however, continued to intrigue in Palestine, whose native states repeatedly joined anti-Babylonian coalitions, all of which collapsed of themselves or were crushed by the Chaldean armies. Jerusalem was twice besieged, in 597 and after 589. Finally, about 587/586, it was stormed and destroyed. The prophet Jeremiah, who had foreseen the tragic denouement and had repeatedly warned his people against their suicidal policy, died in Egypt. Judah was devastated and almost depopulated.

Rise of the Chaldean kings

The Persian empire. In 539 Cyrus II of Persia followed up his triumph over Media by conquering Lydia and Babylonia, thus making himself ruler of the greatest empire thitherto known. In the administrative reforms implemented by Darius I (reigned 522–486 BC), Phoenicia, Palestine-Syria, and Cyprus constituted the fifth province (satrapy) of the Persian empire (Herodotus, *The History*, 3.91).

One of Cyrus' first acts was to decree (*c.* 538) the restoration of Judah and the rebuilding of the First Temple of Jerusalem. A large number of Jewish exiles in Babylonia returned to Jerusalem, and work on the Second Temple was begun. The political situation was extremely unfavourable, however, since Judah south of Hebron had been occupied by Edomites escaping from Arab pressure, while the tiny remainder north of Hebron had passed under the control of the governor of Samaria. In spite of political intrigues to prevent completion of the work of rebuilding the Temple, the Jews took advantage of the civil wars and rebellions that racked the empire after the accession of Darius I to press forward with the work. They were urged on by the fiery prophets Haggai and Zechariah. In 515 BC the Second Temple was finished, but the Jews had meanwhile aroused the suspicion of the Persian authorities, and further efforts to improve their situation were discouraged.

Completion of the Second Temple

Matters rested in this unsatisfactory state until about 445 BC, when the Jewish royal favourite, Nehemiah, deeply stirred by reports of the sorry condition of Judah and Jerusalem, succeeded in obtaining the Persian ruler's support for a mission to Palestine. Under Nehemiah's leadership, Jerusalem's walls were rebuilt. Knowledge of the exact sequence of events is complicated by the confused state of the documentary sources, and the chronology of events in the time of Nehemiah and Ezra, who became a leader of the Jews who had returned to Jerusalem from Babylonia, is not certain. There is good evidence to suggest that Ezra's return to Jerusalem should be dated to 398 BC, early in the reign of the Persian king Artaxerxes II. The mention of Persian intervention in the priestly affairs of the high priest Johanan can reasonably be associated with Ezra's reform activities. In any event, Nehemiah and Ezra were able to establish both the religious autonomy of Judah and the practice of normative Judaism so firmly that they continued with little change for several centuries.

Information concerning the history of Palestine in the period following the age of Nehemiah and Ezra is very scanty. It is known that the province of Judah continued to be administered by high priests who struck their own coins and that the provinces of Samaria and Ammon remained under governors of the houses of Sanballat and Tobiah. In 343 Artaxerxes III Ochus is said to have devastated parts of Palestine in connection with his reconquest of Egypt. Eleven years later the country passed into Macedonian hands after Alexander's conquest of Phoenicia.

(W.F.A./G.R.B.)

FROM ALEXANDER THE GREAT TO AD 70

To Alexander, Palestine was, as to many before him, a corridor leading to Egypt, the outlying Persian province. Consequently, in his attack on that province after the Battle of Issus (333), he confined his attention, in his passage

southward, to reducing the coastal cities that might form bases for the Persian fleet. He left the Jews undisturbed in their religion and customs. The high priest remained the head of the Jewish state, perhaps assisted by a council of elders.

The Ptolemies. After the death of Alexander in 323, Palestine, with much of Syria and Phoenicia, fell to Ptolemy, who established himself as satrap in Egypt in 323 BC and adopted the title of king by 304. (After the death of Ptolemy, the Ptolemaic dynasty ruled Egypt for 300 years.)

Alexander's successors

The successors of Alexander, including Ptolemy and Seleucus, defeated Antigonus Monophthalmus, one of Alexander's generals, who had almost succeeded in recreating under his sole rule Alexander's great empire, at Ipsus in Phrygia in 301 BC. This victory confirmed Ptolemy in his possession of his territory, although he had arrived too late for the battle, and Seleucus, whose participation in it had been decisive, at first disputed Ptolemy's claim to Syria and Phoenicia and actually occupied northern Syria. This early dispute laid the foundations of a century of bitter antagonism between the house of Ptolemy and that of Seleucus, an antagonism that led to war five times within a century and was finally stilled only when Palestine, in 200 BC, became part of the Seleucid kingdom. The northern boundary of the kingdom established by Ptolemy I lay apparently slightly north of the modern Tripoli, perhaps on the course of the Kabir River (ancient Eleutherus), and there are no signs of any important change in this frontier throughout the next century.

Of Ptolemaic rule in the southern part of this territory, Palestine, little is known. What little can be discovered, mainly from writers of a later period, especially the author of the First Book of Maccabees and Josephus, suggests that, unlike the northern region, known as Syria and Phoenicia, the area was left in much its previous state, with considerable power and authority in the hands of the native chieftains.

More is known of taxation than of administration. A story preserved by Josephus (*The Antiquities of the Jews*, xii, § 154 ff.) indicates that tax farming, whereby the right to collect taxes was auctioned or was awarded to privileged persons, was employed for the collection of local taxes. It seems likely that there were additional extraordinary taxes levied by edict from Egypt.

Zenon's records of life in Palestine

Knowledge of the economic and commercial life of Palestine in the middle of the 3rd century is, on the other hand, fuller and more reliable. It is drawn from the dossier of letters received and written by Zenon, the confidential business manager of the chief minister of Ptolemy II Philadelphus (285–246). In 259 Zenon was sent to Palestine and Syria, where his master had commercial interests. His letters speak particularly of a trade in slaves, especially of young girls for prostitution, in whom there appears to have been a brisk trade, with export to Egypt. Zenon's records also testify to a considerable trade in cereals, oil, and wine. Inevitably, like all imports to Egypt, Palestinian exports worked under state monopoly, without which the internal monopolies of Egypt would have been undermined. Palestine, like Egypt and Syria, seems to have had no economic freedom under Ptolemaic rule; in all transactions the hand of the government's agents is clearly visible.

Of the material culture of Palestine in the Ptolemaic period, far less is known. The population seems, as in Syria, to have been divided between the Hellenized cities (*poleis*) of the coast, notably Ascalon and Joppa, and the rural population living in villages (*komai*). The fact that several cities had Ptolemaic dynastic names (Philadelphia, Philoteria, Ptolemais) must not lead to the conclusion that the early Ptolemies wanted to urbanize and raise the standard of living of the people of Palestine. It seldom appears that they did more than rename a previously existing city (*e.g.*, Scythopolis for Bet She'an)—a practice not uncommon in the Hellenistic world. In fact, unlike the Seleucids, the Ptolemies do not appear to have been great city builders. Nor do they appear to have encouraged the outward forms of independence in local government. It seems likely from the story of the Phoenician tax farmers mentioned above that considerable authority lay in the hands of the wealthy; yet the fact that both Ascalon and Joppa issued Ptolemaic regal coinage, but apparently no autonomous bronze coinage, suggests a rigid control. The absence of epigraphic evidence from the cities of Ptolemaic Palestine, however, renders any judgment about the conditions of the cities hazardous. Archaeology, too, helps but little, inasmuch as the buildings of Roman Palestine have swamped most of the Hellenistic remains. One exception must be noted: the tombs of the Hellenized Sidonian military settlers at Marisa in Idumaea, the walls of which are decorated with frescoes of fine hunting scenes, indicate that Hellenic civilization had been embraced by the non-Greek population in this period.

The chronic state of hostility between Egypt and the Seleucid house, which in much of the 3rd century BC had been concerned with the coastal regions of western Asia Minor, received a new impetus with the accession to the Seleucid throne of the energetic Antiochus III the Great (223–187), who aimed to win southern Syria and Palestine from Egypt, now weakly governed, and thereby establish the frontier to which Seleucus I had unwillingly renounced his claim in 301 BC. After his decisive defeat by Ptolemy IV at Raphia in 217, however, Antiochus was for several years occupied with internal troubles, and it was therefore not until about 200 BC that he could think again of an attack on Egypt. There a child—Ptolemy V Epiphanes—had recently ascended the throne, and the government was in the hands of overly powerful ministers who were more concerned with feathering their own nests than with preserving the integrity of the kingdom. At Panion, on the northern boundary of Galilee, the armies of Antiochus and Ptolemy met, and Ptolemy was defeated. Thus, the Ptolemaic possessions north of the Sinai desert, including Palestine, passed into the hands of the house of Seleucus.

The Seleucids. The Seleucids brought to the problem of the administration of Palestine a different tradition from that which had been behind Ptolemaic rule. The latter was based on that careful exploitation of territory that was possible in a small and closely knit land like Egypt and had thence been extended to the Ptolemaic provinces. This had never been possible for the Seleucids, who had always been masters of regions so vast as to render a unified and absolute control impossible. There is no sign of oppression in Seleucid government of the native peoples—they seem, on the contrary, to have aimed at improving the natives' status as far as possible, largely through bringing them into contact with Greek modes of urban existence.

It might then be supposed that Seleucid rule would have been popular in Palestine. In fact, however, it was under Seleucid rule that the great uprising of the Jewish people, the revolt of the Maccabees, occurred. The explanation of this paradox is perhaps twofold. First, the Seleucids were in need of money, and, second, the throne, at a critical time, was occupied by a tactless and neurotic king. Knowledge of Seleucid rule in Palestine before the accession of Antiochus Epiphanes is slight. The years from 188 BC onward were lean years for the dynasty, because the war with Rome, which had ended (in 189) in a complete Roman victory, had cost it not only almost the whole of Asia Minor but also a yearly indemnity of 15,000 talents. It is therefore not surprising that the first glimpse of Seleucid rule in Palestine tells of an attempt by Heliodorus, the leading minister of Seleucus IV (187–175), to deprive the Second Temple in Jerusalem of its treasure. His failure was soon ascribed to divine protection.

Antiochus Epiphanes

With the accession of Antiochus Epiphanes in 175 BC, relations rapidly deteriorated. Antiochus appears to have aimed at a wholesale restoration of the Seleucid empire in the east, including an occupation of Egypt, as a counter to the loss of the western province occasioned by the Peace of Apamea. He made an unwise beginning in Palestine by establishing a philhellene high priest, and it is clear from this and from his whole subsequent policy that he wished to extirpate Jewish religion from its central stronghold (there is no indication that he persecuted Jews of the Diaspora living in the cities of his kingdom).

Antiochus invaded Egypt in 170 or 169 BC, returning to Syria by way of Jerusalem, where he and his army despoiled the Temple of all its wealth. Two years later, after

his humiliating expulsion from the gate of Egypt by the Roman legate, he sent a financial official to exact taxes from the cities of Judaea. Antiochus' official attacked the city of Jerusalem by guile and largely destroyed it. He then built a fortified position on the citadel, called by the Greeks the Akra. This became the symbol of Judah's enslavement, though in itself the presence of a royal garrison in a Hellenistic city was by no means unusual. Its imposition was followed by an open attack on religious practice; many rites were forbidden. Noncompliance with the order—which contained many items calculated to raise the bitterest resistance in the hearts of law-abiding Jews, such as the prohibition of circumcision and the abolition of the observance of the Sabbath—was punishable by death. Finally, on the 25th day of the Hebrew month Kislev (December) in 168 BC, the "abomination of desolation," namely the altar of Zeus, was set up in the Temple in Jerusalem. It was this above all that summoned forth the resistance of the sons of the aged priest Mattathias; thus began the Maccabean revolt, led by Judas Maccabeus.

The Maccabean revolt

The resistance, it must be emphasized, came from only a section of the population. The century and a half of Greek rule had Hellenized much of the upper class of Jerusalem, and some of the characteristic features of Greek city life—such as the ephebic institute, for the training of young men, and the gymnasia—had been established on the initiative of this section of the ruling class, which was able to accept a less radical observance of Judaism and combine it with loyalty to the throne. Throughout the revolt, and indeed until the closing days of the Hasmonean dynasty established by the Maccabeans, this Hellenized element had to be taken into account.

Judas Maccabeus proved himself a leader of high quality. He successfully resisted the weak forces sent by the Seleucid authorities, and after three years of intermittent warfare he succeeded in purifying the Temple (165 BC), though the Akra remained in Seleucid hands until 141 BC.

After the death of Antiochus Epiphanes in 164 BC, the numbers of claimants to the Seleucid throne made a continuous policy toward Palestine impossible, because each claimant felt the need to seek support wherever it might be found. Thus, Jewish high priests were bribed by the kings and dynasts of Syria. This development enabled Judas and those who succeeded him to hold their own and eventually to establish a hereditary dynasty, known as the Hasmonean for their ancestor Hasmoneus. Soon after Antiochus Epiphanes' death, an agreement was reached with the Seleucid regent Lysias (who feared the appearance of a rival in Syria) through which the Jews received back their religious liberty. But, at the same time, the regular practice of pagan worship, beside the Jewish, was established, and a Seleucid nominee was appointed high priest. Thus were laid the seeds of fresh revolt.

Almost immediately Judas again took the field and scored a considerable victory over Nicanor, the Seleucid general, in which Nicanor was killed. Within two months, however, Demetrius I, the Seleucid king, sent Bacchides to take up a position near Jerusalem, and, in the engagement that followed, Judas lost his life (161/160).

The Hasmonean priest-princes. In the following years, dynastic disputes within the Seleucid empire prevented a succession of rulers from settling the Palestinian question. These circumstances allowed first Jonathan (161/160–143/142), the brother and successor of Judas, and then his brother Simon (143/142–134) to attain power. In 153 one of the Seleucid pretenders, Alexander Balas, in order to outplay the legitimate king, Demetrius I, granted Jonathan the office of high priest and gave him the Seleucid rank of a courtier, thereby legitimizing his position.

Beginning of secular rule

When Simon succeeded Jonathan, he acquired the status of a recognized secular ruler; the year of his assumption of rule was regarded as the first of a new era, and official documents were dated in his name and by his regnal year. He secured from the new Seleucid monarch, Demetrius II, exemption from taxation for the Jews.

In 142–141 Simon forced the Syrian garrison on the Akra to surrender, and the Jews passed a decree in his honour, granting the right of permanent incumbency to Simon and to his successors, until "an accredited prophet"



Palestine during the Maccabean period.

should arise. It was thus in Simon's reign that the rule of the priest-prince was transformed into a secular hereditary rule. The Seleucid king recognized this, granting Simon the right to issue his own coins.

Simon's son, John Hyrcanus I (134–104), suffered an initial setback at the hands of the last great Seleucid king, Antiochus VII Sidetes, who set out to reconquer Palestine, but at the latter's death John renewed his father's expansionist program, which resulted in the conquest and destruction of Samaria. In internal policy, however, he committed the grave error of quarreling with one of the two main Jewish ecclesiastical parties, the Pharisees—who followed the Law with great strictness and with whom the Maccabean movement had in origin close affinity—and siding with their opponents, the more liberal Sadducees. This is an early instance of that denial of the revolutionary origin of the movement which became entirely obvious in the reign of Alexander Jannaeus. Hyrcanus I was succeeded by Aristobulus I (104–103), who extended Hasmonean territory northward and is said to have assumed the title of king (basileus), though on his coins he appears, like Hyrcanus I, as high priest.

The reign of his brother and successor, Alexander Jannaeus, was long (103–76) and largely filled with wars. He imposed his rule rigorously over an increasingly large area, including both the cities of the coast and the area east of the Jordan River. Still more clearly than Hyrcanus I, he attests the change in direction and aim of the Hasmonean house. He was the bitter enemy of the Pharisees, his coins bear Greek as well as Hebrew legends, and his title on them is simply "King Alexander." He was succeeded by his widow, Salome Alexandra, who reversed his policy and

was guided by powerful religious advisers, members of the Pharisaic movement. After her death in 67 BC her two sons Aristobulus II and Hyrcanus II fought for the succession. Hyrcanus was defeated but was encouraged to reassert his rights by Antipater, an Edomite, son of the governor of Idumaea and father of the future Herod the Great.

At this stage the Romans appeared on the scene. Pompey the Great, during his reorganization of the lands of the newly conquered Seleucid kingdom, also arranged the affairs of Palestine (63 BC). He attempted to arbitrate between the brothers and eventually, after he had laid siege to and captured Jerusalem, appointed Hyrcanus II as high priest without the title of king adopted by his predecessors. He also imposed taxes on the Jews and curtailed Jewish dominions, granting virtual autonomy to a group of 10 or 11 Hellenized cities in Syria and Palestine, thenceforth to be known as the Decapolis, and placing them under the jurisdiction of the newly appointed governor of Syria. Pliny the Elder lists these cities as Damascus, Philadelphia (modern Amman, Jordan), Raphana, Scythopolis, Gadara, Hippos, Dion, Pella, Gerasa, and Canatha. On the basis of evidence in inscriptions, Abila can be added to the list. Thus, despite the name Decapolis, the actual number appears to have been 11. All of these cities, except for Scythopolis, are located east of the Jordan River, extending from Damascus in the north to Philadelphia in the south. Except for Damascus, all the other cities lie immediately to the east of Galilee, Samaria, or Judaea. Whether the Decapolis geographically belonged to Syria, to Coele Syria ("Hollow Syria"; *i.e.*, the southernmost region of Syria, which may include Palestine and is sometimes mistakenly limited to the modern al-Biqā' valley), or to Arabia, often identified as the land east of the Jordan River, is not clear, especially in the ancient geographers. In any event, the Decapolis ended as a political entity when Arabia was annexed in AD 106; the cities were distributed among the three provinces of Arabia, Judaea, and Syria. It appears that Philadelphia, Gerasa, and probably Dion went to Arabia, Damascus certainly to Syria, and the rest to Judaea or Syria.

The Decapolis

After the death of Pompey, however, the power of Antipater and his family greatly increased. Hyrcanus II became a figurehead of no importance, and Antipater himself, in return for services to Julius Caesar, received Roman citizenship and was awarded the title of "procurator of Judaea," while his sons Phasael and Herod became governors (strategoi) of Jerusalem and Galilee, respectively. The unexpected occupation of Palestine by Parthian troops in 40 BC altered the situation. Antigonus, the son of Aristobulus and therefore a legitimate Hasmonean, won the favour of the Parthians and was established by them as king and high priest of Jerusalem. Phasael was reported to have committed suicide, while his brother Herod escaped to Rome.

The Herodian house and the Roman procurators. Herod, in Rome, was recognized by the Senate, with the approval of Octavian and Mark Antony, as king of Judaea (40 BC) and returned to Palestine in 39 BC. Shortly afterward Roman troops expelled the Parthians, whose popularity in Palestine had been and subsequently remained considerable. After struggles against Antigonus, the Parthian nominee, in which he was assisted by Roman troops, Herod eventually captured Jerusalem. At about the same time, he married a niece of Antigonus, thus probably consoling those who remained loyal to the memory of the almost defunct Hasmonean house. Antigonus, when he fell into the hands of his enemies, was executed by order of Mark Antony.

The accession of Herod, a Roman protégé and an Edomite, brought to Palestine the peace that in the years of independence it had often lacked. His long reign (37–4 BC) was marked by general prosperity; his new city of Caesarea Maritima received lavish praise from Josephus for its spectacular port and sewer system. Between 31 and 20, Augustus restored to him the Jewish territories that Pompey had taken away, and in this enlarged kingdom he created a sound administrative system of Hellenistic type. Toward the end of his life the complex demands of a vast family (he had at least nine wives) led him into difficulties

Herod's reign



Palestine during the time of Herod the Great and his sons.
Adapted from A.H.M. Jones, *The Herods of Judaea*, The Clarendon Press, Oxford

regarding the succession, and it was then that he developed into the gruesome and vicious figure that Christian tradition has made so familiar. He had Mariamne and several of his sons put to death to prevent them from succeeding him; and on his death in 4 BC the country again entered a period of divided rule, which led to the reestablishment of direct Roman government. Augustus decided later that year, in the presence of three surviving sons of Herod, that Archelaus should rule Judaea, Samaria, and Edom (*i.e.*, central and southern Palestine); Herod Antipas rule Galilee and Peraea (east of the Jordan River); and Philip rule Trachonitis, Batanaea, and Auranitis (the area between the Decapolis and Damascus).

The fates of the rulers (of whom Philip and Herod Antipas were called tetrarchs, Archelaus ethnarch) and their territories were different. Philip, the most peaceable of the three, ruled the northern area until his death in AD 34. Antipas reigned in Galilee and Peraea until AD 39 but was then banished by Caligula on the ground that he had parleyed with Rome's enemies. Archelaus reigned for 10 years only; he was removed at the request of his subjects in AD 6. The region under Archelaus' rule (*i.e.*, Judaea, Samaria, and Edom) became the province of Judaea and passed to a series of undistinguished Roman prefects, the last of whom (Pontius Pilate, AD 26–36, under whom Jesus was crucified) lost office for the unnecessary massacre of some Samaritans. Palestine finally was united under the emperor Caligula's protégé, Herod Agrippa I, who succeeded Philip in the north in AD 37. The tetrarchy of Antipas was added soon after his removal in 39, and the territories of Judaea, Samaria, and Edom were added in 41, so that from 41 to his death in 44 Agrippa ruled the kingdom of his grandfather, Herod the Great, from Jerusalem. In 44 the entire kingdom passed under Roman rule and was reconstituted as the procuratorial province of Judaea.

Roman rule in Judea

Disturbances in the early years of procuratorial rule were frequent and largely caused by maladministration. Serious trouble arose under Ventidius Cumanus (48–52); and

under his successor, the imperial freedman Felix (52–60), rebellion was open though sporadic. The incompetence and anti-Jewish posture of Gessius Florus, procurator 64–66, led in 66 to the decisive and final outbreak. Florus, the heir to a long tradition of hostility between the large Hellenized populations of Palestine and the Jews (also a problem in the Diaspora, most notably at Alexandria during the reign of Caligula), allowed the Greek population of Caesarea Maritima to massacre the Jews of that city with impunity. Greeks in other towns of Palestine repeated the assault. In turn, the Jews responded by slaughtering Gentiles in Samaria, Galilee, and elsewhere. Soon Florus had lost control of the situation.

The organization of the Jews was better than it had previously been, and they were successful in an early engagement against the governor of Syria, who had advanced to Palestine with two legions to assist the hard-pressed procurator.

In 67, however, Vespasian, the future emperor, with his son Titus, arrived with a force of about 60,000 men, and the war became increasingly bitter. By the end of 67 Galilee was captured, and Judaea was reduced in three campaigns, which ended with the fall of Jerusalem in AD 70. The Temple was destroyed, though tradition recorded that Titus gave orders that it was to be spared, and the city became the permanent garrison town of a Roman legion. By AD 73 all resistance had ceased. (P.M.F./G.R.B.)

ROMAN PALESTINE

After the destruction of Jerusalem, a legion (X Fretensis) was stationed on the site, and the rank of the provincial governor was raised from *procurator* to *legatus Augusti*, signifying a change from equestrian to senatorial rank. Caesarea Maritima, the governor's residence, became a Roman colony, and, as a reward for the loyalty of the Greeks in the revolt, a new pagan city, Neapolis (modern Nablus), was founded at Shechem, the religious centre of the Samaritans.

The Jews, deprived of the Temple, founded a new religious centre in the rabbinical school of Jamnia (Jabneh). When a revolt broke out in AD 115, the Roman emperor Trajan appointed the first consular legate of Judaea, Lucius Quietus, to put it down. The rank of the legate confirms that two legions were stationed in Judaea, one at Jerusalem, the other at Caparcotna in Galilee, and thenceforth the province must have held consular status.

Bar Kokhba's revolt

In AD 132 the emperor Hadrian decided to build a Roman colony, Aelia Capitolina, on the site of Jerusalem. The announcement of his plan, as well as his ban on circumcision (revoked later, but only for the Jews), provoked a much more serious revolt, led by Bar Kokhba. It was ruthlessly repressed by Julius Severus; almost 1,000 villages were destroyed and more than half a million people killed, according to certain accounts. In Judaea proper the Jews seem to have been virtually exterminated, but they survived in Galilee, which, like Samaria, appears to have held aloof from the revolt. Tiberias in Galilee became the seat of the Jewish patriarchs. The province of Judaea was renamed Syria Palaestina (later simply called Palaestina), and, according to Eusebius (*Historia Ecclesiastica* iv. 6), no Jew was thenceforth allowed to set foot in Jerusalem or the surrounding district. This prohibition apparently was relaxed some time later to permit Jews to enter Jerusalem on one day a year, a Day of Mourning. While this ban was officially still in force as late as the 4th century AD, there is some evidence that from the Severan period onward (after AD 193) Jews visited the city more frequently, especially at certain festival times, and even that there may have been some Jews in residence. About the time the Bar Kokhba revolt was crushed (AD 135), Hadrian proceeded to convert Jerusalem into a Greco-Roman city, with a circus, an amphitheatre, baths, and a theatre and with streets conforming to the Roman grid pattern. He also erected temples dedicated to Jupiter and himself (Aelia was his clan name) on the very site of the destroyed Temple of Jerusalem. To repopulate the city, Hadrian apparently brought in Greco-Syrians from the surrounding areas and even perhaps some legionary veterans. The urbanization and Hellenization of Palestine was continued by the emperor Septimius Severus (reigned AD 193–211), except in Galilee, where the Jewish presence remained strong. New pagan cities were founded in Judaea at Eleutheropolis and Diospolis (formerly Lydda) and at Nicopolis (formerly Emmaus) under one of Severus' successors, Elagabalus (reigned AD 218–222). In addition, Severus issued a specific ban against Jewish proselytism.

With the conversion of Constantine I to Christianity early in the 4th century, a new era of prosperity began for Palestine. The emperor himself built a magnificent church on the site of the Holy Sepulchre, the most sacred of Christian holy places; his mother, Helena, built two others—at the place of the Nativity at Bethlehem and of the Ascension in Jerusalem—and his mother-in-law, Eutropia, built a church at Mamre. Palestine began to attract floods of pilgrims from all parts of the empire. It also became a great centre of the eremitic life; men flocked from all quarters to become hermits in the Judaean wilderness, which was soon dotted with monasteries. Constantine added the southern half of Arabia to the province, but in 357–358 (or perhaps as late as the 390s) the addition was made a separate province under the name of Palaestina or Salutaris (later Palaestina Tertia). At the end of the 4th century, an enlarged Palestine was divided into three provinces: Prima, with its capital at Caesarea; Secunda, with its capital at Scythopolis (Bet She'an); and Salutaris, with its capital at Petra or possibly for a time at Elusa. It is clear that the province of Palaestina underwent several territorial changes in the 4th century AD, but the details and the chronology remain obscure. The governor of Prima bore the high rank of proconsul from 382 to 385 and again from 535 onward. A dux of Palestine commanded the garrison of all three provinces.

Hermits

The bishop of the civil capital, Caesarea, was, according to the usual rule, metropolitan of the province, but the bishops of Jerusalem were claiming special prerogatives as early as the Council of Nicaea. Eventually, Juvenal, bishop of Jerusalem from 421 to 458, achieved his ambition and was recognized by the Council of Chalcedon (451) as patriarch of the three provinces of Palestine.

There was a revolt of the Jews in Galilee in 352, which was suppressed by the Caesar Gallus. Under Marcian (reigned 450–457) and again under Justinian I (reigned 527–565) the Samaritans revolted. Palestine, like Syria and Egypt, was also troubled by the Monophysite Controversy, a debate among Christians who disagreed with the Council of Chalcedon's assertion that the person of Jesus Christ comprised two natures, human and divine. When Juvenal returned from Chalcedon, having signed the Council's canons, the monks of Palestine rose and elected another bishop of Jerusalem, and military force was required to subdue them. Gradually, however, the Chalcedonian doctrine gained ground, and Palestine became a stronghold of orthodoxy. Apart from these disturbances the country enjoyed peace and prosperity until 611, when Khosrow II, king of Persia, launched an invasion. His troops captured Jerusalem (614), destroyed churches, and carried off the True Cross. In 628 the Byzantine emperor Heraclius recovered Palestine, and he subsequently restored the True Cross to Jerusalem, but 10 years later Jerusalem fell to the Arabs. (A.H.M.J./G.R.B.)

FROM THE ARAB CONQUEST TO AD 1900

The rise of Islām. The successful unification of the Arabian Peninsula by the first caliph, Abū Bakr (AD 632–Aug. 23, 634), made it possible to channel the expansion of the Arab Muslims into new directions. Abū Bakr, therefore, summoned the faithful to a holy war (jihad) and quickly amassed an impressive army. He dispatched three detachments of about 3,000 (later increased to about 7,500) men each to start operations in southern and southeastern Syria. He died, however, before he could witness the results of these undertakings. The conquests he started were carried on by his successor, the caliph 'Umar I (634–644).

The first battle took place at Wadi al-'Arabah, south of the Dead Sea. The Byzantine defenders were defeated and retreated toward Gaza but were overtaken and almost annihilated. In other places, however, the natural advantages of the defenders were more effective and the invaders

were hard pressed. Khālid ibn al-Walīd, then operating in southern Iraq, was ordered to the aid of his fellow generals on the Syrian front, and the combined forces won a bloody victory on July 30, 634, at a place in southern Palestine that the sources call Ajnadain. All of Palestine then lay open to the invaders.

Muslim conquest of Syria and Palestine

In the meantime, the emperor Heraclius was mustering a large army and in 636 dispatched it against the Muslims. Khālid concentrated his troops on the Yarmuk, the eastern tributary of the Jordan River. The decisive battle that delivered Palestine to the Muslims took place on Aug. 20, 636. Only Jerusalem and Caesarea held out, the former until 638, when it surrendered to the Muslims, and the latter until October 640. Palestine, and indeed all of Syria, was then in Muslim hands. After the surrender of Jerusalem, 'Umar divided Palestine into two administrative districts (*jund*), similar to the Roman and Byzantine provinces: they were Jordan (al-Urdunn) and Palestine (Filasṭīn). Jordan included Galilee and Acre and extended east to the desert; Palestine, with its capital first at Lydda and later at Ramlah (after 716), covered the region south of the Plain of Esdraelon.

'Umar lost no time in emphasizing Islām's interest in the holy city of Jerusalem as the first qibla toward which, until 623, Muslims had turned their faces in prayer and as the third holiest spot in Islām. (Muhammad himself changed the qibla to Mecca in 623.) On visiting the Temple area and finding the place suffering from neglect, 'Umar and his followers cleaned it with their own hands and declared it a sacred place of prayer.

Under the Umayyads, a Muslim dynasty that gained power in 661 from the Meccans and Medinans who had initially led the Islāmic community, Palestine formed, with Syria, one of the main provinces of the empire. Each *jund* was administered by an emir assisted by a financial officer. This pattern continued, in general, up to the time of Ottoman rule.

For various reasons, the Umayyads paid special attention to Palestine. The process of Arabization and Islāmization was gaining momentum there. It was one of the mainstays of Umayyad power and important in their struggle against both Iraq and the Arabian Peninsula. The caliph 'Abd al-Malik ibn Marwān (685–705) erected the Dome of the Rock in 691 on the site of the Temple of Solomon, which the Muslims believed had been the halting station of the Prophet on his nocturnal journey to heaven. This magnificent structure represents the earliest Muslim monument still extant. Close to the shrine to the south, 'Abd al-Malik's son, al-Walīd I (705–715), built the Aqṣa mosque. The Umayyad caliph 'Umar II (717–720) imposed humiliating restrictions on his non-Muslim subjects, particularly the Christians. Conversions arising from convenience as well as conviction then increased. These conversions to Islām, together with a steady tribal inflow from the desert, changed the religious character of Palestine's inhabitants. The predominantly Christian population gradually became predominantly Muslim and Arabic-speaking. At the same time, during the early years of Muslim control of the city, a small permanent Jewish population returned to Jerusalem after a 500-year absence.

Construction of the Dome of the Rock

'Abbāsid rule. Umayyad rule ended in 750. Along with Syria, Palestine became subject to 'Abbāsid authority, based in Baghdad, and, like Syria, it did not readily submit to its new masters. Unlike the Umayyads, who leaned on the Yemeni (South Arabian) tribes, the 'Abbāsids, in Syria, favoured and indeed used the Qays (North Arabian) tribes. Enmity between the two groups was, therefore, intensified and became an important political factor in Palestine. Pro-Umayyad uprisings were frequent and received Palestinian support. In 840/841 Abū Ḥarb, a Yemenite, unfurled the white banner of the Umayyads and succeeded in recruiting a large number of peasant followers, mainly among the Palestinian population, who regarded him as the saviour whose appearance was to save the land from the hated 'Abbāsids. Though the insurrection was put down, unrest persisted.

Under the 'Abbāsids, the process of Islāmization gained momentum. The 'Abbāsid rulers encouraged the settlement and fortification of coastal Palestine so as to secure it against the Byzantine enemy. During the second half of the 9th century, however, signs of internal decay began to appear in the 'Abbāsid empire. Petty states, and some indeed not so petty, emerged in different parts of the realm. One of the first to affect Palestine was the Tūlūnid dynasty (868–905) of Egypt, which marked the beginning of the disengagement of Egypt and, with it, of Syria and Palestine from 'Abbāsid rule. During this period Palestine also experienced the destructive operations of the Qarmaṭians, an Ismā'īlī Muslim sect that launched an insurrection in 903–906. After a brief restoration of 'Abbāsid authority, Palestine came under Ikhshīdid rule (935–969).

The Fāṭimid dynasty. In the meantime, the Shī'ite Fāṭimid dynasty was rising to power in North Africa. It moved eastward to seize not only Egypt but also Palestine and Syria and to threaten Baghdad itself. The Fāṭimids seized Egypt from the Ikhshīdids in 969 and in less than a decade were able to establish a precarious control over Palestine, where they faced Qarmaṭian, Seljuq, Byzantine, and periodic Bedouin opposition. Palestine was thus often reduced to a battlefield. The country suffered even greater hardship, however, under the Fāṭimid caliph al-Ḥākim (996–1021), whose behaviour was at times erratic and extremely harsh, particularly toward his non-Muslim subjects. He reactivated earlier discriminatory laws imposed upon Christians and Jews and added new ones. In 1009 he ordered the destruction of the Church of the Holy Sepulchre.

In 1071 the Seljuqs captured Jerusalem, which prospered as pilgrimages by Jews, Christians, and Muslims increased despite political instability. The Fāṭimids recaptured the city in 1098 only to relinquish it a year later to a new enemy, the crusaders of western Europe.

Saladin's wars against the crusaders

The Crusades. A year after the capture of Jerusalem by the crusaders, the Latin kingdom of Jerusalem was established (Christmas Day, 1100). Thereafter, there was no effective check to the expansion of the crusaders' power until the capture of their stronghold at Edessa (modern Urfa, Tur.) by the atabeg of Mosul, 'Imād ad-Dīn Zangi ibn Aq Sonqur, in 1144. Zangi's anticrusader campaign was carried on after his death by his son Nureddin (Nūr ad-Dīn Maḥmūd) and, more effectively, by the sultan Saladin, a protégé of the atabeg's family. After consolidating his position in Egypt and Syria, Saladin waged relentless war against the "infidel" Franks (Western Christians). On July 4, 1187, six days after the capture of Tiberias, he dealt the crusaders a crushing blow at the decisive battle of Hattin (Hittin). Most of Palestine was once again Muslim. Further attempts by the crusaders to regain control of Palestine proved ineffective, primarily because of incessant quarrels among the crusaders themselves. Ironically, it was left to an emperor of dubious Christian standing, Frederick II, to negotiate in 1229, while under excommunication, a 10-year treaty that temporarily restored Jerusalem, Nazareth, and Bethlehem to the Christians. In 1244, however, the Ayyūbid sultan aṣ-Ṣaliḥ Ayyūb definitively restored Jerusalem to Islām.

While the Ayyūbids of Saladin's house were losing ground to the Turkish-speaking Mamlūks in Egypt, the Mongol sweep westward continued, placing the crusaders, as it were, between two fires. To make matters worse, the crusaders themselves were hopelessly riddled with dissension. In 1260 the Mamlūk leader Baybars emerged as a champion of Muslim resurgence. After taking part in the defeat of the Mongols at 'Ayn Jalut in Palestine, he became sultan; in the years 1263 to 1271 he carried out annual raids against the harassed Franks. His efforts were continued by the sultan al-Ashraf, during whose reign the last of the crusaders were driven out of Acre (May 18, 1291). A chapter in the history of Palestine thus came to an end. The Mamlūks and subsequent Muslim regimes ruled the area with only brief interruptions for the next 600 years.

Palestine under the Mamlūks in the 14th century saw a period of prosperity for some; this was especially notable in Jerusalem, where the government sponsored an elaborate program of construction of schools, lodgings for travelers and Muslim pilgrims, and renovation of mosques. Tax revenues, collected mainly from the villages, were spent largely on support of religious institutions. Palestine

formed a part of the district of Damascus, second only to Egypt in the Mamlūk domains. The region suffered the ravages of several epidemics, including the great pestilence, the same Black Death that in 1347–51 devastated Europe. The fall of the Baḥrī Mamlūks and the rise of the Burjī Mamlūks (1382–1517) contributed to a gradual economic deterioration and a decrease in security. During the reign of the second Burjī sultan, Nasīr Faraj (1398–1405), the last onslaught of the Mongols, which made the name of Timur (Tamerlane) a synonym of destruction and plunder, took place. Although Palestine was spared the pillage of his hordes, it could not escape its disastrous repercussions as the Mamlūks moved through in a vain attempt to defend Damascus against the invader. The death of Timur in 1405, and the weakness of Iran in the ensuing century, pitted the Mamlūks against the rising power of the Ottoman Turks for the control of western Asia. Hostilities broke out in 1486 when Sultan Qait Bey contested with Bayezid II the possession of some border towns. The climax came three decades later on Aug. 24, 1516, when the Ottoman sultan, Selim I, routed the Mamlūk armies. Palestine began its four centuries under Ottoman domination.

Administrative divisions

Ottoman rule. Under the Ottoman Turks, Palestine continued to be linked administratively to Damascus until 1830, when it was placed under Sidon; then under Acre; then once again under Damascus until 1887–88, at which time the administrative divisions of the Ottoman Empire were settled for the last time. Palestine was divided into the *mutasarrifiyah*s of Nablus and Acre, both of which were linked with the vilayet of Beirut, and the autonomous *mutasarrifiyah* of Jerusalem, which dealt directly with Istanbul. With varying fortunes often accompanied by revolts, massacres, and wars, the first three centuries of Ottoman rule isolated Palestine from and insulated it against most outside influences. The prosperity of 16th-century Ottoman Palestine was followed by an economic and political decline in the 17th century. Ottoman control in the 18th century was indirect. Ḍāhir al-ʿUmar (*c.* 1737–75) dominated the political life of northern Palestine for nearly 40 years. Aḥmad al-Jazzār, the Ottoman governor of Acre, had control of most of Palestine, and in 1799, with English and Ottoman help, he successfully defended Acre against Napoleon.

Both Ḍāhir and al-Jazzār presided over a tightly controlled Palestine, where trade with Europe as well as taxation were growing. They used their new wealth from these sources to gain influence in Istanbul, which allowed them to gain local autonomy and even intermittent control of many areas outside Palestine.

This period came to an end with Napoleon's abortive attempt (1798–1801) to carve for himself a Middle Eastern empire. Egypt, always a determining factor in the fortunes of Palestine, was placed, after the French withdrawal, under the rule of the viceroy Muḥammad (Meḥmet) ʿAlī, who soon embarked upon a program of expansion at the expense of his Ottoman overlord. In 1831 his armies occupied Palestine, and for nine years he and his son Ibrāhīm gave it a centralizing and modernizing administration. Their rule increasingly opened the country to Western influences and enabled Christian missionaries to establish many schools; at the same time, however, taxes were increased, and urban rebellions broke out against the harshness of the regime. When in 1840 the British, the Austrians, and the Russians came to the aid of the Ottomans, the Egyptians were forced to withdraw and Palestine reverted to the Ottoman Empire. Increased European interest, however, led to the establishment of consulates by the powers in Jerusalem and in the ports.

After 1840 the reforms the sultan promulgated gradually took effect in Palestine. Increased security in the countryside and the Ottoman Land Law of 1858 encouraged the development of private property, agricultural production for the world market, the decline of tribal social organization, growth of the population, and the enrichment of the notable families. As the Ottomans extended the central government's new military, municipal, judicial, and educational systems to Palestine, the country also witnessed a marked increase in foreign settlements and colonies—French, Russian, and German. By far the most important, in spite of their initial numerical insignificance, were the Zionist agricultural settlements, which foreshadowed later Zionist endeavours for the establishment of a Jewish national home and still later a Jewish state in Palestine. The earliest of these settlements was established by Russian Jews in 1882. In 1896 Theodor Herzl issued a pamphlet entitled *Der Judenstaat* (*The Jewish State*) and advocated an autonomous Jewish state, preferably in Palestine. Two years later, he himself went to Palestine to investigate its possibilities and, possibly, to seek the help of the German emperor William II, who was then making his spectacular pilgrimage to the Holy Land. (N.A.F./Ed.)

Zionism

FROM 1900 TO 1948

In the last years of the 19th century and the early years of the 20th, the Palestinian Arabs shared in a general Arab renaissance. Palestinians found opportunities in the service of the Ottoman Empire, and Palestinian deputies sat in the Ottoman parliament of 1908. Several Arabic newspapers appeared in the country before 1914. The Arabs sought an end to Jewish immigration and to land purchases by Zionists. The number of Zionist colonies, however, rose from 19 in 1900 to 47 in 1918, even though the majority of the Jews were town dwellers. The population of Palestine, predominantly agricultural, was about 690,000 in 1914 (535,000 Muslims; 70,000 Christians, most of whom were Arabs; and 85,000 Jews).

World War I and after. During World War I the Great Powers made a number of decisions concerning the future of Palestine without much regard to the wishes of the indigenous inhabitants. Palestinian Arabs, however, believed that Great Britain had promised them independence in the Ḥusayn-McMahon correspondence, an exchange of letters from July 1915 to March 1916 between Sir Henry McMahon, British high commissioner in Egypt, and Ḥusayn ibn ʿAlī, then emir of Mecca, in which the British made certain commitments to the Arabs in return for their support against the Ottomans during the war. But in November 1917 Arthur Balfour, the British secretary of state for foreign affairs, addressed a letter to Lord Rothschild (the Balfour Declaration) expressing sympathy for the establishment in Palestine of a national home for the Jewish people on the understanding that "nothing shall be done which may prejudice the civil and religious rights of existing non-Jewish communities in Palestine."

The Balfour Declaration

Palestine was hard-hit by the war. In addition to the destruction caused by the fighting, the population was devastated by famine, epidemics, and Ottoman punitive measures against Arab nationalists. Major battles took place at Gaza before Jerusalem was captured by British and Allied forces under the command of General Sir Edmund (later 1st Viscount) Allenby in December 1917. The remaining area was occupied by the British by October 1918.

At the war's end, the future of Palestine was problematic. On March 20, 1920, delegates from Palestine attended a general Syrian congress at Damascus, which passed a resolution rejecting the Balfour Declaration and elected Fayṣal—son of Ḥusayn ibn ʿAlī, who ruled the Hejaz—king of a united Syria (including Palestine). In April 1920, however, at a peace conference held in San Remo, Italy, the Allies divided the former territories of the defeated Ottoman Empire. Of the Ottoman provinces in the Syrian region, the northern portion (Syria and Lebanon) was mandated to France, and the southern portion (Palestine) was mandated to Great Britain. By July 1920 the French had forced Fayṣal to give up his newly founded kingdom of Syria. The hope of founding an Arab Palestine within a federated Syrian state collapsed and with it any prospect of independence.

The French and British mandates

Uncertainty over the disposition of Palestine affected all its inhabitants and increased political tensions. In April 1920 anti-Zionist riots in the Jewish quarter of Old Jerusalem led to the death of 5 Jews and the wounding of more than 200; 4 Arabs lost their lives and 21 were injured. Following the confirmation of the mandate at San Remo, the British replaced the military administration with a civilian administration in July 1920, and Sir Herbert (later Viscount) Samuel, a Zionist, was appointed the first high

commissioner. The new administration proceeded with the implementation of the Balfour Declaration, announcing in August a quota of 16,500 Jewish immigrants for the first year.

In December 1920, Palestinian Arabs at a congress in Haifa established an executive committee (known as the Arab Executive) to act as the representative of the Arabs. It was never formally recognized and was dissolved in 1934. However, the platform of the Haifa congress, which set out the position that Palestine was an autonomous Arab entity and totally rejected any rights of the Jews to Palestine, remained the basic policy of the Palestinian Arabs until 1948. The arrival of more than 18,000 Jewish immigrants between 1919 and 1921 and land purchases in 1921 by the Jewish National Fund (established in 1901), which led to the eviction of Arab peasants (*fellahin*), further aroused Arab opposition. On May 1, 1921, anti-Zionist riots broke out in Jaffa, spreading to Petaḥ Tiqwa and other Jewish communities, in which 47 Jews and 48 Arabs were killed and 140 Jews and 73 Arabs wounded. Alarmed by the extent of Arab opposition, the British government issued a White Paper in June 1922 declaring that Great Britain did "not contemplate that Palestine as a whole should be converted into a Jewish National Home, but that such a Home should be founded in Palestine." Immigration would not exceed the economic absorptive capacity of the country, and steps would be taken to set up a legislative council. These proposals were rejected by the Arabs.

The White Paper of June 1922

The British mandate. In July 1922 the Council of the League of Nations approved the mandate instrument for Palestine, including its preamble incorporating the Balfour Declaration and stressing the Jewish historical connection with Palestine. Although Transjordan—*i.e.*, the lands east of the Jordan River—constituted three-fourths of the British mandate of Palestine, it was, despite protests from the Zionists, excluded from the clauses covering the establishment of a Jewish national home. On Sept. 29, 1923, the mandate officially came into force.

Palestine was a distinct political entity for the first time in centuries. This created problems and challenges for Palestinian Arabs and Zionists alike. Both communities realized that by the end of the mandate period the region's future would be determined by size of population and ownership of land. Thus the central issues throughout the mandate period were Jewish immigration and land purchases, with the Jews attempting to increase both and the Arabs seeking to slow down or halt both. Conflict over these issues often escalated into violence, and the British were forced to take action—a lesson not lost on either side.

Central issues during the mandate

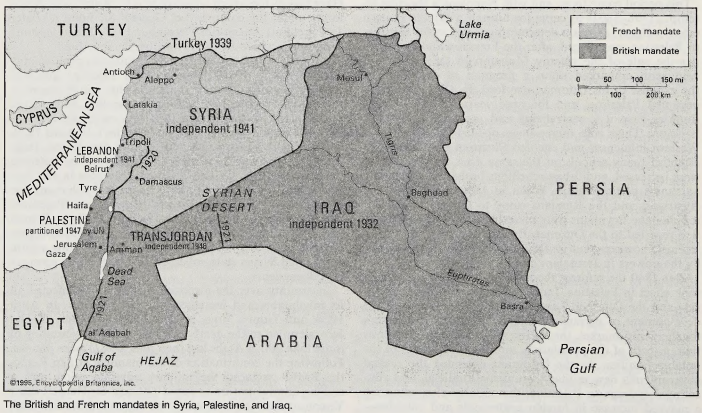
Several Arab organizations in the 1920s opposed Jewish immigration, including the Palestine Arab Congress, Muslim-Christian associations, and the Arab Executive. Most Arab groups were led by the strongly anti-British al-Ḥusaynī family, while the National Defense Party (founded 1934) was under the control of the more accommodating an-Nashashibi family. In 1921 the British high commissioner appointed Amīn al-Ḥusaynī to be the (grand) mufti of Jerusalem and made him president of the newly formed Supreme Muslim Council, which controlled the Muslim courts and schools and a considerable portion of the funds raised by religious charitable endowments.

Amīn al-Ḥusaynī

Initially, the Jews of Palestine thought it best served their interests to cooperate with the British administration. The World Zionist Organization (founded 1897) was regarded as the de facto Jewish Agency stipulated in the mandate, although its president, Chaim Weizmann, remained in London, close to the British government; David Ben-Gurion became the leader of a standing executive in Palestine. Throughout the 1920s most British local authorities in Palestine, especially the military, sympathized with the Palestinian Arabs, while the British government in London tended to side with the Zionists. The Jewish community in Palestine, the Yishuv, formed a military organization called the Haganah. Although the Jewish Agency was controlled by Labour Zionists who, for the most part, believed in cooperation with the British and Arabs, the Revisionist Zionists, founded in 1925 and led by Vladimir Jabotinsky, fully realized that their goal of a Jewish state in all of Palestine (*i.e.*, both sides of the Jordan River) was inconsistent with that of Palestinian Arabs. They formed their own military arm, Irgun Zvai Leumi, which did not hesitate to use force against the Arabs.

The years from 1923 to 1929 were relatively quiet; Arab passivity was partly due to the drop in Jewish immigration in 1926–28. In 1927 the number of Jewish emigrants exceeded that of immigrants, and in 1928 there was a net Jewish immigration of only 10 persons.

Nevertheless, the Jewish national home continued to consolidate itself in terms of urban, agricultural, social, cultural, and industrial development. Large amounts of land were purchased from Arab owners, who often were absentee landlords. In August 1929 a dispute in Jerusalem concerning religious practices at the Western Wall—sacred to Jews as the only remnant of the Second Temple of Jerusalem and to Muslims as the site of the Dome of the Rock—flared up into communal clashes in Jerusalem, Zefat, and Hebron, in which 133 Jews were killed and 339 wounded, the Arab casualties, mostly at the hands of British security forces, being 116 killed and 232 wounded. A royal commission of inquiry under the aegis of Sir



The British and French mandates in Syria, Palestine, and Iraq.

Walter Shaw attributed the clashes to the fact that "the Arabs have come to see in Jewish immigration not only a menace to their livelihood but a possible overlord of the future." A second royal commission, headed by Sir John Hope Simpson, issued a report stating that there was at that time no margin of land available for agricultural settlement by new immigrants. These two reports formed the basis of the Passfield White Paper, issued on Oct. 20, 1930, which accorded some priority to Britain's obligations to the Arabs. Not only did it call for a halt to Jewish immigration, but it also recommended that land be sold only to landless Arabs and that the determination of "economic absorptive capacity" be based on levels of Arab as well as Jewish unemployment. In response to protests from Palestinian Jews and London Zionists, the British prime minister, Ramsay MacDonald, in February 1931 addressed an explanatory letter to Chaim Weizmann nullifying the Passfield White Paper, which virtually meant a return to the policy of the 1922 White Paper.

The Passfield White Paper

From the early 1930s onward, developments in Europe once again began to impose themselves more forcefully on Palestine. The Nazi accession to power in Germany in 1933 and the widespread persecution of Jews throughout central and eastern Europe gave a great impetus to Jewish immigration, which jumped to 30,000 in 1933, 42,000 in 1934, and 61,000 in 1935. By 1936 the Jewish population of Palestine had reached almost 400,000, or 30 percent of the total. This new wave of immigration provoked major acts of violence against Jews and the British in 1933 and 1935. The Arab population of Palestine also grew rapidly, largely by natural increase. Most of the Arabs (nearly 90 percent) continued to be employed in agriculture. By the mid-1930s, however, many landless Arabs had joined the expanding Arab proletariat working in the construction trades on the edge of rapidly growing Jewish urban centres. This was the beginning of a shift in the foundations of Palestinian economic and social life that was to have profound immediate and long-term effects. In November 1935 the Arab political parties collectively demanded the cessation of Jewish immigration, the prohibition of land transfer, and the establishment of democratic institutions. A boycott of Zionist and British goods was proclaimed. In December the British administration offered to set up a legislative council of 28 members, in which the Arabs (both Muslim and Christian) would have a majority, but London rejected the proposal. This, together with the example of rising nationalism in neighbouring Egypt and Syria, increasing unemployment in Palestine, and a poor citrus harvest, touched off a long-smoldering Arab rebellion.

The Arab Revolt. The Arab Revolt of 1936–39 was the first sustained violent uprising of Palestinian Arabs for more than a century. Thousands of Arabs from all classes were mobilized, and nationalistic sentiment was fanned in the Arabic press, schools, and literary circles. The British, taken aback by the extent and intensity of the revolt, shipped more than 20,000 troops into Palestine, and by 1939 the Zionists had armed more than 15,000 Jews in their own nationalist movement.

The revolt began with spontaneous acts of violence committed by the religiously and nationalistically motivated followers of Sheikh ʿIzz ad-Dīn al-Qassām, who had been killed by the British in 1935. In April 1936 the murder of two Jews led to escalating violence, and Qassamite groups initiated a general strike in Jaffa and Nablus. At this point the Arab political parties formed an Arab High Committee presided over by the mufti of Jerusalem, Amīn al-Ḥusaynī. It called for a general strike, nonpayment of taxes, and the shutting down of municipal governments, although government employees were allowed to stay at work, and demanded an end to Jewish immigration, a ban on land sales to Jews, and national independence. Simultaneously with the strike, Arab rebels, joined by volunteers from neighbouring Arab countries, took to the hills, attacking Jewish settlements and British installations in the northern part of the country. By the end of the year the movement had assumed the dimensions of a national revolt, the mainstay of which was the Arab peasantry. The strike was called off in October 1939; however, even though the arrival of British troops restored some semblance of order, the armed rebellion, arson, bombings, and assassinations continued.

A royal commission of inquiry presided over by Lord Robert Peel, which was sent to investigate the volatile situation, reported in July 1937 that the revolt was caused by Arab desire for independence and fear of the Jewish national home. It declared the mandate unworkable and Britain's obligations to Arabs and Jews mutually irreconcilable. In the face of what it described as "right against right," the commission recommended the partition of the country. The Arabs were horrified by the idea of the dismemberment of the region and particularly by the suggestion of their forcible transfer (to Transjordan). As a result, the momentum of the revolt increased during 1937 and 1938. It continued well into 1939, but high casualty rates and firm British measures gradually eroded its strength.

The Peel Commission

However, the prospect of war in Europe alarmed the British government and caused it to reassess its policy in Palestine. If Britain went to war, it could not afford to face Arab hostility in Palestine and in neighbouring countries. On May 17, 1939, the British government issued a White Paper, which essentially yielded to Arab demands. It stated that the Jewish national home should be established within an independent Palestinian state. During the next five years 75,000 Jews would be allowed into the country; thereafter, Jewish immigration would be subject to Arab "acquiescence." The Zionists were shocked and enraged by the paper, which they considered a death blow to their program and to Jews who desperately sought refuge in Palestine from the growing persecution they were enduring in Europe. The 1939 White Paper marked the end of the Anglo-Zionist entente.

World War II. With the outbreak of World War II in September 1939 Zionist and British policies came into direct conflict. Throughout the war Zionists sought with growing urgency to increase Jewish immigration to Palestine, while the British sought to prevent such immigration, regarding it as illegal and a threat to the stability of a region essential to the war effort. British measures led to the disastrous sinking of two ships carrying Jewish refugees, the *Patria* (November 1940) and the *Struma* (February 1942). In response, the Irgun, under the leadership of Menachem Begin, and a small terrorist splinter group, LEHI (Fighters for the Freedom of Israel), known for its founder as the Stern Gang, embarked on widespread attacks on the British, culminating in the murder of Lord Moyne, British minister of state, by two LEHI members in Cairo in November 1944. Nonetheless, the moderate wing of the Jewish community supported the British; in September 1944 a Jewish brigade was formed—a total of 27,000 Jews having enlisted in the British forces—and attached to the British 8th Army.

The Arabs of Palestine remained largely quiescent throughout the war. Amīn al-Ḥusaynī had fled—by way of Iraq, Iran, Turkey, and Italy—to Germany, whence he broadcast appeals to his fellow Arabs to ally with the Axis powers against Britain and Zionism. Yet the mufti failed to rally Palestinian Arabs to the Axis cause. Although some supported Germany, the majority supported the Allies, and approximately 23,000 Arabs enlisted in the British forces (especially in the Arab Legion).

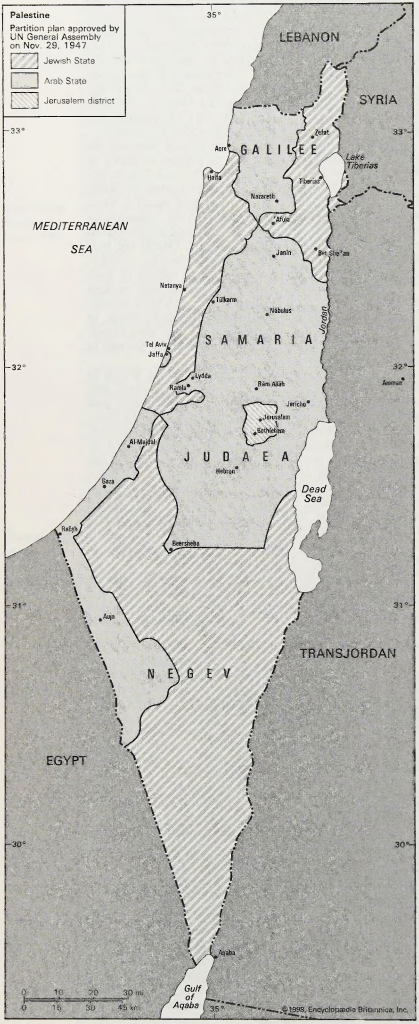
The discovery of the Nazi death camps at the end of World War II and the undecided future of Holocaust survivors led to an increasing number of pro-Zionist statements from U.S. politicians. The question of Palestine, now linked with the fate of Holocaust survivors, became once again the focus of international attention.

As the war came to an end, the neighbouring Arab countries began to take a more direct interest in Palestine. The covenant of the League of Arab States, or Arab League, formed in March 1945, contained an annex emphasizing the Arab character of Palestine. The Arab League appointed an Arab Higher Executive for Palestine (the Arab Higher Committee), which included a broad spectrum of Palestinian leaders, to speak for the Palestinian Arabs. In December 1945 the league declared a boycott of Zionist goods. The pattern of the postwar struggle for Palestine was unmistakably emerging.

The Arab League

Post-World War II. The major issue between 1945 and

1948 was, as it had been throughout the mandate, Jewish immigration to Palestine. The Yishuv was determined to remove all restrictions to Jewish immigration and to establish a Jewish state. The Arabs were determined that no more Jews should arrive and that Palestine should achieve independence as an Arab state. The primary goal



United Nations partition plan for Palestine adopted in 1947.

of British policy following World War II was to secure British strategic interests in the Middle East and Asia. Because the cooperation of the Arab states was considered essential to this goal, British Foreign Secretary Ernest Bevin opposed Jewish immigration and the foundation of an independent Jewish state in Palestine.

While Zionists pressed ahead with immigration and attacks on the government, and Arab states mobilized in response, British resolve to remain in the Middle East was collapsing. World War II had left Britain victorious but exhausted. After the war it lacked the funds and political will to maintain control of colonial possessions that were agitating for independence. When a conference called in London in February 1947 failed to resolve the impasse, Great Britain referred the Palestine question to the United Nations. On August 31 a majority report of the UN Special Committee on Palestine (UNSCOP) recommended the partition of the country into an Arab and a Jewish state, which, however, should retain an economic union. Jerusalem and its environs were to be international. These recommendations were substantially adopted by a two-thirds majority of the UN General Assembly in a resolution dated Nov. 29, 1947, a decision made possible partly because of the agreement of the United States and the Soviet Union on partition and partly because of pressure on some small countries by Zionist sympathizers in the United States. All the Islāmic Asian countries voted against partition, and an Arab proposal to query the International Court of Justice on the competence of the General Assembly to partition a country against the wishes of the majority of its inhabitants (in 1946 there were 1,269,000 Arabs and 678,000 Jews in Palestine) was narrowly defeated.

UN partition resolution

The Zionists welcomed the partition proposal both because it recognized a Jewish state and because it allotted 55 percent of (west-of-Jordan) Palestine to it. As in 1937, the Arabs fiercely opposed partition both in principle and because a substantial minority of the population of the Jewish state would be Arab. Great Britain was unwilling to implement a policy that was not acceptable to both sides and refused to share with the UN Palestine Commission the administration of Palestine during the transitional period. It set May 15, 1948, as the date for ending the mandate.

Civil war in Palestine. Soon after the UN resolution, fighting broke out in Palestine. The Zionists mobilized their forces and redoubled their efforts to bring in immigrants. In December 1947 the Arab League pledged its support to the Palestinian Arabs and organized a force of 3,000 volunteers. Civil war spread and external intervention increased as the disintegration of the British administration progressed.

The Zionists made a major effort to establish their state. They launched two offensives during April. The success of these operations coincided roughly with the failure of an Arab attack on the Zionist settlement of Mishmar Haemek, the death in battle of an Arab national hero, ʿAbd al-Qādir al-Ḥusaynī, in command of the Jerusalem front, and the massacre, by Irgunists and members of the Stern Gang, of civilian inhabitants of the Arab village of Deir Yāsīn. On April 22 Haifa fell to the Zionists, and Jaffa, after severe mortar shelling, surrendered to them on May 13. Simultaneously with their military offensives, the Zionists launched a campaign of psychological warfare. The Arabs of Palestine, divided, badly led, and reliant on the regular armies of the Arab states, became demoralized, and their efforts to prevent partition collapsed.

On May 14 the last British high commissioner, General Sir Alan Cunningham, left Palestine. On the same day the State of Israel was declared. Early on May 15 units of the regular armies of Syria, Transjordan, Iraq, and Egypt crossed the frontiers of Palestine.

Declaration of the State of Israel

In a series of campaigns alternating with truces between May and December 1948, the Arab units were routed. By the summer of 1949 Israel had concluded armistices with its neighbours. It had also been recognized by more than 50 governments throughout the world, joined the United Nations, and established its sovereignty over about 8,000 square miles (21,000 square kilometres) of formerly man-

dated Palestine west of the Jordan River. The remaining 2,000 square miles were divided between Transjordan and Egypt. Transjordan retained the lands on the west bank of the Jordan River, including East Jerusalem, although its annexation of those lands in 1950 was not generally recognized as legitimate. In 1949 the name of the expanded country was changed to the Hashemite Kingdom of Jordan. Egypt retained control of, but did not annex, a small area on the Mediterranean coast that became known as the Gaza Strip. The Palestinian Arab community ceased to exist as a social and political entity. (W.A.K./I.J.B.)

PALESTINE AND THE PALESTINIANS (1948–67)

The partition of Palestine and its aftermath. If one chief theme in the post-1948 pattern was embattled Israel (for greater detail on the history of Israel, see ISRAEL) and a second the unremitting hostility of its Arab neighbours, a third was the plight of the huge number of Arab refugees. The violent birth of Israel led to a major displacement of the Arab population. Many wealthy merchants and leading urban notables from Jaffa, Tel Aviv, Haifa, and Jerusalem, disproportionately Christian, fled to Lebanon, Egypt, and Jordan, while the middle class tended to move to all-Arab towns such as Nablus and Nazareth. The majority of peasants ended up in refugee camps. More than 350 Arab villages disappeared, and Arab life in the coastal cities (especially Jaffa and Haifa) virtually disintegrated. The centre of Palestinian life shifted to the Arab towns of the hilly eastern region later called the West Bank.

Like everything else in the Arab-Israeli conflict, population figures are hotly disputed. About 1,300,000 Arabs lived in Palestine before the war. Estimates of the number of Arabs displaced from their original homes, villages, and neighbourhoods during the period from December 1947 to January 1949 range from about 520,000 to about 1,000,000. Some 276,000 moved to the West Bank; by 1949 more than half the prewar Arab population of Palestine lived in the West Bank (from 400,000 in 1947 to more than 700,000). Between 160,000 and 190,000 fled to the Gaza Strip. More than 20 percent of Palestinian Arabs left Palestine altogether. About 100,000 of these went to Lebanon, 100,000 to Jordan, between 75,000 and 90,000 to Syria, 7,000 to 10,000 to Egypt, and 4,000 to Iraq.

The term "Palestinian." Henceforth the term Palestinian will be used when referring to the Arabs of the former mandated Palestine, excluding Israel. Although the Arabs of Palestine had been creating and developing a Palestinian identity for about 200 years, the idea that Palestinians form a distinct people is relatively recent. The Arabs living in Palestine had never had a separate state. Until the establishment of Israel, the term Palestinian was used by Jews and foreigners to describe the inhabitants of Palestine, but it was rarely used by the Arabs themselves; mostly they saw themselves as part of the larger Arab or Muslim community. But after 1948—and even more so after 1967—for Palestinians themselves the term came to signify not only a place of origin but, more importantly, a sense of a shared past and future. The Arabs of Palestine, and then of the West Bank and Gaza only, began widely using the term Palestinian to indicate the nationalist concept of a Palestinian people and, after 1967, of a Palestinian state.

Diverging histories for Palestinian Arabs. *The Israeli Arabs.* Approximately 150,000 Arabs remained in Israel. The Israeli Arabs represented about one-eighth of all Palestinians and by 1952 roughly the same proportion of the Israeli population. The majority of them lived in villages in western Galilee. Because much of their land was confiscated, Arabs were forced to abandon agriculture and become unskilled wage labourers, working in Jewish industries and construction companies. As citizens of the State of Israel, in theory they were guaranteed equal religious and civil rights with Jews. In reality, however, until 1966 they lived under a military jurisdiction that imposed severe restrictions on their political options and freedom of movement. Most of them remained politically quiescent, and some accepted Zionist Israel as a reality and sought to ameliorate their circumstances through electoral participation, education, and economic integration. For almost 20 years after Israel was established, Israeli Arabs were isolated from other Arabs.

West Bank (and Jordanian) Palestinians. The Jordanian monarchy saw in the events of 1948–49 the opportunity to expand Jordanian territory and to integrate Palestinians into its population and thereby create a new Jordanian nationality. It provided education and, in 1949, extended citizenship to Palestinians; indeed, two-thirds of all Palestinians became Jordanian citizens. However, tensions soon developed between original Jordanian citizens and the better-educated, more skilled newcomers. Wealthy Palestinians lived in the towns of the East and West banks, competing for positions within the government, while the peasants, *fellahin,* filled the UN refugee camps.

Palestinians constituted about two-thirds of the population of Jordan. Half the seats in the Jordanian Chamber of Deputies were reserved for representatives from the West Bank, but this measure and similar attempts to integrate the West Bank with the area lying east of the Jordan River were made difficult by the significant social, economic, educational, and political differences between the residents of each (for greater detail on the history of Jordan, see JORDAN). Jordanian Palestinians, other than the notable families favoured by the Jordanian monarchy, tended to support the radical pan-Arab and anti-Israeli policies of Egypt's President Gamal Abdel Nasser rather than the more cautious and conciliatory position of Jordan's King Hussein.

Palestinians in the Gaza Strip. During the 20 years the Gaza Strip was under Egyptian control (1948–67), it remained little more than a reservation. Egyptian rule was brutal and repressive. Palestinians living in the region were denied citizenship, which rendered them stateless (*i.e.,* it left them without citizenship of any nation), and they were allowed little real control over local administration. They were, however, allowed to attend Egyptian universities.

The Gaza Strip, 25 miles long and 4–5 miles wide, became one of the most densely populated areas of the world, with more than 80 percent of the population urban. Poverty and social misery became characteristic of life in the region. The rate of unemployment was high; many of the Palestinians lived in refugee camps, depending primarily on UN aid (see below). Most of the agricultural lands they had formerly worked were now inaccessible, and little or no industry was allowed, but commerce flourished, as Gaza became a kind of duty-free port for Egyptians. Although some Gaza Strip Palestinians were able to leave the territory and gain an education and find employment elsewhere, most had no alternative but to stay in the area. Economic conditions

The UNRWA camps. In December 1949 the UN General Assembly created the United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA) to assist the Palestinian refugees. In May 1950 UNRWA established a total of 53 refugee "camps" in the West and East banks of Jordan, the Gaza Strip, Lebanon, and Syria to assist the 650,000 or more Arab refugees it calculated needed help. Initially refugees in the improvised camps lived in tents, but after 1958 these were replaced by small houses of concrete blocks with iron roofs. Conditions were extremely harsh, and poverty and misery were widespread. Work was scarce, even though UNRWA sought to integrate the Palestinians into the depressed economies of the "host" countries.

Palestinians who continued to live in refugee camps felt a greater sense of alienation and dislocation than the more fortunate ones who found jobs and housing and became integrated into the national economies of the countries in which they resided.

Palestinians outside mandated Palestine. Palestinians found employment in Syria, Iraq, Lebanon, and the Persian Gulf states, but only a few were able to become citizens of these countries. Often they were the victims of discrimination, as well as being closely supervised by the respective governments intent on limiting their political activities.

Emergence of a Palestinian identity. The catastrophe of 1948 (also called *an-nakba*) and the experience of exile shaped Palestinian political and cultural activity for the next generation. The central task of reconstruction fell to The new leadership

Palestinians living outside Israel—both in the West Bank and Gaza communities and in the new Palestinian communities outside the former British mandate. The new leaders came disproportionately from among those who had moved to various Middle Eastern states and to the West, even though four out of five Palestinians had remained within the borders of the former mandate. By the mid-1960s, despite Israeli efforts to forestall the emergence of a new Palestinian identity, a young, educated leadership had arisen, replacing the discredited traditional local and clan leaders.

The role of camps. Palestinian refugee camps differed depending on the country in which they were located, but they shared one common development—the emergence of a "diaspora consciousness." In time this consciousness grew into a new national identity and reinvigorated social institutions, leading to the establishment of more complex social and political structures by the 1960s. Palestinians living in the UNRWA-administered refugee camps felt isolated, politically powerless, disoriented, bitter, and resentful. They remained unassimilated and were generating a new sense of identity based on a Nasser-inspired pan-Arabism, the cultivated memory of a lost paradise (Palestine), and an emerging pan-Islāmic movement.

The role of Palestinians outside formerly mandated Palestine. By the late 1960s a class of educated and mobile Palestinians had emerged, with fewer than half of them living in the West Bank or Gaza. They were working in the oil companies, civil services, and educational institutions of most Arab states in the Middle East. Having successfully resisted the efforts of Jordan and Egypt to control and speak for them, they joined the process of reshaping Palestinian consciousness and institutions.

The Palestine Liberation Organization. An Arab summit meeting in Cairo in 1964 led to the formation of the Palestine Liberation Organization (PLO). A political umbrella organization of several Palestinian groups, the PLO thereafter consistently claimed to be the sole representative of all Palestinian people. Its first leader was Aḥmad Shuqayri, a protégé of Egypt. In its charter (the Palestine National Charter, or Covenant), the PLO delineated its basic principles and goals, the most important of which were the right to an independent state, the total liberation of Palestine, and the destruction of the State of Israel.

Palestine National Council

A Palestine National Council (PNC) was established to serve as the supreme body, or parliament, of the PLO, and an executive committee was formed to manage PLO activities. Initially, the PNC consisted of civilian representatives from various areas, including Jordan, the West Bank, and the Persian Gulf states, but representatives of guerrilla organizations were added in 1968.

Fatah and other guerrilla organizations. Several years before the creation of the PLO, a secret organization had been formed: the Palestine National Liberation Movement (Ḥarakāt at-Taḥrīr al-Waṭani al-Filasṭīnī), known from a reversal of its Arabic initials as Fatah. Both the PLO and Fatah undertook the training of guerrilla units for raids on Israel. In addition to Fatah, the largest and most influential guerrilla organization, several others emerged in the late 1960s. The most important ones were the Popular Front for the Liberation of Palestine (PFLP); the Popular Front for the Liberation of Palestine–General Command (PFLP–General Command, a splinter group from the PFLP); aṣ-Ṣa'iqah, backed by Syria; the Democratic Front for the Liberation of Palestine (DFLP); and the Popular Struggle Front (PSF). In 1969 Yāsir 'Arafāt, leader of Fatah, became chairman of the PLO's executive committee and thus the chief of the Palestinian national movement.

Despite their differences in tactics and ideology, the guerrilla organizations were united in rejecting any political settlement that did not include what they characterized as the total liberation of Palestine and the return of the refugees to their homeland, goals that were to be achieved through armed struggle.

THE ARAB-ISRAELI WAR OF 1967 AND ITS CONSEQUENCES

In the Arab-Israeli war of 1967, Israel dispatched the combined forces of Egypt, Syria, and Jordan and also overran large tracts of territory, including East Jerusalem, the West Bank (known to Israelis as Judaea and Samaria), and Gaza. Israel's victory gave rise to another exodus of Palestinians, with more than 250,000 people fleeing to the East Bank of Jordan. However, roughly 600,000 Palestinians remained in the West Bank and 300,000 in Gaza. Thus, the 3,000,000 Israeli Jews came to rule some 1,200,000 Arabs (including the 300,000 already living in the State of Israel).

The PLO's rise as a revolutionary force. Throughout the 1970s and '80s the PLO, dominated by Fatah, acted as a state in the making, launching frequent military attacks on Israel. The strategy of armed struggle had been inaugurated as early as the 1950s, and, indeed, guerrilla raids against military and civilian targets in Israel had been a factor leading to the Suez Crisis of 1956 and the 1967 war. However, after 1967 Fatah launched a new wave of violent activities, and the Palestinian guerrilla groups, under the umbrella of the PLO, emerged as an element of major importance in the Middle East. The Arab-Israeli war had discredited Nasser's pan-Arabism, and Fatah quickly permeated and mobilized the reunited Palestinian population, providing social services and organizations. One result was an escalating cycle of raids and reprisals between the Palestinian guerrillas and Israel; terrorism (unacknowledged by the PLO until renounced by 'Arafāt in 1988) became a key element in the struggle against Israel.

The PLO's struggle for Palestinian autonomy. Although generally recognized as the symbol of the Palestinian national movement, the PLO lacked organizational cohesion. It championed Palestinian distinctiveness and autonomy, but circumstances forced the leadership to remain somewhat distant from the lives of Palestinians. This made it difficult to impose a unified policy.

Palestinians and the PLO in Jordan. Within Jordan an understanding was reached whereby the royal government allowed the Palestinian guerrillas independent control of their own bases in the Jordan Valley, but relations between the government and the guerrillas remained uneasy. The guerrillas suspected King Hussein—who had to face the damage inflicted upon his country by Israeli reprisals—of preparing to reach a direct settlement with Israel at their expense.

"Black September"

Tensions between the Jordanian army loyal to King Hussein and the Palestinian guerrillas erupted in a brief but bloody civil war in September 1970 that became known as "Black September." By September 17, the army was fighting the guerrillas in Amman and in northern Jordan, where the guerrillas were reinforced by Syrian armoured units. Hostilities formally ended on September 25, with the guerrillas still in control of their northern strongholds. In 1971, however, the fighting resumed, and the Jordanian army established full control over the whole country and crushed the Palestinian military. From that point on, the king and the PLO were deeply suspicious of each other, and the majority of Palestinians holding Jordanian citizenship, especially those living on the West Bank under Israeli control, became highly critical of the king and his policies.

Palestinians and the PLO in Lebanon. Driven from Jordan, the PLO intensified its activities in Lebanon. The presence of more than 235,000 Palestinians was a source of tension and conflict in Lebanon, as it had been in Jordan. Palestinian guerrillas continued to carry out attacks against Israel. Israel countered with raids, primarily into southern Lebanon. The Lebanese government sought to restrict the activities of the Palestinian guerrillas, which led to sporadic fighting between the Lebanese army and the guerrillas.

Israeli attack on Beirut

In 1973 the Palestinian movement suffered a severe blow from an Israeli commando attack in Beirut on April 10 that killed three of its leaders. After 1973 many Palestinians from Lebanon and Jordan moved to the increasingly prosperous and labour-hungry Arab oil states, such as Kuwait and Saudi Arabia.

International recognition. In its international relations the PLO made important gains in the 1970s. By the end of the decade the organization had representatives in more than 80 countries. On Sept. 22, 1974, the UN General Assembly, overriding strong Israeli objections, included

on its agenda for the first time the "Palestine question" as a separate subject for debate rather than as part of the general question of the Middle East. On November 13 the assembly heard ʿArafāt plead for the Palestinian people's national rights.

International recognition of the PLO had important repercussions within the Arab camp. At an Arab summit conference held in Rabat, Morocco, on Oct. 26–28, 1974, King Hussein accepted a resolution stating that any "liberated" Palestinian territory "should revert to its legitimate Palestinian owners under the leadership of the PLO." The Rabat decision was denounced by the more radical "rejection front," composed of the PFLP, the PFLP–General Command, the pro-Iraq Arab Liberation Front, and the Front for the Popular Palestinian Struggle, which sought to regain all of Palestine. Although the decision was recognized as enhancing the less extreme position of the PLO elements led by ʿArafāt, the United States continued adamantly to refuse to recognize or deal with the PLO so long as the PLO refused to recognize Israel's right to exist.

Palestinians and the civil war in Lebanon. Palestinian guerrilla activity against Israel in 1975 was largely confined to the southern Lebanese border area, where it provoked heavy Israeli reprisals. Israeli raids, however, were overshadowed by the civil strife in Lebanon that developed along Muslim-Christian lines. Civil war between the militias of Lebanon's confessional groups erupted in 1975, bringing 15 years of bloodshed. The presence of Palestinians was a factor contributing to the civil war, and the war proved disastrous for them.

The PLO initially tried to stay out of the fighting, but by the end of 1975 groups within the overall organization, particularly the "rejection front" groups, were being drawn into an alliance with the Muslim side fighting against the Christians. In August 1976, after a two-month siege, Christian forces killed an estimated 2,000 to 3,000 Palestinians in Tall Zaʿtar camp northeast of Beirut. A peace agreement was negotiated in October 1976. The settlement provided for the creation of a 30,000-member Arab Deterrent Force (ADF), a cease-fire throughout the country and withdrawal of forces to positions held before April 1975, and implementation of a 1969 agreement limiting Palestinian guerrilla operations in Lebanon.

Although the Palestinian guerrillas suffered heavy losses in the Lebanese civil war, they continued to mount attacks against Israel in the late 1970s. On March 11, 1978, a Palestinian raid into Israel left 37 civilian tourists dead and 82 wounded; there followed the Israeli invasion of southern Lebanon three days later (Operation Litani). On March 19 the UN Security Council passed Resolution 425, calling for Israel's withdrawal and the establishment of the UN Interim Force in Lebanon (UNIFIL). The Israelis withdrew in stages and had evacuated Lebanese territory by mid-June. They had been only partly successful in their aim of destroying Palestinian guerrillas and their bases south of the Litani River. Several hundred Palestinian guerrillas had been killed, but most of them had escaped northward. Estimates of civilian casualties ranged from 1,000 to 2,000.

Operation Litani

The PLO in the occupied territories. As Israeli occupation of the West Bank solidified, many local Palestinian leaders turned to building social organizations, labour unions, and religious, educational, and political institutions. By the early 1980s the PLO had set up an extensive rival bureaucratic structure providing health, housing, educational, legal, media, and labour services for Palestinians both inside and outside the camps. Israeli and Jordanian attempts to encourage alternatives to the PLO failed in the 1980s. Active opposition to Israeli control in the West Bank spread, while frequent demonstrations, strikes, and other incidents occurred, particularly among students.

NEGOTIATIONS, VIOLENCE, AND INCIPIENT SELF-RULE

The late 1970s was a period of more active negotiation on Arab-Israeli disputes. In March 1977 U.S. President Jimmy Carter spoke of the need for a Palestinian homeland, and he later stated that it was essential for the Palestinians to take part in the peace process.

The Camp David Accords and the PLO. In November 1977 the Egyptian president, Anwar el-Sādāt, initiated peace negotiations that led to the Camp David Accords in September 1978 and the Egyptian-Israeli peace treaty signed on March 26, 1979. Provisions of the accords (named for Camp David, Md., U.S., where they were negotiated) included the establishment of a self-governing authority in the West Bank and the Gaza Strip and a transitional period of not more than five years, at the end of which the inhabitants would become autonomous. The Soviet Union during the time of the peace negotiations recognized the PLO as the sole legitimate representative of the Palestinians and in 1981 extended formal diplomatic recognition. The nations of western Europe announced their support of PLO participation in peace negotiations in June 1980. The PLO continued to seek diplomatic recognition from the United States, but the Carter administration honoured a secret commitment made by the former U.S. secretary of state Henry Kissinger to Israel not to deal with the PLO so long as it declined to renounce terrorism and to recognize Israel's right to exist.

The dispersal of the PLO from Lebanon. The Likud government of Israel viewed the possibility of peace and compromise with suspicion. On June 6, 1982, Israel, in an attempt to end attacks on its territory, dislodge the PLO, and encourage the installation of a Lebanese government friendly to Israel, launched an invasion of Lebanon. PLO and Syrian forces were defeated by Israeli troops, and the remaining PLO forces were contained in West Beirut. After a lengthy siege and bombardment by Israel in late August 1982, some 11,000 Palestinian fighters were allowed to leave Beirut for various destinations, under international guarantees for their own safety and that of their civilian dependents. However, after Israeli troops had occupied West Beirut, the Phalangists, Israel's rightist Lebanese allies, despite these guarantees, massacred hundreds (variously estimated to be between 700 and 3,000) of Palestinian civilians in the Beirut refugee camps of Sabra and Shatila.

Israel's invasion of Lebanon

The dispersal of the PLO from Lebanon significantly weakened the organization's military strength and political militancy. It was unable to operate freely from any of the nations bordering Israel. ʿArafāt and the other PLO leaders were also threatened by the emergence within Fatah of a faction encouraged by Syria. In December 1983 ʿArafāt was driven out of northern Lebanon by the Syrians and their protégés inside the PLO.

After having established himself near Tunis, Tunisia, ʿArafāt turned to diplomatic initiatives. He sought Egyptian and Jordanian support against Syria. He also looked to King Hussein as an intermediary for negotiations with the United States and Israel that might lead to a Palestinian ministate on the West Bank within a Jordan-Palestine confederation—an idea that had been favoured by the dominant factions in the PLO since the early 1980s.

Years of mounting violence. Violence escalated from the mid-1980s onward. Rejectionist elements within the PLO renewed their activities, attracting worldwide attention. On Oct. 7, 1985, members of the Palestine Liberation Front, a small faction within the PLO headed by Abul Abbas, hijacked an Italian cruise ship, the *Achille Lauro,* and murdered one of its passengers.

In Israel itself several new developments strengthened Arab feelings of alienation from the Jewish majority. Ultranationalists increased their demands for more Jewish settlements and the annexation of the West Bank and advocated the forceful removal of the Palestinian Arabs. Israelis—especially West Bank settlers—became more antagonistic and militant toward Palestinians as armed attacks against Israel and Israelis living abroad increased in number. Throughout 1986–87 attacks by Arabs on Israeli settlers and by Israeli settlers on Arabs mounted. By 1988 more than half of the West Bank and about a third of the Gaza Strip had been transferred to Jewish control, and the Israeli population of the West Bank—mostly concentrated in 15 metropolitan satellites of Tel Aviv and Jerusalem—had reached about 100,000.

By the late 1980s a whole generation of Palestinian youth had grown up under Israeli occupation (more accurately described as military rule after 1978). More than 70 per-

cent of Palestinians were younger than 25 years of age. Their political status was uncertain, their civil rights diminished, and their economic status low and dependent upon Israel's economy. Between 100,000 and 120,000 Palestinians crossed daily from the occupied territories into Israel to work. Lacking faith in Arab governments and in the PLO, Palestinians increasingly came to rely on their own efforts. (I.J.B.)

The *intifada*. When the Palestinians saw no improvement in support from other countries and no favourable change coming from the Israelis, they engaged, throughout 1987, in small-scale demonstrations, riots, and occasional violence directed against Israelis. Israeli authorities responded with university closings, arrests, and deportations. Large-scale riots and demonstrations continued for more than five years thereafter. This uprising, which became known as the *intifada* ("shaking off"), inspired a new era in Palestinian mass mobilization. Demonstrators turned to throwing stones at Israeli troops, and the soldiers responded by shooting and arresting them. Women, and women's organizations, were prominent.

An Islāmic group, Ḥamās—an acronym for the Islāmic Resistance Movement—challenged the authority of the secular nationalist movement and sought to take over the leadership of the *intifada*. Ḥamās was a wing of the Sunnite Muslim Brotherhood, which had built up a network of charitable institutions in the occupied territories, and rejected any accommodation with Israel.

Tactics and tolls. The tactics of the *intifada* were highly sophisticated. Union-organized strikes, commercial boycotts and closures, and demonstrations were carried out in one part of the territories and then, after Israel had reestablished its local power there, were transferred to a previously quiescent area. Israel closed universities and schools, destroyed houses, and imposed curfews, yet it was unable to quell the uprising. By mid-1990 hundreds of Palestinians and dozens of Israelis had been killed in the violence. Thousands of Palestinians had been imprisoned by Israeli security forces, and hundreds more—branded as Israeli collaborators—had been killed by their compatriots. Political paralysis gripped Israel.

PLO declaration of independence. In 1988 ʿArafāt took the diplomatic initiative. At the 19th session of the PNC, held near Algiers, Algeria, in November, he succeeded in having the council issue a declaration of independence for a state of Palestine in the West Bank and Gaza Strip. ʿArafāt proclaimed the state (without defining its borders) on November 15. Within days more than 25 nations had extended recognition to the government-in-exile.

In December 1988 ʿArafāt announced that the PNC recognized Israel as a state and condemned terrorism in all its forms—including state terrorism, the PLO's term for Israel's actions. He addressed a special meeting of the UN General Assembly convened at Geneva and proposed an international peace conference under UN auspices. He publicly accepted UN resolutions 242 and 338, thereby recognizing, at least implicitly, the State of Israel. Despite their ambiguities, UN resolution 242 (1967), which encapsulated the principle of land for peace, and resolution 338 (1973), which called for direct negotiations, were regarded by both parties as the starting points for negotiations. Although Israeli Prime Minister Shamir stated that he was still not prepared to negotiate with the PLO, the U.S. government announced that it would open dialogue with the organization.

Moving toward self-rule. The approaching end of the Cold War left the Palestinians diplomatically isolated, as did PLO support for Iraq during the Persian Gulf War (1990–91). Funds from the Persian Gulf states dried up, and most Palestinians living in Kuwait were expelled. Moreover, ongoing violence led Israel to impose severe travel restrictions on Palestinian day labourers. Support for the PLO leadership in Tunis declined.

However, prospects for peace were altered by several factors: the convening of an international peace conference between Israeli and Arab delegates (including Palestinians from the occupied territories as part of a joint Jordanian-Palestinian delegation) at Madrid in October 1991; the dissolution of the Soviet Union in December; and the election of a Labour Party government in 1992 that was committed to quickly implementing Palestinian autonomy.

Although progress at the Madrid conference was discouraging, secret meetings in Norway between PLO and Israeli officials produced an understanding known as the Oslo Accords. On Sept. 13, 1993, the PLO and Israel signed a Declaration of Principles in Washington, D.C. It included mutual recognition and terms whereby governing functions in the West Bank and Gaza would be progressively handed over to a Palestinian Council for an interim period of five years, during which time Israel and the Palestinians would negotiate a permanent settlement on the most contentious issues—such as the status of Jerusalem, the return of refugees, the presence of Israeli settlements in the occupied territories, and the issue of Palestinian sovereignty and statehood. Declaration of Principles

The Israelis completed their withdrawal from the West Bank town of Jericho and parts of the Gaza Strip in May 1994. On July 1 ʿArafāt entered Gaza in triumph. Four days later he swore in members of the Palestinian Authority (PA) in Jericho. Despite continuing violence, ʿArafāt and Israeli Prime Minister Yitzhak Rabin and Foreign Minister Shimon Peres signed an agreement in Washington on Sept. 28, 1995, providing for expanding Palestinian self-rule in the West Bank and for elections of a chairman and a legislative council. The PA would gain control over the larger towns in the West Bank, and Israel would then gradually redeploy from some 440 villages, which would then come under Palestinian rule. Permanent-status negotiations were to be concluded by 1999.

In October 1995, right-wing Israelis stepped up their rhetoric against Rabin and the peace process, and in early November Rabin was assassinated by a Jewish extremist. Peres, Rabin's successor, quickly expressed his determination to continue the planned Israeli deployments. In Palestinian elections held in January 1996, ʿArafāt won the presidency, and Fatah won a majority in the legislative council. Ḥamās continued its violent opposition to the peace process, however, which stalled for nearly three years when the Likud Party's Benjamin Netanyahu was elected prime minister of Israel in May. (I.J.B./Ed.)

In May 1999 Netanyahu left office following defeat at the hands of the Labour Party led by Ehud Barak, who ran on a pro-peace platform. However, a number of negotiating deadlines passed without an agreement. Notable among these was the May 1999 date set as an end of the interim period and a completion of permanent-status talks on the most contentious issues.

Also, by that time only a small fraction of the occupied territories had reverted to full Palestinian control, the number of Israelis living in West Bank settlements had grown significantly, and the gross domestic product per capita in Palestinian areas had actually declined. Life for most residents of PA-controlled areas was worse, in many respects, than it had been before the peace process.

A summit meeting sponsored by the United States in July 2000 failed to resolve these issues and led to an increasingly strained situation; Likud leader Ariel Sharon's visit to Jerusalem's Al-Ḥaram al-Sharīf (the Temple Mount) in September triggered a renewed uprising. The new *intifada* soon reached a level of violence unseen in the first outbreak and engulfed the still largely occupied West Bank and Gaza Strip. Thereafter, suicide bomb attacks by Palestinians in Israeli cities reached an unprecedented number, and many Israelis, most of them civilians, were killed and wounded. Israeli attacks on PA targets (most located within population centres) raised the already high casualty rate among the Palestinian populace. In the spring of 2002 Israeli troops reoccupied all the towns and cities of the West Bank, reclaiming security control from the PA and tightening restrictions on movement. (Ed.)

BIBLIOGRAPHY. Geography is discussed in *Atlas of Israel: Cartography, Physical and Human Geography*, 3rd ed. (1985; originally published in Hebrew, 3rd ed., 1985); *Geography* (1973), compiled from material originally published in the *Encyclopaedia Judaica*; and W.B. FISHER, *The Middle East: A Physical, Social, and Regional Geography*, 7th ed. rev. (1978).

The dating of Palestine's early history is linked to Egyptian dating as interpreted by ROLF KRAUSS, *Sothis- und Monddaten*

(1985). Prehistory and early history are examined by JACQUES CAUVIN, *Les Premiers Villages de Syrie-Palestine du IXe au VIIe millénaire avant J.C.* (1978); LORENZO VIGANÒ, "Literary Sources for the History of Palestine and Syria: The Ebla Tablets," *Biblical Archaeologist,* 47(1):6–16 (March 1984); A.T. OLMSTEAD, *History of Palestine and Syria to the Macedonian Conquest* (1931, reprinted 1972); MICHAEL GRANT, *The History of Ancient Israel* (1984), a good general account; JOHN BRIGHT, *A History of Israel,* 3rd ed. (1981); KATHLEEN M. KENYON, *Archaeology in the Holy Land,* 4th ed. (1979, reprinted 1985); and F.E. PETERS, *Jerusalem: The Holy City in the Eyes of Chroniclers, Visitors, Pilgrims, and Prophets from the Days of Abraham to the Beginnings of Modern Times* (1985).

Hellenistic and Roman Palestine are the subject of VICTOR TCHERIKOVER, *Hellenistic Civilization and the Jews,* trans. from Hebrew (1959, reprinted 1975); M. ROSTOVTZEFF, *The Social & Economic History of the Hellenistic World,* 3 vol. (1941, reprinted 1986); and A.H.M. JONES, *The Herods of Judaea* (1938, reprinted 1967). The most valuable general modern work is EMIL SCHÜRER, *The History of the Jewish People in the Age of Jesus Christ (175 B.C.–A.D. 135),* rev. ed., trans. from German (1973–). Other useful works are E. MARY SMALLWOOD, *The Jews Under Roman Rule: From Pompey to Diocletian* (1976, reprinted 1981); and M. AVI-YONAH, *The Jews of Palestine: A Political History from the Bar Kokhba War to the Arab Conquest,* trans. from Hebrew (1976, reprinted as *The Jews Under Roman and Byzantine Rule,* 1984).

Studies of the periods of Muslim rule before 1516 include GUY LE STRANGE (trans.), *Palestine Under the Moslems: A Description of Syria and the Holy Land from A.D. 650 to 1500* (1890, reprinted 1975), a collection of medieval Arabic sources; SHLOMO D. GOITEIN, "Jerusalem in the Arab Period (638–1099)," *The Jerusalem Cathedra,* 2:168–196 (1982); and AMNON COHEN and GABRIEL BAER (eds.), *Egypt and Palestine: A Millennium of Association (868–1948)* (1984).

Among many recent works on Ottoman Palestine, the following are the broadest and most valuable: MOSHE MA'OZ (ed.), *Studies on Palestine During the Ottoman Period* (1975); AMNON COHEN, *Palestine in the 18th Century: Patterns of Government and Administration* (1973); DAVID KUSHNER (ed.), *Palestine in the Late Ottoman Period: Political, Social, and Economic Transformation* (1986); and NEVILLE J. MANDEL, *The Arabs and Zionism Before World War I* (1976, reprinted 1980).

Some useful general studies that cover the 20th century are BARUCH KIMMERLING and JOEL S. MIGDAL, *Palestinians: The Making of a People* (1993); FRED J. KHOURI, *The Arab-Israeli Dilemma,* 3rd ed. (1985); HOWARD M. SACHAR, *A History of Israel,* 2 vol. (1979–87); BRUCE R. KUNIHOLM and MICHAEL RUBNER, *The Palestinian Problem and United States Policy: A Guide to Issues and References* (1986), with an extensive bibliography; and IAN J. BICKERTON and CARLA L. KLAUSNER, *A Concise History of the Arab-Israeli Conflict,* 2nd ed. (1995).

The period of the British mandate is covered by TARIF KHALIDI, "Palestinian Historiography: 1900–1948," *Journal of Palestine Studies,* 10(3):59–76 (Spring 1981); SAMI HADAWI, *Bitter Harvest: A Modern History of Palestine,* 4th rev. and updated ed. (1991); ANN MOSELY LESCH, *Arab Politics in Palestine, 1917–1939: The Frustration of a Nationalist Movement* (1979); Y. PORATH, *The Emergence of the Palestinian-Arab National Movement, 1918–1929,* trans. from Hebrew (1974), and *The Palestinian Arab National Movement: From Riots to Rebellion, 1929–1939,* trans. from Hebrew (1977); KENNETH W. STEIN, *The Land Question in Palestine, 1917–1939* (1984); BENNY MORRIS, *The Birth of the Palestinian Refugee Problem, 1947–1949* (1987); CONOR CRUISE O'BRIEN, *The Siege: The Saga of Israel and Zionism* (1986); and WM. ROGER LOUIS and ROBERT W. STOOKEY (eds.), *The End of the Palestine Mandate* (1985).

Among the many studies on the Palestinian Arabs from 1948 to the present, the following are particularly worth consulting: HELENA COBBAN, *The Palestinian Liberation Organisation: People, Power, and Politics* (1984); F. ROBERT HUNTER, *The Palestinian Uprising: A War By Other Means,* rev. and expanded 2nd ed. (1993); ANN MOSELY LESCH, *Transition to Palestinian Self-Government* (1992); SHAUL MISHAL, *The PLO Under 'Arafat: Between Gun and Olive Branch* (1986); DON PERETZ, *The West Bank: History, Politics, Society, and Economy* (1986), and *Intifada: The Palestinian Uprising* (1990); JULIE M. PETERS, *Gender in Crisis: Women and the Palestinian Resistance Movement* (1991); EDWARD W. SAID and JEAN MOHR, *After the Last Sky: Palestinian Lives* (1986); ZE'EV SCHIFF and EHUD YA'ARI, *Intifada: The Palestinian Uprising—Israel's Third Front,* ed. and trans. by INA FRIEDMAN (1990; originally published in Hebrew, 1990); CHARLES D. SMITH, *Palestine and the Arab-Israeli Conflict,* 3rd ed. (1996); and ELIA T. ZUREIK, *The Palestinians in Israel: A Study in Internal Colonialism* (1979).

(G.R.B./I.J.B.)

Paraguay

The Republic of Paraguay (Spanish: República del Paraguay) is an independent nation of South America, located in the south-central part of the continent. A landlocked country, Paraguay is dwarfed by the larger bordering countries of Bolivia to the northwest and north; Brazil to the northeast and east; and Argentina to the southeast, south, and west. The country has a total area of some 157,048 square miles (406,752 square kilometres). The national capital is Asunción, which is located on the east bank of the Paraguay River opposite the mouth of its primary western tributary—the Pilcomayo River. Rivers play an extremely important role in the nation's economic life, providing the country with access to the distant Atlantic Ocean and with sites for hydroelectric power plants. Indeed, the name of the country is said to derive from the Guaraní Indian word meaning "river that gives birth to the sea."

The article is divided into the following sections:

Physical and human geography 426
- The land 426
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant and animal life
- Settlement patterns 426
  - Rural settlement
  - Urban settlement
- The people 426
  - Ethnic composition
  - Language groups
  - Religious groups
  - Demography
- The economy 428
  - Resources
  - Agriculture, forestry, and fishing
  - Industry
  - Finance
  - Trade
  - Transportation
- Government and social conditions 430
  - Government
  - The political process
  - The judicial system
  - Armed forces
  - Education
  - Health and welfare
- Cultural life 431
  - Daily life
  - The arts
  - Cultural institutions
  - Recreation
  - The press and broadcasting

History 431
- Exploration, settlement, and independence 431
- Establishment of the nation 432
- Paraguay's conflicts with its neighbours 432
- Since the Chaco War 433

Bibliography 433

Physical and human geography

THE LAND

Relief. The Paraguay River, which runs from north to south, divides Paraguay into two distinct geographic regions—the Región Oriental (Eastern Region), also known as the Paraneña region, and the Región Occidental (Western Region), also known as the Chaco Boreal (Northern Chaco).

The Región Oriental and the Chaco Boreal

The Región Oriental, with an area of about 61,700 square miles, is an extension of the Brazilian Plateau and varies in elevation from about 165 feet (50 metres) above sea level in the southwest to a few hills that rise to 2,500 feet in the east. The Cordillera de Amambay (in Brazil called Serra de Amambai) runs approximately north to south along part of the border with Brazil and then runs eastward as the Cordillera de Mbaracayú (in Brazil called Serra de Maracaju). From the northeast, other ranges extend southward toward Encarnación, diminishing to hills in the south. The highest peak is Mount San Rafael at 2,789 feet (850 metres), in the Cordillera de San Rafael in southeastern Paraguay. To the east of these mountains lies the Paraná (Alto [Upper] Paraná) River valley. To the west lies the broad valley of the Paraguay River. The area from Encarnación northward to the Brazilian border, comprising one-third of eastern Paraguay, is called the Paraná Plateau. The western part of the Región Oriental and the Paraná valley north and east of Encarnación are the areas most favourable to human settlement.

The Chaco Boreal, which covers more than 95,000 square miles, about two-thirds of the country, forms the northeastern part of the Gran Chaco, a flat and largely featureless tropical region that extends into Bolivia and Argentina.

Drainage. Four-fifths of the country's perimeter is traced by the Paraguay, Apa, Paraná, and Pilcomayo rivers. The Paraná forms both the eastern and southern borders of the country. Multiple tributaries of the Paraguay and Paraná cross the eastern and central regions. The mountain ranges of Amambay and Mbaracayú form the watershed between the Paraguay and the Paraná rivers. Important eastern tributaries of the Paraguay River include, from north to south, the Apa, Aquidabán, Ypané, Jejuí Guazú, and Tebicuary. Except for the Acaray and Ytambey rivers, the streams that flow into the Paraná have little economic significance. The Paraná joins the Paraguay River at the country's southwestern corner. The only important tributary flowing from the west is the sluggish Pilcomayo, which joins the Paraguay near Asunción. Rising to the northwest in Bolivia, the Pilcomayo forms the southern border of the Chaco Boreal and is navigable in its lower reaches by small boats. Other Chaco rivers, including the Verde and Monte Lindo, are slow, sluggish, intermittent streams that drain into swamps or disappear during dry periods.

Lakes

Paraguay has only two lakes of consequence. The largest, Lake Ypoá, about 40 miles (65 kilometres) south of Asunción, merges into Lake Verá; it is drained by channels of the Tebicuary and feeds the marshes of the Ñeembucú plain. Lake Ypacaraí, about 30 miles east of Asunción, has a favourite summer resort at San Bernardino.

Soils. A large part of eastern Paraguay is covered by a residual soil mantle so deep that bedrock is rarely exposed. This soil is generally red and sandy and is low in nitrogen and other basic plant foods. About 40 percent of eastern Paraguay, in a belt running from the Brazilian border south to the Tebicuary and including the Asunción area, is covered by soils underlain by sandstone. Soils from basaltic lava, which generally are the most fertile, cover the Paraná Plateau. Transported soils cover a band along the Paraguay River, extending from the Apa River to the southern border and covering the Ñeembucú plain. Soils of the Gran Chaco are largely alluvial mud, clay, and sand that have been transported from the Bolivian highlands.

Climate. The climate is subtropical in most of eastern Paraguay, which mainly lies south of the Tropic of Capricorn, and tropical in most of the Chaco Boreal extending to its north. Masses of humid air blanket the country in the summers, and the winters are subject to cold southern winds. Summer temperatures between October and March generally range from 77° F to 104° F (25° C to 40° C). Winter temperatures usually range from 61° F to 73° F (16° C to 23° C), although extremes of 30° F and 104° F (−1° C and 40° C) are not uncommon. Frost occurs frequently in the Paraneña region. The annual average rainfall in eastern Paraguay varies from 65 inches (1,600 millimetres) in the southeast to about 55 inches along the Paraguay River. It diminishes gradually westward across the Chaco Boreal, averaging about 30 inches. The heaviest rainfall is from October to April. The entire country is subject to periodic floods and droughts, both of which cause severe agricultural losses.

Plant and animal life. More than half of Paraguay was forested in the 1940s, particularly the northern and eastern parts of the country, but by the end of the 1980s the proportion had dropped to nearly one-third. Rapid deforestation began in the 1970s, particularly with the extension of the agricultural frontier in the eastern border region. Widespread environmental damage ensued, as reafforestation has been minimal. Official estimates of the rate of deforestation suggest that, if the present rate is maintained, Paraguay is in danger of losing virtually all its forests by the middle of the 21st century.

Deforestation

More than 500 species of hardwoods have been identified, among which are the urunday, peterebi, curupay, lapacho, and many kinds of palms. The supply of cedars has been exploited extensively for furniture, boxes, and general use. Large stands of native yerba (*Ilex paraguayensis*) occur in many areas. Forest resources of the Chaco include many species of hardwoods that yield tannin, of which the red quebracho is the most important. Perhaps the most famous is the *samuú* on account of its bottle-shaped trunk. Numerous palms, such as the caranday, are commercially useful. The very hard palo santo ("holy wood") yields a valuable oil. Much of the Chaco is covered by cacti and a thorny scrub growth similar to the *caatinga* of Brazil's northeast. Medicinal plants abound in Paraguay and formerly were the basis of an extensive native pharmacopeia.

Wildlife includes marsh deer, monkeys, armadilloes, anteaters, otters, wild boars, tapirs, jaguars, ocelots, and the coypu, a South American aquatic rodent. In the Chaco there is a small reserve of Chacoan peccaries, thought by scientists to be long extinct until the early 1970s when they "discovered" living representatives of the species. Some types of caiman (particularly the yacare), parrots, and macaws are species threatened because of illegal trade. There is also extensive trading in snake and iguana skins for export. The birdlife is spectacular and includes parakeets, rheas, ibises, herons, toucans, eagles, falcons, doves, and bats. Insect life is extensive and includes locusts, mosquitoes, cockroaches, and the tarantula spider. The rivers abound with fish, including the piranha.

Birds and insects

SETTLEMENT PATTERNS

Paraguay has one of the lowest population densities in the world. Less than 5 percent of the population lives west of the Paraguay River in the Chaco.

Rural settlement. The principal areas of rural settlement are in the east. Paraguayans and Brazilians have settled there in large numbers, particularly in the departments of Alto Paraná and Canindeyú. There are several German-speaking Mennonite colonies in Paraguay, about half of them in the Chaco. The first were established about 120 miles west of Puerto La Victoria (also Puerto Casado) in the 1920s and '30s. Japanese immigrants, especially after the 1930s, established thriving agricultural colonies southeast of Asunción and near Encarnación. By the 1980s the number of Japanese rural settlers had risen to about 8,000.

Urban settlement. The largest city is the capital, Asunción; the metropolitan area of Asunción includes the cities of San Lorenzo and Fernando de la Mora. Ciudad del Este (formerly Puerto Presidente Stroessner) and Hernandarias, both in the east, grew rapidly in the 1970s. Other important urban centres are Concepción, Encarnación, Pedro Juan Caballero, Coronel Oviedo, and Villarrica.

THE PEOPLE

Ethnic composition. Paraguay has the most homogeneous population of any South American country. The



MAP INDEX

Political subdivisions

Alto Paraguay . . . . 21 10 s 58 50 w
Alto Paraná . . . . . 25 30 s 54 50 w
Amambay . . . . . 23 00 s 56 00 w
Asunción . . . . . 25 16 s 57 40 w
Boquerón . . . . . 21 10 s 61 00 w
Caaguazú . . . . . 25 20 s 55 50 w
Caazapá . . . . . 26 10 s 56 00 w
Canindeyú . . . . . 24 15 s 55 15 w
Central . . . . . 25 30 s 57 30 w
Concepción . . . . . 23 00 s 57 00 w
Cordillera . . . . . 25 15 s 57 00 w
Guairá . . . . . 25 45 s 56 30 w
Itapúa . . . . . 26 50 s 55 50 w
Misiones . . . . . 27 00 s 57 00 w
Ñeembucú . . . . . 27 00 s 58 00 w
Paraguarí . . . . . 26 00 s 57 10 w
Presinte Hayes . . . . . 23 30 s 58 50 w
San Pedro . . . . . 24 15 s 56 30 w

Cities and towns

Acahay . . . . . 25 55 s 57 09 w
Alberdi . . . . . 26 10 s 58 09 w
Areguá . . . . . 25 18 s 57 25 w
Asunción . . . . . 25 16 s 57 40 w
Ayolas . . . . . 27 24 s 56 54 w
Bella Vista . . . . . 22 08 s 56 31 w
Benjamín Acevai . . 24 58 s 57 34 w
Caacupé . . . . . 25 23 s 57 09 w
Caaguazú . . . . . 25 26 s 56 02 w
Caapucú . . . . . 26 13 s 57 12 w
Caazapá . . . . . 26 09 s 56 24 w
Capitán Bado . . . . 23 16 s 55 32 w
Capitán Pablo Lagerenza . . . . . 19 55 s 60 47 w
Carapeguá . . . . . 25 48 s 57 14 w
Carmen del Paraná) . . . . . 27 14 s 56 13 w
Ciudad del Este (Puerto Presidente Stroessner) . . . . . 25 31 s 54 37 w
Concepción . . . . . 23 25 s 57 17 w
Coronel Bogado . . 27 11 s 56 18 w
Coronel Martínez . . 25 45 s 56 37 w
Coronel Oviedo . . . 25 25 s 56 27 w
Doctor Cecilio Báez . . . . . 25 03 s 56 19 w
Doctor Juan León Mallorguín . . 25 24 s 55 16 w
Doctor Pedro P. Peña . . . . . 22 27 s 62 20 w
Emboscada . . . . . 25 09 s 57 21 w
Encarnación . . . . . 27 20 s 55 54 w
Eusebio Ayala . . . . 25 24 s 56 58 w
Filadelfia . . . . . 22 21 s 60 02 w
Fuerte Olimpo . . . . 21 02 s 57 54 w
General Artigas . . . 26 56 s 56 13 w
General Eugenio A. Garay . . . . . 20 31 s 62 08 w
Hernandarias . . . . 25 22 s 54 45 w
Hohenau . . . . . 27 05 s 55 45 w
Horqueta . . . . . 23 24 s 56 53 w
Itá . . . . . 25 29 s 57 21 w
Itacurubí del Rosario . . . . . 24 29 s 56 41 w
Iturbe . . . . . 26 01 s 56 30 w
Jesús . . . . . 27 03 s 55 47 w
La Colmena . . . . 25 53 s 56 48 w
Limpio . . . . . 25 11 s 57 32 w
Luque . . . . . 25 16 s 57 34 w
Mariscal Estigarribia . . . . 22 02 s 60 38 w
Paraguarí . . . . . 25 38 s 57 09 w
Pedro Juan Caballero . . . . . 22 34 s 55 37 w
Pilar . . . . . 26 52 s 58 23 w
Piribebuy . . . . . 25 29 s 57 03 w
Poro Colorado . . . 23 26 s 58 58 w
Presidente Franco . . . . . 25 32 s 54 37 w
Puerto La Victoria . . . . . 22 20 s 57 59 w
Puerto Presidente Stroessner, see Ciudad del Este
Quiindy . . . . . 25 58 s 57 16 w
Rosario . . . . . 24 27 s 57 03 w
Salto del Guairá . . . . . 24 05 s 54 20 w
San Bernardino . . . 25 16 s 57 19 w
San Estanislao . . . 24 39 s 56 26 w
San Ignacio . . . . . 26 52 s 56 57 w
San Isidro de Curuguaty . . . . . 24 31 s 55 42 w
San Juan Bautista . . . . . 26 38 s 57 10 w
San Juan Nepomuceno . . . . 26 06 s 55 58 w
San Lázaro . . . . . 22 10 s 57 58 w
San Lorenzo . . . . . 25 20 s 57 32 w
San Miguel . . . . . 26 30 s 57 03 w
San Pedro de Ycuamandiyú . . . 24 06 s 57 05 w
San Pedro del Paraná . . . . . 26 50 s 56 12 w
Santa Rosa . . . . . 26 52 s 56 51 w
Santiago . . . . . 27 09 s 56 47 w
Sapucai . . . . . 25 40 s 56 55 w

Villa del Rosario, see Rosario 25 06 s 57 34 w
Villa Hayes 25 06 s 57 34 w
Villarrica 25 45 s 56 26 w
Villeta 25 28 s 57 36 w
Ybycuí 26 01 s 57 03 w
Yhú 24 59 s 55 59 w
Ypacaraí 25 23 s 57 16 w
Yuty 26 32 s 56 15 w

Physical features and points of interest

Acaray, *river* 25 29 s 54 42 w
Alto Paraná, see Paraná
Amambay, Cordillera de (Serra de Amambai) 23 10 s 55 30 w
Apa, *river* 22 06 s 58 00 w
Aquidabán, *river* 23 11 s 57 32 w
Aracanguy (San Joaquin) Hills 24 30 s 54 30 w
Caaguazú, Cordillera de 25 36 s 55 58 w
Caaguazú National Park 25 55 s 55 51 w
Cerro Corá National Park 22 38 s 56 00 w
Chaco Boreal, *region* 23 00 s 60 00 w
Curupaytiy, *river* 22 03 s 57 59 w
Defensores del Chaco National Park 20 40 s 60 15 w
Fernheim, *Mennonite colony* 22 40 s 60 00 w
Friendship (Amistad or Amizade) Bridge 25 31 s 54 37 w
Gran Chaco (Chaco), *region* 23 00 s 60 00 w
Guairá (Sete Quedas) Falls 24 02 s 54 16 w
Itaipu Dam 25 22 s 54 34 w
Itaipu Reservoir 25 00 s 54 30 w
Jejuí Guazú (Jejuí), *river* 24 13 s 57 09 w
León Hill 20 23 s 60 19 w
Mbaracayú, Cordillera de (Serra de Maracaju) 23 57 s 55 01 w
Mbaracayú Nature Reserve 24 09 s 55 28 w
Melo, *river* 21 24 s 57 54 w
Menno, *Mennonite colony* 22 15 s 59 40 w
Monte Lindo, *river* 23 56 s 57 12 w
Neembucú, *river* 26 53 s 58 17 w
Negro, *river* 24 23 s 57 11 w
Neuland, *Mennonite colony* 22 40 s 60 30 w
Oriental, *region* 25 00 s 56 00 w
Paraguay, *river* 27 18 s 58 38 w
Paraná (Alto Paraná), *river* 27 15 s 58 00 w
Paraná Plateau 25 30 s 55 30 w
Patiño Marsh 24 05 s 59 50 w
Pilcomayo, *river* 25 21 s 57 42 w
San Joaquin, see Aracanguy Hills
San Joaquin, Sierra de 24 48 s 56 00 w
San Rafael, Cordillera de 26 25 s 55 38 w
San Rafael, Mount 26 44 s 55 50 w
Sete Quedas, see Guairá Falls
Talavera Island 27 32 s 56 26 w
Tebicuary, *river* 26 36 s 58 16 w
Teniente Encisco National Park 22 28 s 61 32 w
Tinfunque National Park 24 29 s 59 20 w
Trinidad, Lake 20 21 s 61 35 w
Verá, Lake 26 05 s 57 39 w
Verde, *river* 23 09 s 57 37 w
Yacaré Norte, *river* 22 43 s 58 14 w
Yacyretá Island 27 25 s 56 30 w
Ybycuí, Mount 25 59 s 57 01 w
Ybycuí National Park 25 58 s 56 59 w
Ypacarai, Lake 25 17 s 57 20 w
Ypané, *river* 23 29 s 57 19 w

vast majority of inhabitants are native Paraguayans, who are almost all mestizo (a mixture of Spanish and Indian). They pride themselves on their Guaraní descent, although the admixture of European strains is prominent. About 300,000 Brazilians, many of them small farmers, arrived in eastern Paraguay in the 1970s, because land in Paraguay was cheaper than in Brazil. Other immigrants have come from western Europe, particularly Germany, Italy, and Spain, and from Japan, Korea, Hong Kong, and Taiwan.

Indian groups

Estimates of the number of Indians in the country range from 40,000 to 70,000. They comprise six linguistic groups, five in the Chaco and one in the east. The 17 ethnically distinct Indian groups include the Pai-Tavyterá, Mbya, Aché, and Chiripá (in the east) and the Toba, Mascoy, Lengua, Chulupí, Tapieté, Ayoreo, and Chamacoco (in the Chaco). Some are threatened with extinction through forced assimilation and the destruction of their traditional habitat. The welfare of the Indians is the official responsibility of the Paraguayan Indian Institute. Many Indians receive support from missionary groups.

Language groups. Until 1992 Spanish was the official language, although Guaraní was spoken by nearly 90 percent of the population. But the 1992 constitution recognizes both Spanish and Guaraní as official languages. Spanish is used predominantly in government and business, but both Spanish and Guaraní are used as media of instruction. At least half of the population is bilingual. The constitution also recognizes Indian languages as part of the country's heritage.

Religious groups. More than 95 percent of the population professes adherence to Roman Catholicism. There is also a sizable minority of Protestants. The constitution recognizes no official religion and emphasizes the state's independence from the Catholic church.

Demography. Between 1970 and 1990 the population of Paraguay increased from 2.4 million to 4.3 million. Although migration to the cities is common, more than half the population lives in the country. Emigration has been high since the mid-20th century, especially among men seeking employment in neighbouring countries. About 25 percent of all Paraguayans live outside their country.

Life expectancy at birth is about 65 years for men and about 70 years for women. The population is young, with more than two-thirds under 30 years of age. Infant mortality has decreased significantly since the 1960s but remains high. The distribution of the population is not uniform, the highest density occurring in the Asunción metropolitan area.

THE ECONOMY

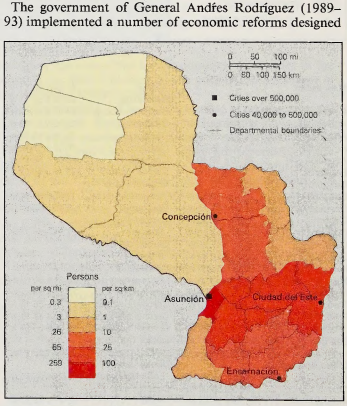
Agriculture is the most common economic activity, employing more than two-fifths of the workforce. It also accounts for about one-fourth of the gross domestic product (GDP) and the vast majority of exports. The economy is therefore highly dependent on the vagaries of climate and world commodity prices for its main agricultural products.

General Alfredo Stroessner Matiauda (president 1954–89) encouraged domestic and foreign private investment, particularly in commercial agriculture. Until the mid-1970s, public-sector investment was low by Latin American standards and was concerned mainly with improving roads, telecommunications, and air transport. This changed with the establishment of several state companies, most notably Itaipú Binacional, set up in 1973 to build a huge hydroelectric dam on the Paraná, and steel, cement, and alcohol-distillation plants. Public-sector employment grew rapidly, making up about one-tenth of the labour force during the late 20th century. Until 1982, when the construction of Itaipu was completed, Paraguay was able to offset its current account and trade deficit with international loans. For the rest of the decade, the country was faced with a growing public-sector deficit, high debt repayments on commercial borrowing, and dwindling international reserves.

Impressive economic growth, particularly in the 1970s, was not matched by government efforts to distribute its benefits equitably. Most Paraguayans, particularly in rural areas, remained poor by Latin American standards. Income tax was kept low, partly in order to attract foreign capital, and foreign investors were exempt from many taxes. Farmers growing cash crops for export were given state subsidies. The police and armed forces absorbed a large portion of the budget.

Economic reforms under Rodríguez

The government of General Andres Rodríguez (1989–93) implemented a number of economic reforms designed



Population density of Paraguay.

to introduce a market-based economy. They included the abolition of a multiple exchange rate, the end of cheap subsidies to state companies, the elimination of export taxes, and plans to privatize several state companies.

Under General Stroessner, labour unions were strictly controlled, which helped to keep wage increases low. For most of his rule, the country had one large, government-recognized trade union, the Confederation of Paraguayan Workers (Confederación Paraguaya de Trabajadores; CPT). After Stroessner's fall, a number of independent union groupings emerged, most notably the Unified Workers Central (Central Unitaria de Trabajadores; CUT). In 1991 the United States government restored Paraguay's trade benefits under the Generalized System of Preferences (GSP), from which it had been suspended in 1987 as a result of the lack of trade union freedom.

Resources. *Mining and quarrying.* Most mineral deposits are found east of the Paraguay River. Manganese is located near Emboscada; malachite and azurite (copper ores) near Caapucú, Encarnación, and San Miguel; feldspar and mica near Concepción; and talc and *pirofilita* (hard, iron-bearing flagstone) near Caapucú and San Miguel. Ochre is found in the Cordillera region, and gypsum and limestone near the Paraguay River; there is some peat near Pilar. Marble, clay (kaolin), and salt are quarried. Copper, bauxite, iron, and uranium ores have been reported. Extensive drilling in the Paraguayan Chaco has failed to find any hydrocarbons. Despite the varied mineral resources, mining and quarrying are the least-developed economic activities.

Energy. Paraguay's most important natural resource is its hydroelectric potential. Most electricity in Paraguay came from wood- and oil-burning thermoelectric plants in Asunción until the Acaray hydroelectric power plant began operating in 1968. The plant's capacity was expanded to 190,000 kilowatts. Paraguay's total production increased more than 15-fold from 1970 to 1990. Nearly all of this increase came from hydroelectric sources. Distribution of electricity is controlled by the National Power Company, which was created in 1949.

Itaipu hydroelectric project

A dramatic and far-reaching economic event in Paraguay's history was the construction, in partnership with Brazil, of the hydroelectric project at Itaipu on the Paraná, about 10 miles north of the Friendship Bridge at Ciudad del Este. Under a 1973 treaty a major portion of the cost was borne by Brazil. It was agreed that Paraguay would receive one-half of the electricity that was generated and would sell its excess power to Brazil at predetermined rates. Financing, provided by both private and international banks, reached about $20 billion. The Treaty of Itaipu was revised in 1985, giving fairer compensation to Paraguay, although the terms are still questioned by Paraguayan nationalists.

Work was completed in 1982 on the main gravity dam, 620 feet high and 4,045 feet long, spanning the Paraná. The reservoir created by the dam covers about 870 square miles of Paraguayan and Brazilian territory. By 1990 the plant was operating at about 60 percent of its capacity. The last of its 18 turbines was completed in 1991. The dam is one of the largest in the world and has a capacity of 12,600 megawatts, which is one of the world's highest planned generating capacities.

The joint Paraguayan-Argentine hydroelectric project in the Yacyretá-Apipé islands zone of the Paraná is smaller but still important, with planned capacity of 2,700 megawatts. It is expected to be fully operational in the mid-1990s. Because domestic demand is likely to absorb only 5 percent of the combined output of Itaipu and Yacyretá, Paraguay will become one of the world's largest exporters of electricity.

Agriculture, forestry, and fishing. Although about one-fifth of Paraguay's total land area is suitable for intensive cultivation, only 5.6 percent of this is utilized steadily, and virtually all of it is in the east. Most farm units are occupied by owners, though there are small groups of tenant farmers and squatters. Paraguay has a highly skewed system of land tenure, which is largely a legacy of land sales after the War of the Triple Alliance. During the late 20th century more than three-quarters of the land was owned by 1 percent of landholders. The Rural Welfare Institute has helped several thousand farmers acquire land, particularly in the eastern region, but the number of landless families remains high.

Distribution of land

Until the 1970s the economy was largely dependent on the export of tannin, meat products, yerba maté (a type of tea), and tobacco. Whereas these products have declined, the cultivation of soybeans and cotton, which are grown in the eastern region, has increased significantly. Other important crops are cassava, sugarcane, corn (maize), rice, wheat, peanuts (groundnuts), coffee, and citrus fruits. A significant amount of marijuana is grown in Paraguay as well. The country is self-sufficient in most foodstuffs.

Some pigs, sheep, chickens, and horses are raised, but cattle are the most important livestock. Cattle raising, a traditional activity, is particularly prevalent in the Chaco and in the southern departments of Misiones and Ñeembucú. Meat and dairy products and hides contribute to both domestic consumption and exports. Timber products have long been an important export.

Fish are plentiful in the rivers, and *surubí, pacú,* and dorado are popular domestically. There is no large-scale commercial fishing industry, however.

Industry. Although the industrial sector registered high growth rates in the late 1970s and early 1980s, Paraguay is one of the least industrialized countries in South America. Because of the limited quantities of proven mineral reserves, there is mining of only limestone, gypsum, and clays, mostly for the building trade. Manufacturing is generally small-scale and directed toward processing agricultural products. These include refined soybean oil, sugar, tinned meat, textiles, leather products, alcohol, beer, and cigarettes. The construction and cement industry boomed in the late 1970s and early 1980s because of the Itaipu and other hydroelectric projects. There are a small steel mill, inaugurated in 1986, and a factory that has produced ethyl alcohol (ethanol) from sugar cane since 1980; both are owned by the government.

Tourism plays a minor role in the economy. Most visitors are from Brazil and Argentina; they are attracted by Paraguay's status as a major importer of consumer goods.

Finance. The main state banks are the Banco Central del Paraguay, which handles all monetary functions, and the Banco Nacional de Fomento, which grants credits to agricultural enterprises and the manufacturing and lumber industries. There also are branches of Latin American, European, and U.S. commercial banks. Foreign currency is freely available at banks and exchange houses. In 1992 the government approved laws encouraging foreign investment and the development of a stock market. The currency, the guaraní, has been relatively stable by Latin American standards.

Trade. The principal exports—soybeans, cotton, meat products, tung and other industrial oilseeds, essential oils (mainly petit grain), and sawn timber—are marketed in Brazil, Argentina, The Netherlands, Italy, Germany, and Switzerland. The main imports are machinery, petroleum and petroleum products, chemicals, and automobiles and buses, principally from Brazil, the United States, Japan, and Argentina. During the early 1980s, Paraguay ran a trade deficit. Because of widespread smuggling, official figures severely underestimated the real size of Paraguay's trade. When in 1989 the government of General Rodríguez replaced the multiple exchange rate with a single, fluctuating rate and old tax rates with new unified rates, it weakened some of the causes of smuggling.

Smuggling

Paraguay is a member of the Latin American Integration Association (LAIA; formerly the Latin American Free Trade Association [LAFTA]). In 1991 Paraguay, Brazil, Uruguay, and Argentina signed a free trade agreement creating the Mercado Común del Cono Sur (Mercosur), a common market of countries from the "Southern Cone."

Transportation. During the mid-20th century, most international freight was transported along the Paraguay and Paraná rivers, which linked Asunción and other Paraguayan ports to the Atlantic Ocean via Brazil. Freight is increasingly taken by road, particularly to Buenos Aires, Arg., and the Brazilian cities of Santos and Paranaguá.

Roads. Paraguay has a sizable road network equipped

with adequate bridges, but a considerable portion remains unpaved. The country's major highway connects Asunción with Ciudad del Este, where the Friendship Bridge spans the Paraná and carries the highway into Brazil. This paved road continues to the port of Paranaguá, from which Paraguay exports most of its soybeans. Another major paved highway joins Asunción with Encarnación. A suspension bridge, part of the Pan-American Highway, links Asunción and Clorinda, Arg. Another bridge links Encarnación to Posadas in Argentina. The Trans-Chaco highway runs northwest from Villa Hayes to the Bolivian border. It is paved as far as Filadelfia.

Railways. Most of the railway system is made up of the Ferrocarril (Railway) Presidente Carlos Antonio López, which runs for 274 miles from Asunción southeastward to Encarnación, where it connects with a train ferry to Posadas, Arg.

Water transport. Paraguay's merchant marine, the state-owned Flota Mercante del Estado, was created in 1945 and runs cargo vessels on the Paraguay and Paraná rivers. In addition, there are a fleet of merchant ships and a number of small shipping enterprises in the private sector. Asunción is the largest port and has modern facilities.

Air services. Air transport is provided by the government-owned national airline, Líneas Aéreas Paraguayas, which has greatly expanded its services since its creation in 1962. Líneas Aéreas de Transporte Nacional and Transporte Aéreo Militar serve interior cities. Nine miles from Asunción there is a modern international airport named Silvio Pettirossi (formerly Presidente Stroessner); it was opened in 1980.

GOVERNMENT AND SOCIAL CONDITIONS

The 1992 Constitution

Government. *Constitutional framework.* The 1992 constitution is the basic charter of Paraguay. It was drawn up by a Constituent Assembly, which was elected in December 1991, and it replaced the previous constitution of 1967. The constitution states that Paraguay is a representative and pluralist democracy and that government is exercized by the separate powers of the legislature, executive, and judiciary.

The president is elected by a simple majority of votes for a five-year term. He must be a Paraguayan by birth and more than 35 years of age. There is no runoff election if the leading candidate fails to obtain an absolute majority. President Stroessner amended the 1967 constitution in 1977 to allow his reelection, but the 1992 constitution specifically rules out a president's reelection. The president is the commander in chief of the armed forces and is authorized to appoint and remove commanders of the army and police. The 1992 constitution created the post of vice president. A council of ministers is appointed by the president. A state of exception can be declared, by the executive or the Congress, only in cases of international armed conflict or serious internal unrest threatening the rule of the constitution.

The legislative body is the Congress, composed of the Chamber of Deputies and the Senate. All its members are elected by popular vote for five-year terms on the same date the presidential elections are held. The Chamber of Deputies is made up of a minimum of 80 members elected from the 17 departments, and the Senate of a minimum of 45 members.

The constitution guarantees rights to Indians, the right to strike, and basic civic liberties, including freedom of expression, association, and religion. The death penalty has been abolished.

Local government. Paraguay is divided into 17 departments, 14 in the east and 3 in the Chaco. Each department is divided into *distritos* ("districts"). Until 1991 the central government used to appoint departmental governors and local mayors, but in May of that year direct municipal elections were held for the first time. The 1992 constitution, in a further innovation, provided for elections for a governor and a departmental board for each department, also to be held every five years.

The political process. *Elections.* Elections are governed by an electoral code, which can be changed by Congress. Voting is compulsory for all Paraguayans from age 18 onward. Resident aliens are allowed to vote in municipal elections. Until 1990 the party winning a simple majority was awarded two-thirds of the seats of both chambers. This was replaced by a system of proportional representation.

Political parties. Under General Stroessner, political parties were closely controlled, including the dissident factions of the ruling Colorado Party (Partido Colorado), officially named the National Republican Association (Asociación Nacional Republicana; ANR). The police kept dossiers on citizens, particularly political opponents, and political repression was widespread. Senior generals played a major part in government. Political freedom improved significantly under General Rodríguez. The internal factions of the Colorado Party are openly tolerated. Other parties include the Authentic Radical Liberal Party (Partido Liberal Radical Auténtico; PLRA), the February Revolutionary Party (Partido Revolucionario Febrerista; PRF), the Christian Democratic Party (Partido Demócrata Cristiano; PDC), the Paraguayan Communist Party (Partido Comunista Paraguayo; PCP), and several independent groupings. Parties dedicated to the substitution of force for democracy may not be organized. No party may receive aid or instructions from foreign organizations or states or establish structures that directly or indirectly embrace violence as a political methodology.

The Supreme Court

The judicial system. The judicial system is headed by the Supreme Court. The 1992 Constitution increased its members from five to nine, who are chosen by the Senate and the executive. Supreme Court justices are appointed for a term of five years. Judges who are confirmed for two terms following the terms of appointment cannot be removed from their post until they reach the mandatory retirement age established for Supreme Court justices. The Supreme Court appoints judges of lower courts and magistrates. There is also an ombudsman and an attorney general appointed by the executive. The judiciary has budgetary autonomy.

Armed forces. Although the armed forces are not large (16,500 members in the early 1990s), they and the police force receive a disproportionately high share of the total budget. The 1992 constitution reduced military service, which is compulsory, from 18 to 12 months. During the Stroessner years, all military officials were obliged to become members of the Colorado Party. The 1992 constitution banned all military personnel in active service from belonging to political parties or engaging in any political activity.

Education. Primary education is free and, where possible, compulsory for children between the ages of 7 and 13. Although the official enrollment figures are high (about 70 percent by 1990), the dropout rate is also high. The 4,600 primary schools are supplemented by some 800 secondary schools in all departments. The two universities—the public National University of Asunción (1890) and the private Our Lady of the Assumption Catholic University (1960)—are located in Asunción; attendance is free at the National University. Paraguay also has schools of agriculture and veterinary science. Half of all university graduates are women. The United Nations Educational, Scientific and Cultural Organization (UNESCO) estimated adult literacy at 90 percent in the late 20th century, although functional literacy is probably lower. In the late 1980s spending on education was slightly below defense spending, at about 12 percent of the total budget. The number of schools is insufficient, especially in rural areas, and adequate teaching resources are scarce everywhere.

Health and welfare. Typhoid, measles, tuberculosis, acute respiratory infections, cancer, dysentery, hookworm, and hepatitis are particularly prevalent. There also is some occurrence of leprosy and malaria. Although infant mortality rates have declined significantly since the 1960s, malnutrition and poor implementation of immunization programs have led to thousands of preventable deaths, particularly in rural areas. The United Nations Environmental Programme (UNEP) estimates that 65 percent of the urban population had access to drinking water in the late 1980s, compared with 8 percent of the rural population. For the same years, 90 percent of the ur-

ban population and 38 percent of the rural population had access to health services. Health spending in the late 1980s was low—only about 3 percent of the total budget. In the late 1980s there were approximately 150 hospitals and clinics and about 4,000 physicians, most of them in Asunción. The state-run Institute of Social Provision (IPS) is funded by contributions from government, employers, and employees. It offers pensions, medical care, and subsidies during illness. However, its service is inefficient and reaches only a small percentage of the salaried workers.

CULTURAL LIFE

Guaraní and Spanish traditions

The main characteristic of Paraguayan culture is its fusion of both the Guaraní and Spanish traditions. Folklore, the arts, and literature reflect this dual origin. The country's outstanding handicraft is the production of *ñandutí* lace, which is thought to represent a combination of 16th-century needlepoint lacemaking techniques from Europe with Guaraní traditions.

Daily life. Social life tends to revolve around the family. Godparents are particularly important; if parents become unable to provide for their children, godparents are expected to assume responsibility for them. The family usually determines the individual's allegiance to political parties; a change of political affiliation is often considered an act of betrayal. Party membership means less the adherence to a political ideology than the unswerving support of the party's candidates. Especially in the rural areas, such loyalty is often the route to employment.

Outside Asunción the pace of life is slow. A common pastime is drinking *tereré* (a bitter tea of yerba leaves) from a shared gourd or a hollowed cow's horn, or *guampa,* which often is beautifully carved.

The arts. Paraguay has a distinctive musical tradition, especially of songs and ballads. Typical music for dancing includes polkas and languid *guaranias,* played on the native harp. José Asunción Flores (1904–72) was an outstanding composer and harpist. Perhaps the most famous dance is the *galopa,* a variant of which is the bottle dance, in which dancers balance bottles on their heads. The number of books published in Paraguay increased significantly in the 1980s and particularly after the coup in 1989. Paraguay's most famous author is Augusto Roa Bastos (b. 1917), whose novel *Yo, el supremo* (1974; *I, the Supreme*), based on the life of the 19th-century dictator José Gaspar de Francia, has won wide acclaim.

Cultural institutions. Paraguay's principal cultural institutions are located in Asunción. There are learned societies concerned with Paraguayan and Guaraní history and culture, as well as various other societies and research institutes. The Normal School of Music, the Conservatory of Music, the National Academy of Fine Arts, and the Asunción Symphony Orchestra are major arts organizations. Paraguay has museums of ethnography, natural history, and military history, as well as museums with collections of the work of Paraguayan artists.

Library services are centred in Asunción. The largest collections are in the National Library and Archive and the Museum of Natural History and Ethnography.

Recreation. Paraguayans are fond of sports. While football (soccer) is the most popular sport, fishing, tennis, basketball, and horse racing are also common.

Religious celebrations are also important. Thousands of Paraguayans visit Caacupé in December to participate in the city's annual celebration of the festival of the Virgin of Miracles.

The press and broadcasting. Censorship was practiced periodically during the Stroessner years but was relaxed considerably under President Rodríguez. Virtually all newspapers and periodicals are published in Spanish. Most important are the four Asunción daily newspapers, the *ABC Color* (founded 1967; closed March 1984–March 1989), *El Diario Noticias* (1984), *Hoy* (1977), and *Ultima Hora* (1973); they have a combined daily circulation of about 250,000. The Colorado Party publishes *Patria* (1946) and the Roman Catholic church the weekly *Sendero* (1973).

The National Telecommunications Administration oversees radio and television broadcasting. Radio Nacional is the government network, but there are many privately operated stations. Commercial television networks transmit from Asunción, Encarnación, and Ciudad del Este.

(J.E.Pa.)

For statistical data on the land and people of Paraguay, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

EXPLORATION, SETTLEMENT, AND INDEPENDENCE

The Guaraní. Indian tribes speaking the Guaraní language occupied the region between the Paraguay and Paraná rivers long before the arrival of Europeans. They were members of the Tupian language stock, which was widespread in South America, and in most respects resembled the other Indian tribes of the tropical forests. The women cultivated corn, manioc, and sweet potatoes, and the men hunted and fished. They were warlike, semi-nomadic people who lived in large thatched dwellings grouped in villages, and each village was surrounded by a defensive palisade. In the 15th century raiders from the Gran Chaco region made frequent attacks upon Guaraní tribes. The Guaraní retaliated, crossing the Paraguay River, and subdued their enemies, carrying the conflict into the margins of the Inca empire. They were, therefore, the natural allies of early European explorers who were seeking short routes to the mineral wealth of Peru. Alejo García, making his way from the Brazilian coast in 1524, and Sebastian Cabot, sailing up the Paraná in 1526, were the earliest of these explorers to reach the area.

Colonial period. The first colonial settlements were established by Domingo Martínez de Irala in the period 1536–56. The first Spanish colonists, unsuccessful in their search for gold, settled peacefully among the Guaraní in the region of Asunción, the present capital of Paraguay. These first settlers established their notorious "harems" of Guaraní women. Their racially mixed descendants gradually grew into the rural population of modern Paraguay, which still considers itself to be Guaraní in custom and habit. With Asunción as his principal base, Irala laid the foundations of Paraguay and made it the centre of Spanish power in southeastern South America. Irala's colonization policy involved the delimitation of the boundary with Brazil by a line of forts against Portuguese expansion, the foundation of villages, the settlement of the Guaraní to provide food, labour, and soldiers, and extensive Guaraní-Spanish intermarriage. Rapidly a national and fairly homogeneous amalgam of Indian and Spanish cultures came into being.

Jesuit missions to the Indians

From early in the 17th century, for more than 150 years, Jesuit communal missions in the Paraná and Uruguay basins of southeastern Paraguay governed the lives of 150,000 Indians in 30 *reducciones,* or settlements. These were centres of religious conversion, agricultural and pastoral production, and manufacturing and trade, which served also as strategic outposts against Portuguese expansion from southern Brazil. Isolated from the heart of Paraguay, which centred on Asunción, the mission became an autonomous military, political, and economic "state within a state," which increasingly excited the envy of the Spanish landowners in the Asunción area. In the period 1721–35 the latter waged a struggle to overthrow the Jesuit monopoly of Indian trade and labour. Unaided, the *reducciones* also had to defend themselves against slave raiders from São Paulo and, in 1754–57, a combined Spanish-Portuguese attack designed to enforce a territorial partition of the mission settlements. Defiance of such powerful groups paved the way for the expulsion of the Jesuits in 1767. The *reducciones* were abandoned; the Indians were absorbed by either the landed estates or the jungle; the settlements fell into ruin; and economic activity ceased.

In 1776 the new Viceroyalty of Río de la Plata was created, with its capital in Buenos Aires. This effectively made Asunción and all of Paraguay dependent on Buenos Aires, thus ending the region's colonial dominance.

Struggle for independence. As the power of Buenos Aires grew, the leaders of Paraguay began to resent the decline in their province's significance, and, although they had early challenged Spanish authority, they refused to ac-

cept the declaration of Argentine independence in 1810 as applying to Paraguay. Nor could an Argentine army under General Manuel Belgrano enforce Paraguayan acceptance, as Paraguayan militia repulsed Belgrano's forces in 1811. Later, however, when the Spanish governor sought assistance from the Portuguese in defending the colony from further attacks from Buenos Aires, he underestimated the nationalistic spirit of the Paraguayans. Under the leadership of the militia captains Pedro Juan Caballero and Fulgencio Yegros, they promptly deposed the governor and declared their independence on May 14, 1811.

(E.R.Se./G.J.B./J.H.Wi.)

ESTABLISHMENT OF THE NATION

The Francia regime. A governing junta was soon established, led by Yegros but in reality dominated by a civilian lawyer, José Gaspar Rodríguez de Francia. Francia proposed the idea of a confederation of equals to Buenos Aires, which, hoping for eventual domination, settled for a vague military alliance, signed in October 1811. This constituted de facto recognition of Paraguayan independence, and, when Buenos Aires attempted to use the alliance to acquire Paraguayan troops for its own interprovincial quarrels, the accord became void. Buenos Aires' response was to blockade Paraguay. In the face of regional fragmentation, Buenos Aires sent Nicolás de Herrera to Asunción to frighten, bluff, or bribe Paraguay into a union of unequals. Francia responded by convening a congress, which on Oct. 12, 1813, formally declared Paraguay an independent republic and rejected further treaties with Buenos Aires. A consulate of two men, Yegros and Francia, was established to rule the nation for a year.

Formal declaration of independence

At the end of that year, a new congress met and proclaimed Francia supreme dictator of the republic for a period of five years; in 1816 a third congress made him perpetual dictator, and his will was the law in Paraguay for an additional 24 years. El Supremo, as he was known, prohibited any political activity, stripped the church of its holdings and power, confiscated the wealth of the small Spanish elite, abolished the municipal government of Asunción, and generally isolated Paraguay from its rather hostile neighbours. In 1820 El Supremo discovered a plot to depose him and restore the native elite to power. Hundreds of arrests were made, and in the following year at least 68 men of the traditional Paraguayan aristocracy (including Fulgencio Yegros) were executed. Their wealth in land and slaves became part of the national patrimony, and well before Francia's death (1840) the state owned a vast proportion of the nation. With the borders sealed, Paraguay became, of necessity, almost self-sufficient; and only a small, carefully regulated commerce was permitted with Argentina and Brazil. Uninvited foreigners were often held for years under loose arrest in the interior.

Carlos Antonio López. When Francia died, he left behind a quietly prosperous nation that had adjusted well to what amounted to state socialism, but he also left a nation of rustics with no political experience and a strong tradition of dictatorial rule. In 1841 a second consulate emerged from the chaos in the figures of a civilian, Carlos Antonio López, and a soldier, Mariano Roque Alonso. It was soon clear that López was the true ruler of Paraguay, and in 1844 a congress named him president. The same congress promulgated a constitution, notable for the great powers accorded the president and the absence of the word liberty from its text. López devoted much of his two decades in power to opening the country slowly to the wider world and to modernization. Doing so provoked international crises, and it was not until after the fall of the Argentine dictator Juan Manuel de Rosas in 1852 that Argentina recognized Paraguayan sovereignty and eased its stranglehold on the rivers leading to the sea.

PARAGUAY'S CONFLICTS WITH ITS NEIGHBOURS

Modernization of the military forces

While he was "opening" his nation, López also had to attend to border crises with Brazil and Argentina. These crises convinced him that Paraguayan modernization should proceed along military avenues. Thus hundreds of foreign engineers, medical men, scientists, machinists, and advisers were put to work on military projects. López was threatened by a major Brazilian naval expedition on the Paraná River in 1855; in 1858 a large flotilla of the U.S. Navy appeared to force a solution to a complex diplomatic issue; and finally British war vessels captured and held for a time the flagship of the small Paraguayan navy. In most of these contretemps, López was forced to give in, and the consequent humiliation lent greater urgency to his desire to strengthen Paraguay's defenses. By the time of his death, in September 1862, he had created a major regional military machine. A cautious man, he warned his eldest son, Francisco Solano López, who was to succeed him, not to use the new military might capriciously but to settle disputes through diplomacy and negotiation.

Francisco Solano López and the War of the Triple Alliance. Francisco Solano López in 1862 was the inexperienced, spoiled son of an iron-willed dictator. He overestimated the military strength of his nation and felt that Paraguay should have a larger voice in the affairs of the region. Thus, when Uruguay, wracked by civil war, was threatened with intervention by Brazil, López took an increasingly bellicose position. When Brazil ignored his warnings and ultimatums and invaded Uruguay in August 1864 to support a pro-Brazilian faction in the civil war, López decided to use the strength of his military machine. In November he ordered the capture of a Brazilian war steamer and sent units of his army and navy north to invade the Mato Grosso, simultaneously preparing a larger army corps to strike south to destroy the Brazilian army in Uruguay. When Argentina denied his request for transit of a Paraguayan army, he declared war on Argentina as well, in March 1865. In May, as Paraguayan troops were approaching, a puppet Uruguayan government signed the Treaty of the Triple Alliance with Brazil and Argentina, committing all three to the war against Paraguay.

War of the Triple Alliance

The Paraguayan force heading southward was destroyed at Uruguaiana, in Brazil, and a strike into northeast Argentina resulted in heavy Paraguayan casualties and the virtual destruction of López' fleet in 1865. Much of the rest of the war was fought in southwestern Paraguay, near and around Humaitá. In May 1866 López threw the cream of his army into suicidal attacks against allied forces at Tuyutí, losing almost 20,000 of his best men. Other lost battles in 1866–68, as well as widespread epidemics of Asiatic cholera, devastated the population of the country. In 1869 and 1870 the tragedy was completed as López, pursued by large allied forces, retreated through the interior of his nation with a shattered army and thousands of civilian refugees, dragging famine, disease, and death in his wake. Perhaps at this point unhinged, he ordered the executions of hundreds of people, including his own two brothers and two brothers-in-law and scores of his officers. Finally, on March 1, 1870, his last camp was attacked at Cerro Corá by Brazilian cavalry, and López died in combat. His country by then lay in ruins, with one-third of its former population of 450,000 dead. A Brazilian occupation army remained, further draining the country, until 1876. This Paraguayan War, or War of the Triple Alliance, was the bloodiest in Latin American history.

Reconstruction. Under a liberal constitution promulgated in 1870, Paraguay began a painful reconstruction. Only the mutual jealousies of Brazil and Argentina prevented the country from losing much of its territory. As a result, Brazil gained no lands that it had not actually occupied before the war, and Argentina's claims to most of the Chaco were reduced considerably when, in arbitration, U.S. President Rutherford B. Hayes decided one key boundary issue in 1878 in favour of Paraguay. When the army of occupation was removed in 1876, it left a crowd of Paraguayan politicians noted for their corruption and ambition. In 1887 Paraguay's two major political parties, the Liberals and the Colorados, were born. The Colorados were in power from 1887 until a Liberal revolt unseated them in 1904, and the Liberals, in their turn, dominated the presidency for the next 30 years.

Formation of major political parties

The Chaco War. Paraguay's reconstruction was complicated by a dispute with Bolivia concerning boundaries in the Chaco. The dispute was exacerbated when in the 1880s Bolivia lost its seacoast in the War of the Pacific with Chile and, seeing the Chaco as a possible outlet to

the sea via the Paraná River, began to penetrate it with soldiers and colonists. By the 1920s, armed clashes began to take place as Paraguay moved into the region in greater force. As Paraguay was frantically trying to arm itself, a Bolivian force stormed a Paraguayan fort on June 15, 1932, and the war began. The Paraguayan president, Eusebio Ayala, gave a military carte blanche to General José Félix Estigarribia, who gradually pushed the Bolivians back until they were almost entirely ejected from the Chaco. Through foreign mediation, a cease-fire was attained on June 12, 1935, and a peace treaty was signed three years later, awarding Paraguay three-quarters of the Chaco.

SINCE THE CHACO WAR

In February 1936, Ayala and Estigarribia were imprisoned following a military coup, the Febrerista revolt, conducted by radical officers. The inept new government soon fell, however, and Estigarribia was elected president in 1939.

Estigarribia and Morínigo and their successors. On Sept. 7, 1940, before he could actually implement a new constitution that gave him great authoritarian powers, Estigarribia was killed in an air crash. He was replaced by General Higinio Morínigo, a harsh opportunist, who immediately persecuted the Liberals and rewarded the Colorados. A revolt of Liberals and other groups in 1947 caused a civil war that again devastated the country. Morínigo was deposed by the Colorados themselves in 1948. In the next six years, Paraguay had six weak presidents, and then, in 1954, General Alfredo Stroessner, supported by both Colorados and the army, seized power.

The Stroessner regime. The authoritarian Stroessner, with aid from the United States and later Brazil, managed to stabilize one of the world's least stable currencies, attract foreign investment, and embark on large public works projects. Paraguayan isolation was broken down. Harsh rule was relaxed after 1960, and elections on all levels were permitted; however, the Colorado Party never lost, and Stroessner was duly reelected every five years with a huge plurality. The church alone continued to object to the repressive aspects of the regime, such as the inhumane treatment of the Indian minority and censorship. Relations with the United States deteriorated throughout the 1970s, and U.S. aid was much reduced. Partly because of this, the Stroessner government aligned itself closely with the authoritarian regime in Brazil, which offered aid and political support. The two nations cooperated in the building of the immense Itaipu hydroelectric plant on their shared border. As a result of this project, the national economy briefly improved, but it took a downturn in the early 1980s, causing some protests against the Stroessner regime. The government showed little tolerance for opposition to its policies; most of the main opposition leaders were kept forcibly in exile. Such repression focused international attention on Paraguay for human rights violations, further hampering the country's foreign relations and intensifying economic stagnation. The aging Stroessner, who had been elected in 1983 for a seventh term, also had to deal with dissension within his own Colorado Party that pitted the traditionalist, or "moderate," wing of the party against the "militant" wing. The former sought to open the political and economic system somewhat, whereas the latter favoured continuation of the policies of the Stroessner regime and wanted Gustavo Stroessner to succeed his father whenever he stepped down from office.

Factional discord rocked the Colorado Party, resulting in a partial purge of the traditionalists in 1988, and it appeared briefly that the militants were firmly in control. The traditionalists, however, were simply lining up their forces for the inevitable conflict. On Feb. 3, 1989, General Stroessner was overthrown in a coup led by his erstwhile top military commander, General Andrés Rodríguez, who announced that democracy had come to Paraguay.

The end of the Stroessner era

The return of democracy. Elections were held on May 1, 1989, and General Rodríguez was elected president (by a 74 percent plurality). The opposition parties had not had much time to organize for the electoral contest, and control of Congress remained with the Colorado Party. The party also remained in control of the judiciary, and Rodríguez' cabinet included a number of military officers. Moreover, there was some concern over the fact that Rodríguez had never changed his active-duty military status.

Nonetheless, a new constitution went into effect on June 20, 1992, and the president adopted certain democratic measures. He declared freedom of the press, legalized all political parties, repealed a number of repressive laws, ratified the human rights treaties of the United Nations and Organization of American States, and freed the nation's remaining political prisoners. Still, much in Paraguay remained as under Stroessner. There has been no significant land reform in a nation in which a handful of the rich and a few corporations own 78 percent of the farm and ranch land; smuggling is still rampant; journalists have been arrested; and the government has interfered with unions and strikes. As under Stroessner, the Colorado Party and the army remain the key powers in Paraguay. Rodríguez' willingness to step down at the end of his term and permit free elections surprised many, and the election of Juan Carlos Wasmosy of the Colorado Party on May 9, 1993, bodes well for democracy in Paraguay.

For later developments in the history of Paraguay, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 964, 966, and 974, and the *Index*. (J.H.Wi.)

BIBLIOGRAPHY

Physical and human geography. General surveys of Paraguayan history, economics, and politics include ADOLF N. SCHUSTER, *Paraguay: Land, Volk, Geschichte, Wirtschaftsleben, und Kolonisation* (1929), on early 20th-century Paraguay; and PAUL H. LEWIS, *Paraguay under Stroessner* (1980), and *Socialism, Liberalism, and Dictatorship in Paraguay* (1982). The best bibliographies are those by DAVID LEWIS JONES, *Paraguay* (1979); and R. ANDREW NICKSON (compiler), *Paraguay* (1987); and the extended listing in R. ANDREW NICKSON, *Historical Dictionary of Paraguay*, 2nd ed., rev., enlarged, and updated (1993). Although dated, JOSEPH PINCUS, *The Economy of Paraguay* (1968), is one of the best general surveys of the economy in English. GEORGE PENDLE, *Paraguay: A Riverside Nation*, 3rd ed. (1967), is brief but informative. J.M.G. KLEINPENNING, *Man and Land in Paraguay* (1987), discusses agriculture and land tenure patterns through the Stroessner era. BRANISLAVA SUŠNIK, *Los aborígenes del Paraguay* (1978–), gives a comprehensive survey of Indian groups. RICARDO CANESE, *Itaipú y la cuestión energética en el Paraguay* (1983), is a critique of the Itaipú power project. Social conditions and culture are dealt with in RIORDAN ROETT and RICHARD SCOTT SAKS, *Paraguay: The Personalist Legacy* (1991).

History. General treatments include EFRAÍM CARDOZO, *Paraguay independiente* (1949, reissued 1987), and *Breve historiá del Paraguay* (1965, reprinted 1987); CARLOS R. CENTURIÓN, *Historia de la cultura paraguaya*, 2 vol. (1961); and HARRIS GAYLORD WARREN, *Paraguay: An Informal History* (1949, reprinted 1982), with a useful bibliography. The best treatment of the independence epoch is FULGENCIO R. MORENO, *Estudio sobre la independencia del Paraguay*, 3rd ed. edited by CARLOS SCHAUMAN (1985). Also useful is RAFAEL ELADIO VELÁZQUEZ, *El Paraguay en 1811* (1965). The Francia period is addressed by JULIO CÉSAR CHAVES, *El supremo dictador*, 4th ed. (1964); and JOHN HOYT WILLIAMS, *The Rise and Fall of the Paraguayan Republic, 1800–1870* (1979), which also deals with the two Lópezes. See also THOMAS WHIGHAM, *The Politics of River Trade: Tradition and Development in the Upper Plata, 1780–1870* (1991), with a comprehensive bibliography. The first López and his modernization of Paraguay are treated well in JUAN FRANCISCO PÉREZ ACOSTA, *Carlos Antonio López, obrero máximo, labor administrativa y constructiva* (1948); and JULIO CÉSAR CHAVES, *El Presidente López*, 2nd. ed. (1968). The War of the Triple Alliance is treated in EFRAÍM CARDOZO, *Vísperas de la guerra del Paraguay* (1954), and *Hace cien años*, 13 vol. (1967–82), a day-by-day chronicle of the war years; and HARRIS GAYLORD WARREN, *Paraguay and the Triple Alliance: The Postwar Decade, 1869–1878* (1978). The postwar era is chronicled by HARRIS GAYLORD WARREN and KATHERINE F. WARREN, *Rebirth of the Paraguayan Republic: The First Colorado Era, 1878–1904* (1985); and JOSÉ RODRIGUEZ ALCALÁ, *El Paraguay moderno* (1915). Coverage of the Chaco War can be found in DAVID H. ZOOK, JR., *The Conduct of the Chaco War* (1960); and ROBERTO QUEREJAZU CALVO, *Masamaclay*, 4th ed. (1981). Treatments of the more recent decades include MICHAEL GROW, *The Good Neighbor Policy and Authoritarianism in Paraguay: United States Economic Expansion and Great-Power Rivalry in Latin America During World War II* (1981); and the up-to-date CARLOS R. MIRANDA, *The Stroessner Era: Authoritarian Rule in Paraguay* (1990). (J.H.Wi./E.R.Sc./J.E.Pa.)

Paris

Paris is the capital of France and one of Europe's largest conurbations. The city was founded more than 2,000 years ago on an island in the Seine River, some 233 miles (375 kilometres) upstream from the river's mouth on the English Channel. The modern city has spread from the island (the Île de la Cité) and far beyond both banks of the Seine. The City of Paris itself covers an area of 41 square miles (105 square kilometres); the Greater Paris conurbation, formed of suburbs and other built-up areas, extends around it in all directions to cover approximately 890 square miles. Paris occupies a central position in the rich agricultural region known as the Paris Basin, and it constitutes one of eight *départements* of the Île-de-France administrative region. It is by far the nation's most important centre of commerce and culture.

This article is divided into the following sections:

Physical and human geography 434
- Character of the city 434
- The landscape 434
  - The city site: Paris and its river
  - Climate
- Neighbourhoods and sights 435
  - The Chaillot Palace
  - The Invalides
  - The Louvre
  - The Institute of France
  - Pont-Neuf (New Bridge)
  - Île de la Cité
  - Notre-Dame de Paris
  - City walls
  - Île Saint-Louis
  - The Bastille
  - The Marais
  - Hôtel de Ville
  - The Halles
  - Rue de Rivoli
  - The "Triumphal Way"
  - Modern business quarters
  - The ministry quarter
  - The Latin Quarter (Quartier Latin)
  - The Buttes
  - Parks and open spaces
- The people 444
- The economy 444
  - Industry
  - Commerce and finance
  - Transportation
- Administration and social conditions 445
  - Government
  - Services
  - Health
  - Education
- Cultural life 446

History 446
- Foundation and medieval growth 446
- From the Renaissance to the Revolution 447
- Evolution of the modern city 448
  - The Revolution and Napoleon I
  - Napoleon III and Haussmann
  - The Third Republic and after

Bibliography 448

Physical and human geography

CHARACTER OF THE CITY

For centuries Paris has been one of the world's most important and attractive cities. It is appreciated for the opportunities it offers for business and commerce, for study, for culture, and for entertainment: its gastronomy, haute couture, painting, literature, and intellectual community especially enjoy an enviable reputation. Its sobriquet "the City of Light" (la Ville Lumière), earned during the Enlightenment, remains appropriate, for Paris has retained its importance as a centre for education and intellectual pursuits.

Paris' site at a crossroads of both water and land routes significant to Europe as well as to France has had a continuing influence on its growth. Under Roman administration, in the 1st century BC, the original site on the Île de la Cité was designated the capital of the Parisii tribe and territory. The Frankish king Clovis took Paris from the Gauls by AD 494 and later made his capital there. Under Hugh Capet (ruled 987–996) and the Capetian dynasty the preeminence of Paris was firmly established, and Paris became the political and cultural hub as modern France took shape. France has long been a highly centralized country, and Paris has come to be identified with a powerful central state, drawing to itself much of the talent and vitality of the provinces.

The three main parts of historical Paris are defined by the Seine. At its centre is the Île de la Cité, which is the seat of religious and temporal authority (the word *cité* connotes the nucleus of the ancient city). The Seine's left bank (*rive gauche*) has traditionally been the seat of intellectual life; and its right bank (*rive droite*) contains the heart of the city's economic life, but the distinctions have become blurred in recent decades. The fusion of all these functions at the centre of France and, later, at the centre of an empire, resulted in a tremendously vital environment. In this environment, however, the emotional and intellectual climate that was created by contending powers often set the stage for great violence in both the social and political arenas (the years 1358, 1382, 1588, 1648, 1789, 1830, 1848, and 1870 being notable for such events).

In its centuries of growth Paris has for the most part retained the circular shape of the early city. Its boundaries have spread outward to engulf the surrounding towns (*bourgs*), usually built around monasteries or churches and often the site of a market. From the mid-14th to the mid-16th century the city's growth was mainly eastward; since then it has been westward. It comprises 20 *arrondissements* (municipal districts), each of which has its own mayor, town hall, and particular features. The numbering begins in the heart of Paris and continues in the spiraling shape of a snail shell, ending to the far east. Parisians refer to the *arrondissements* by number as the first (*premier*), second (*deuxième*), third (*troisième*), and so on. Adaptation to the problems of urbanization, such as immigration, housing, social infrastructure, public utilities, suburban development, and zoning, has produced the vast urban agglomeration. (Ed.)

THE LANDSCAPE

The city site: Paris and its river. Paris' city proper is small; no corner is farther than six miles from the square in front of Notre-Dame cathedral. The city occupies a depression hollowed out by the Seine, and the surrounding heights have been respected as the limits of the city. Elevation varies from 430 feet (130 metres) at Montmartre to 85 feet at Grenelle. The river flows for eight miles through the centre of the city and 10 of the 20 *arrondissements*. It enters the city at the southeast corner, flows northwestward, and turns gradually southwestward, eventually leaving Paris at the southwest corner. As a result, what starts out as the stream's east bank becomes its north bank and ends as the west bank, and the Parisians



The Eiffel Tower, from the Centre Beaubourg on the Right Bank, looking westward over the rooftops of the city.
Nancy Cohen—Black Star

therefore adopted the simple, unchanging designation of Right and Left Bank (when facing downstream). Specific places are usually indicated by *arrondissement* or by quarter (*quartier*).

The river bank

At water level, some 30 feet below street level, the river is bordered—at least on those portions not transformed into expressways—by cobbled quays graced with trees and shrubs. From street level another line of trees leans toward the water. Between the two levels, the retaining walls, usually made of massive stone blocks, are decorated with the great iron rings once used to moor merchant vessels, and some are pierced by openings left by water gates for old palaces or inspection ports for subways, sewers, and underpasses. At intermittent points the walls are shawled in ivy.

The city is ringed with great forests of beech and oak; they are called the "lungs of Paris," for they help to purify the air in the heavily industrialized region. Paris' position at the centre of the Île-de-France region, bordered by the Seine, Oise, and Marne rivers, has long made it the dominant settlement.

Climate. In its location on the western side of Europe and in a plain relatively close to the sea, Paris benefits from the balmy influences of the Gulf Stream and has a fairly temperate climate. The weather can be very changeable, however, especially in winter and spring, when the wind can be sharp and cold. The annual average temperature is 53° F (12° C); the July average is about 67° F (19° C), and the January average is about 38° F (3° C). The temperature drops below freezing about 35 days a year, and snow falls on an average of 15 days. Since the 1950s the city has taken measures to decrease air pollution, and a system of water purification has made tap water safe for drinking.

NEIGHBOURHOODS AND SIGHTS

Paris' many old buildings, monuments, gardens, plazas, boulevards, and bridges compose one of the world's grandest cityscapes. An impressive spot from which to view the city is the Chaillot Palace, which stands on a rise on the Right Bank of the Seine to the west, where the river begins its southwestward curve.

The Chaillot Palace. The Chaillot Palace dates from the International Exposition of 1937. It replaced the Trocadéro Palace, a structure left over from the 1878 International Exposition. The Chaillot Palace is made up of two separate pavilions, from each of which extends a curved wing. The Musée de l'Homme, the Naval Museum, the Museum of French Monuments, and the Cinema Museum are located there. Under the terrace that separates the two sections are two theatres, the National Theatre of Chaillot and a small hall that serves as one of the two motion-picture houses of the national film library.

The terrace, which is lined by statues, gives a splendid view across Paris. The slope descending to the river has been made into a terraced park, the centre of which is alive with fountains, cascades, and pools. The Trocadéro Aquarium is in a grotto a few steps away in the park.

The Eiffel Tower

From the bottom of the slope the five-arched Jena Bridge (Pont d'Iéna) leads across the river. It was built for Napoleon in 1813 to commemorate his victory at the Battle of Jena in 1806. On the Left Bank rises the Eiffel Tower, an unclad metal truss tower designed by Gustave Eiffel. The tower was built for the International Exposition of 1889, against the strident opposition of national figures who thought it unsafe or ugly or both. When the exposition concession expired in 1909, the 984-foot (300-metre) tower was to have been demolished, but its value as an antenna for radio transmission saved it. Additions made for television transmission have added 56 feet to the height. From the topmost of the three platforms the view extends for more than 40 miles.

From the two-acre base of the tower the Champ-de-Mars ("Field of Mars"), an immense field, stretches to



The city of Paris and (inset) its metropolitan area.

the Military Academy (École Militaire), which was built from 1769 to 1772 and is still used by the War College (École Supérieure de Guerre). The Champ-de-Mars, which originally served as the school's parade ground, was the scene of two vast rallies during the French Revolution: that of the Federation (1790) and that of the Supreme Being (1794). From 1798 there were annual national expositions of crafts and manufactures, which were followed by world's fairs between 1855 and 1900.

Behind the Military Academy stands the headquarters of the United Nations Educational, Scientific and Cultural Organization (UNESCO). The building, erected in 1958, was designed by an international trio of architects and decorated by artists of member nations.

The Invalides. One street to the northeast is the Hôtel des Invalides, founded by Louis XIV to shelter 7,000 aged or invalid veterans. The enormous range of buildings was completed in five years (1671–76). The gold-plated dome (1675–1706) that rises above the hospital buildings belongs to the church of Saint-Louis. The dome was designed by Jules Hardouin-Mansart, who employed a style known in France as "Jesuit" because it derives from the Jesuits' first church in Rome, built in 1568. The churches of the French Academy (Académie Française), the Val-de-Grâce Hospital, and the Sorbonne, as well as three others in Paris, all of the 17th century, followed this style. By using the classical elements more freely than had been done in Rome, the French made it something recognizably Parisian.

In the chapels of Saint-Louis are the tombs of Napoleon's brothers Joseph and Jérôme, of his son (whose body was returned from Vienna in 1940 by Adolf Hitler), and of the marshals of France. Immediately beneath the cupola is a red porphyry sarcophagus that covers the six coffins, one inside the other, enclosing the remains of Napoleon, which were returned from St. Helena in 1840 through the efforts of King Louis-Philippe. Napoleon's uniforms, personal arms, and deathbed are displayed in the Army Museum at the front of the Invalides. A portion of the Invalides still serves as a military hospital; facilities have been modernized since World War II.

Napoleon's tomb

The vast, tree-lined Invalides Esplanade slopes gently to the Quai d'Orsay and the Alexandre III Bridge. The first stone for the bridge, which commemorates the Russian tsar Alexander III, was laid in 1897 by Alexander's son, Tsar Nicholas II. The bridge was finished in time for the International Exposition of 1900, and it leads to two other souvenirs of that year's fair, the Grand Palais and the Petit Palais. The buildings are still used for annual shows and for major visiting art exhibits.

The Louvre. Vikings camped on the Right Bank across from the western tip of the Île de la Cité in their unsuccessful siege of Paris in 885, and in about 1200 King Philip II had a square crusader's castle built on the same site, just outside the new city wall, to buttress the western defenses. Over the following centuries many additions and renovations were made, and from the castle grew one of the world's largest palaces, completed only in 1852. From the original square, known as the Cour Carrée ("Square Court"), two galleries extend westward for 1,640 feet, one along the river, the other along the rue de Rivoli. In 1871, only 19 years after the huge oblong was completed, its western face, the Tuileries Palace (begun 1563), was destroyed by the insurrectionists of the Commune.

The Cour Carrée

Two of the facades of the Cour Carrée had strong influence on French architecture. Pierre Lescot began his inner courtyard facade in 1546, adapting the Renaissance rhythms and orders he had observed in Italy and adding purely French decoration to the classical motifs. The physician and architect Claude Perrault collaborated with Louis Le Vau, architect to the king, to design the outer east face of the palace in 1673. It, too, employs classic elements, making especially graceful use of coupled columns and a pediment.

The Louvre Museum occupies the four sides of the palace around the Cour Carrée as well as the south gallery, which stretches along the river. Among the treasures of the museum are the Victory of Samothrace, the Venus de Milo, and the Mona Lisa. The enormous collections contain works from the 7th century BC to the mid-19th century, with a huge cultural and geographic spread. The north gallery, along the rue de Rivoli, houses a separate museum, the Museum of Decorative Arts, as well as the Ministry of Finance.

Extensive remodeling has been undertaken throughout the Louvre to increase space for art works. Construction began in the early 1980s to create a new main entrance and underground reception hall in the vast Napoleon Courtyard, between the two galleries; the 70-foot-high glass pyramid designed by I.M. Pei to cover the entrance aroused both strong support and spirited criticism.

The Arts Bridge (Pont des Arts), which crosses the Seine from the Louvre to the Left Bank, is one of the most charming of all the Parisian bridges. It was the first (1803) to be made of iron, and it has always been reserved for pedestrians: it provides an intimate view of riverside Paris and of the Seine itself.

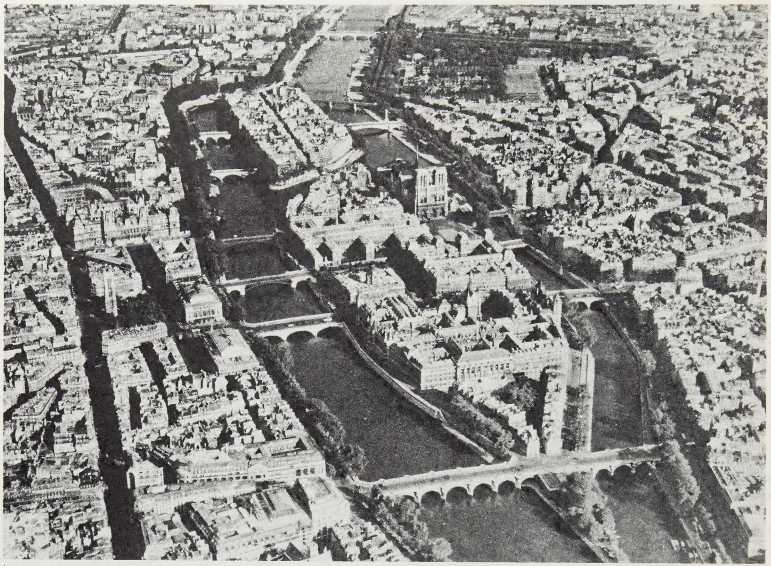
The Institute of France. At the point where the Arts Bridge meets the Left Bank stands the Institute of France, which since 1806 has housed the five French academies. The site was originally occupied by the Nesle Tower (Tour de Nesle), a defense work for the Left Bank terminus of the city wall of 1220. Louis Le Vau designed the additional buildings in 1663 to house the College of the Four Nations (Collège des Quatre-Nations), paid for by a legacy from Louis XIV's minister Cardinal Mazarin, who had brought the four entities in question—Piedmont, Alsace, Artois, and Roussillon—under the French crown. Le Vau based his designs on Italian models. The five contemporary academies are the French Academy (Académie Française), founded by Cardinal de Richelieu in 1635, which edits the official French dictionary, awards literary prizes, and has a membership of "40 Immortals"; the Academy of Inscriptions and Belles Lettres, founded in 1663 by Louis XIV's finance minister, Jean-Baptiste Colbert; the Academy of Sciences, founded in 1666, also by Colbert; the Academy of the Fine Arts, two sections formed at different times by Mazarin and Colbert and joined in 1795; and the Academy of Ethics and Political Science, created by the Convention (a governing body during the French Revolution) in 1795 to ponder questions of philosophy, economics, politics, law, and history.

The French Academy

Almost next door is the Mint (Hôtel des Monnaies). In this sober late 18th-century building visitors may watch coins being struck and tour a museum of coins and medals.

Pont-Neuf (New Bridge). Despite its name, the Pont-Neuf, which was built from 1578 to 1604, is the oldest of the Paris bridges (others predate it but have been rebuilt). Its sturdiness has become axiomatic: Parisians still say, "Solid as the Pont-Neuf." The bridge is supported in the middle by the tip of the Île de la Cité, to which it extends five arches from the Left Bank and seven from the Right. The parapet corbels are decorated with more than 250 different grotesque masks. The parapet curves out toward the water at each bridge pier, forming half-moon bays along what was the first sidewalk in Paris; in these bays street vendors set up shop. For 200 years this bridge was the main street and the perpetual fair of Paris. The custom of holding summer fairs on the bridge has been revived to some extent. Although the structure undergoes regular repair, in the main Pont-Neuf as it exists today is the original bridge.

Île de la Cité. Downstream and just below the bridge is the tip of the Île de la Cité, fashioned into a triangular gravel-pathed park bordered by flowering bushes, with rustic benches under the ancient trees. It is surrounded by a wide cobbled quay that is especially popular with sunbathers and lovers. Where the steps come onto the bridge from the park there is a bronze equestrian statue of King Henry IV, who insisted on completion of the Pont-Neuf. The statue is an 1818 reproduction of the 1614 original, which was the first statue to stand on a public way in Paris. Opposite is the narrow entrance to the Place Dauphine (1607), named for Henry's heir, the future Louis XIII. The *place* was formerly a triangle of uniform red-brick houses pointed in white stone, but the row of houses along its base was ripped out in 1871 to make room for construction of part of the Palace of Justice.



Île de la Cité and Île Saint-Louis, on the Seine River at the heart of Paris.
By courtesy of the French Government Tourist Office

The ship-shaped Île de la Cité is 10 streets long and five wide. Eight bridges link it to the riverbanks and a ninth leads to the scow-shaped Île Saint-Louis, which lies to the southeast.

The palace of the early Roman governor (now the Palace of Justice) was rebuilt on the same site by Louis IX (St. Louis) in the 13th century and enlarged 100 years later by Philip IV the Fair, who added the grim, gray-turreted Conciergerie, with its impressive Gothic chambers. The Great Hall, which, under the kings, was the meeting place of the Parlement (the high court of justice), was known throughout Europe for its Gothic beauty. Fires in 1618 and 1871 destroyed much of the original room, however, and most of the rest of the palace was devastated by flames in 1776. The Great Hall now serves as a waiting room for the courts, in one of which, the adjoining first Civil Chamber, the Revolutionary Tribunal sat from 1793, condemning 2,600 persons to the guillotine. After sentencing, the victims were taken back down the stone stairs to the dungeons of the Conciergerie to await the tumbrils. The Conciergerie still stands and is open to visitors.

In the palace courtyards is found one of the great monuments of France, the 13th-century Sainte-Chapelle ("Holy Chapel"). Built at Louis IX's direction between 1243 and 1248, it is a masterpiece of Gothic Rayonnant style. With great daring, the architect (possibly Pierre de Montreuil) poised his vaulted ceilings on a trellis of slender columns, the walls between being made of stained glass. The exquisite chapel was designed to hold the Crown of Thorns, thought to be the very one worn by Jesus at his crucifixion. Louis IX had purchased the relic from the Venetians, who held it in pawn from Baldwin, the Latin king of Byzantium. Other holy relics, such as nails and pieces of wood from the True Cross, were added to the chapel's collection, the remnants of which are now in the treasury of Notre-Dame.

Under King Louis-Philippe (1830–48), the "sanitization" of the island was begun, and it was continued for his successor, Napoleon III, by Baron Georges Haussmann. The project involved a mass clearing of antiquated structures, widening of streets and squares, and the erection of massive new government offices, including parts of the Palace of Justice. The portion of the palace that borders the Quai des Orfèvres—formerly the goldsmiths' and silversmiths' quay—became the headquarters of the Paris municipal detective force, the Police Judiciaire ("Judicial Police"), which keeps a small museum on the fourth floor.

Across the boulevard du Palais is the Police Prefecture, another 19th-century structure. On the far side of the prefecture is the Place du Parvis-Notre-Dame, an open space enlarged six times by Haussmann, who also moved the Hôtel-Dieu, the first hospital in Paris, from the riverside to the inland side of the square. Its present buildings date from 1868.

Notre-Dame de Paris. At the east end of the square is the cathedral of Notre-Dame de Paris, which is situated on a spot that Parisians have always reserved to the practice of religious rites. The Gallo-Roman boatmen of the *cité* erected their altar to Jupiter there (it is now in the Cluny Museum), and, when Christianity was established, a church was built on the temple site. The first bishop of Paris, St. Denis, became its patron saint. The red in the colours of Paris represents the blood of this martyr who, in popular legend, after decapitation, picked up his head and walked.

When Maurice de Sully became bishop in 1159 he decided to replace the decrepit cathedral of Saint-Étienne and the 6th-century Notre-Dame with a church in the new Gothic style. The style was conceived in France, and a new structural development, the flying buttress, which added to the beauty of the exterior and permitted interior columns to soar to new heights, was introduced in the building of Notre-Dame. Construction began in 1163 and continued until 1345.

Restoration of Notre-Dame

After being damaged during the French Revolution, the church was sold at auction to a building-materials merchant. Napoleon came to power in time to annul the sale, and he ordered that the edifice be redecorated for his coronation as emperor in 1804. Louis-Philippe later initiated restoration of the neglected church. The architect Eugène Viollet-le-Duc worked from 1845 to 1864 to restore the monument. Like all cathedrals in France, Notre-Dame is the property of the state, although its operation as a religious institution is left entirely to the Roman Catholic Church.

A few 16th- and 17th-century buildings survive north of the cathedral. They are what remains of the Cloister of the Cathedral Chapter, whose school was famous long before the new cathedral was built. Early in the 12th century, one of its theologians, Peter Abelard, left the cloister with his disciples, crossed to the Left Bank, and set up an independent school in the open air in the Convent of the Paraclete near the present Place Maubert. After a prolonged struggle with the monks of Saint-Denis the followers of Abelard in 1200 won the right, from both the king and the pope, to form and govern their own community. This was the beginning of the University of Paris.

City walls. After the Roman town on the Left Bank was sacked by barbarians in the 3rd century AD, the fire-blackened stones were freighted across to the Île de la Cité, where a defensive wall was constructed. Neglected in times of peace, it was rebuilt several times over the course of the centuries. The earliest of the bridges to the Left Bank, the Petit Pont (which has been rebuilt several times), was guarded by a fortified gate, the Petit Châtelet, and the bridge to the Right Bank, the Pont au Change, was guarded by the Grand Châtelet, which served as a fort, prison, torture chamber, and morgue until it was demolished in 1801.

From 1180 to 1225 Philip II built a new wall that protected the settlements on both banks. In 1367–70 the Right Bank enclosure was enlarged by Charles V, with the massive Bastille protecting the eastern approaches as the Louvre protected the west. In 1670 Louis XIV had the Charles V extensions transformed into the tree-planted Grands Boulevards, embellished at the Saint-Denis Gate (Porte Saint-Denis) and the Saint-Antoine Gate (Porte Saint-Antoine) with triumphal arches that still stand. (The word *boulevard,* related to "bulwark," originally was a military engineering term for the platform of a defensive wall.)

A century later a new wall was begun. The wall was built with 57 tollhouses to enable the farmers-general, a company of tax "farmers," or collectors, to collect customs duties on goods entering Paris. The tollhouses are still standing at Place Denfert-Rochereau.

The last wall

The last wall, built by Adolphe Thiers for Louis-Philippe, was a genuine military installation with outlying forts. By the time it was finished it enclosed a number of hamlets outside Paris, among them Auteuil, Passy, Montmartre, La Villette, and Belleville. Post-Revolutionary rebuilding and post-Napoleonic economic recovery, with expansion of employment because of the Industrial Revolution, drew more and more people to Paris, and—as railways developed—with ever-increasing facility. Between 1852 and 1870 Haussmann razed the walls of the farmers-general and built the outer boulevards. The 1845 walls were finally knocked down, and the boulevards were extended in 1925.

Île Saint-Louis. In 1627 Louis XIII granted a 60-year lease on two mudbanks behind the Île de la Cité to a contractor, Christophe Marie, and two financiers. It was 37 years before Marie was able to unite the islets, dike the circumference, lay out a central avenue with 10 lateral streets, and rent space to householders. The church of Saint-Louis-en-l'Île was begun the same year, 1664, but one of the finest houses, by Louis Le Vau, had been completed as early as 1640. Another, the Hôtel de Lauzun, a few yards upstream on the Quai d'Anjou, was completed in 1657. The Marie Bridge to the Right Bank, which was completed as part of the contract, is the original span, although it has been modified for modern traffic. The Île Saint-Louis constitutes a tranquil neighbourhood in the centre of the busy city.

The Bastille. The road off the upper end of the Île Saint-Louis leads to the Place de la Bastille on the Right Bank. From the river to the *place* runs a canal, the Arsenal Basin, which formerly supplied water to the moat around the Bastille. At the Place de la Bastille the waterway goes underground for almost a mile, then emerges to form the Saint-Martin Canal, which, with its bridges and locks, and its barges sailing slowly down the centre of city streets, constitutes one of the least known and most picturesque sections of Paris.

The capture of the Bastille on July 14, 1789, was a symbolic blow at tyranny rather than an act of liberation for tyranny's victims. The prison, which had been virtually unused for years and was scheduled for demolition by the monarchy, held on that day only four counterfeiters, two madmen, and a young aristocrat who had displeased his father. The fortress was demolished after its capture.

Napoleon had the *place* laid out in 1803. A railway station was built there in 1859 and was razed in 1984 to allow construction of a new opera house (begun two years later).

The neighbourhood between the Bastille and the Place de la Nation, eastward along the rue du Faubourg Saint-Antoine, has been one of skilled craftsmen since the mid-15th century, when the self-governing royal abbey gave space within its wide domains to those cabinetmakers who refused to abide by the restrictions of Paris guilds as to styles and types of wood to be used. This neighbourhood was always among the first to revolt when revolution was in the air and was noted for the speed with which it raised barricades of impressive height. The character of the area has changed, however, as most of the small workshops have closed.

The Marais. To the west of the Bastille lies a triangular area with its base along the river up to the Hôtel de Ville (City Hall) and its apex just short of the Place de la République to the north. It keeps its name—*le marais* ("the marsh")—from the Middle Ages, and, because it became the market garden of Paris, it gave its name to all market gardens in France (*culture maraîchère*). Extension of the city walls along the Right Bank led to diking of the shore and drainage of the soil. In 1107 the Knights Templar established their Paris Temple, a vast fortified enclosure, at the top of the triangle. In 1360 Charles V moved into his new royal residence in the lower right-hand corner, where the rue des Lions marks the former location of the menageries.

Charles VII preferred to live just behind the Bastille, in the Hôtel des Tournelles, which Henry II had enlarged and beautified by Philibert Delorme in 1550. Great nobles, such as the dukes of Guise and Lorraine, followed the King and had palaces built in the vicinity. When Henry II was killed in a joust on the rue Saint-Antoine in 1559, his widow, Catherine de Médicis, had the Tournelles razed. On the site in 1607 construction began on the first residential square to be designed in Paris. Henry IV had it built and reserved a house there for himself. The three-story houses are made of red brick with white-stone quoins (solid-corner angles) and window surrounds, and the ground floors form arcades over the sidewalks. The square was named Place Royale, but since 1800 it has been called Place des Vosges. Another wave of building by the rich, anxious to be close to a royal project, endowed the Marais with 200 more private palaces.

The first residential square

In 1792 the Knights Hospitaller of St. John of Jerusalem (also known as the Knights of Malta) were turned out of the Temple, which had been given to them in 1313 when the Templar order was dissolved. On Aug. 13, 1792, the royal family was incarcerated in the Temple's tower keep; the King was taken off to his death on Jan. 21, 1793, and the Queen was removed to the Conciergerie on August 2 (and executed on October 16). The Temple's tower was leveled in 1808 to discourage rallies there by royalists.

After the 17th-century construction boom the Marais remained virtually untouched. Three-quarters of the buildings there date from before 1870. Toward the end of the 19th century, while some of the oldest and most imposing of the palaces were being demolished by private developers, other owners managed to restore a few mansions, and the nation and the municipality also restored a handful of

fine buildings. Many Jewish refugees from eastern Europe settled in the district. Scores of fine houses were subdivided into tiny apartments to profit from the poverty-stricken newcomers, and workshops were installed on the lower floors and in courtyard sheds. Through centuries of neglect the Marais became one of the worst slums in Paris.

In 1969 the Municipal Council approved an urban renewal scheme for ending slum conditions while preserving the workaday life and animation and restoring the undeniable beauty of the quarter. The scheme has been very successful, and property prices in the Marais have soared.

Among the restored ancient buildings open to the public are the Museum of the History of Paris (Musée Carnavalet), built in 1545 and enlarged by François Mansart in 1645; the Museum of the History of France (National Archives, Hôtel de Soubise), parts of which date from 1375, 1553, and 1704–15; the Museum of Nature and the Hunt (Fondation Somer, Hôtel Guénégaud), built by François Mansart in 1648–51; the National Bureau of Historic Monuments (Caisse Nationale des Monuments Historiques, Hôtel de Sully) by Jean I Androuet du Cerceau; and the Picasso Museum (Hôtel Salé).

Closer to the Hôtel de Ville is the Gothic Hôtel de Sens, built at the end of the 15th century for the bishops of Sens, then also bishops of Paris. It was restored after 40 years of work and now serves as a city library of specialized collections. Nearby, behind facades of a much later date, two half-timbered medieval houses have been uncovered. In the Marais, portions of the 13th-century city wall, including one of the watchtowers, are still to be seen.

Hôtel de Ville. Three city halls have stood on this site, each grander than its predecessor. The present building (1874–82) replaces the Renaissance structure that was in use from the 16th century until it was burned by the insurrectionary Communards in 1871. The first was the House of Pillars (Maison aux Piliers), used by the municipality from 1357 to 1533. The present building contains the official apartments of the mayor of Paris.

The massacre of thousands of Protestants on St. Bartholomew's Day, 1572; the Day of the Barricades, May 12, 1588, when the Catholic League drove Henry III out of Paris; and the civil war known as the Fronde, 1648–53, all involved the Paris masses, but they were palace revolutions rather than popular uprisings. When the people rose in their own cause, the focus of their action was inevitably the Hôtel de Ville, as was indeed the case in 1789, 1830, 1848, 1871, and 1944.

In July 1789, for example, after killing the provost, the revolutionary mob took the Hôtel de Ville as it had already taken the Invalides and the Bastille. Three days later Louis XVI appeared on the balcony, wearing the tricolour cockade (royal white added to municipal red and blue), and was cheered by the crowd. Four years later the building was taken as headquarters for the Commune, which directed mob action to control the Convention. On 9 Thermidor, year II (July 27, 1794), the Convention's guards entered the Hôtel de Ville and seized Robespierre and his Communards; all were executed soon after.

The Hôtel de Ville as a centre of revolution

In August 1830 Louis-Philippe appeared on the balcony and was acclaimed by the revolutionary crowd. The name of the square was changed in that year to Place de l'Hôtel de Ville. As the Place de Grève (*grève* meaning "strand," or "bank") it had been the principal port of Paris for centuries. (The refusal of boatmen to put out gave the French worker the phrase for going on strike: *faire la grève.*) From 1310 to 1832 it was Paris' principal place of execution.

In 1871, after Napoleon III's defeat at Sedan, a new republic was declared from the steps of the Hôtel de Ville, but, when the national government, in its turn, capitulated, Parisians refused to accept defeat and in March formed the Commune of Paris. In May troops of the government entered the town and fought sharp engagements with the Communards, who set fire to the Hôtel de Ville, the Tuileries Palace, the Palace of Justice, the Police Prefecture, the Arsenal, and other government buildings. The Communards executed 67 hostages, including the archbishop of Paris. The government subsequently rounded up and executed 20,000 Parisians, shipping thousands of others to penal colonies.

In 1944, as the city was being liberated from German occupation, the National Council of Resistance made the Hôtel de Ville its headquarters. At the climax of the Liberation General Charles de Gaulle appeared on the balcony and was acclaimed by the crowd.

The Halles. On the rue de Rivoli westward from the Hôtel de Ville is the Saint-Jacques bell tower (Tour Saint-Jacques), all that remains of a church in the Flamboyant Gothic style that was torn down in 1797. A few streets farther northwest toward the Louvre is the quarter of the Halles, from 1183 to 1969 the central market of Paris. When the markets moved out to a new location at Rungis, near Orly, the distinctive 19th-century iron-and-glass market halls—10 originals designed by Victor Baltard and built between 1854 and 1866, and two 1936 reproductions—and their neighbourhood were designated for renewal. The renewal projects were delayed for several years, however, by bitter disagreements over how the area—which had been thought of as the heart of Paris—should be used. The old market halls were used temporarily for exhibitions and cultural events, but in 1971 they were torn down. Their demolition left an enormous hole in the ground that became a symbol to many Parisians of the end of an era. Construction at the site began in 1971, and in 1977 a station linking the city's subway system (the Métro) with the regional express system was opened. The Forum des Halles, a partly subterranean, multistoried commercial and shopping centre, was opened in 1979. The nearby streets were converted to a traffic-free zone for pedestrians, and a small formal garden was laid out.

The markets of the Halles

A few of the neighbourhood's old houses have been renovated or restored to keep some of the old flavour, but many buildings have been demolished and the area has been largely transformed. Many of its poorer residents were forced to move as renovation and improvement made the neighbourhood unaffordable. The church of Saint-Eustache (1532–1637) remains, as does the circular Grain Exchange (Halle au Blé; 1811–13), which was burned by the Commune and restored in the 1880s as the Commercial Exchange (Bourse de Commerce).

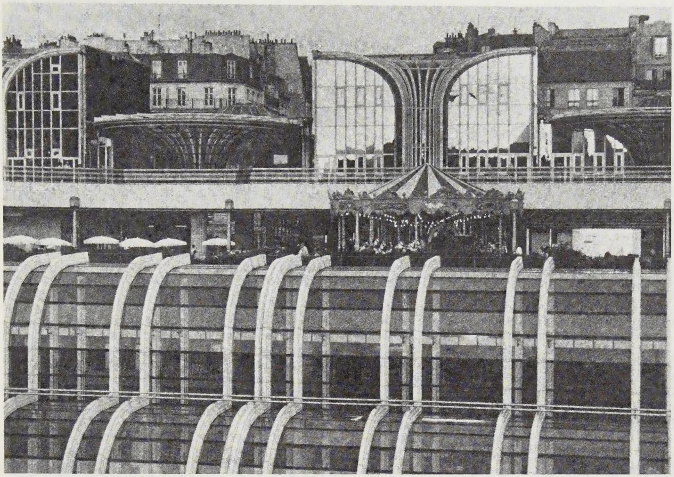
Another development in the area is the Georges Pompidou National Centre for Art and Culture, popularly called Centre Beaubourg, a vast glass-and-metal structure of distinctive design inaugurated in 1977. It soon proved its popularity and remains a successful attraction for Parisians and tourists alike. The centre houses the National Museum of Modern Art, temporary exhibits, the multimedia Public Information Library, the Industrial Design Centre, the Institute for Acoustic and Musical Research, and workshops for children.

Rue de Rivoli. The Louvre and the Tuileries Gardens take up the south side of this street, and on the other side runs an arcade more than a mile long. Opposite the middle of the Louvre, the Place du Palais-Royal leads to the palace of Cardinal de Richelieu, which he willed to the royal family. Louis XIV lived there as a child, and during the minority of Louis XV the kingdom was ruled from there by the debauched but gifted regent Philippe II, duc d'Orléans from 1715 to 1723. Late in the 18th century Louis-Philippe d'Orléans, who was popularly renamed Philippe-Egalité during the Revolution for his radical opinions, undertook extensive building around the palace garden. It was a commercial operation, and the Prince hoped to pay his debts from the property rents. Around the garden he built a beautiful oblong of colonnaded galleries and at each end of the gallery farthest from his residence a theatre. The larger playhouse has been the home of the Comédie-Française, the state theatre company, since Napoleon's reign. The princely apartments now shelter high state bodies such as the Council of State.

Just behind the courtyard is the Bibliothèque Nationale (National Library), the national library of deposit, with an enormous collection of books and prints.

Haussmann greatly enlarged the Place du Palais-Royal in 1852, and he was careful to preserve the palace when he laid out the avenue de l'Opéra. At the top of the avenue, where the Grands Boulevards crossed an enormous new *place,* the new opera house was built from 1825 to 1898. The Paris Opera House, a splendid monument

Garnier's Opera



Portion of the multistoried Forum des Halles commercial and shopping centre, with glass-enclosed subterranean stories, a café and carousel at street level, and an elevated promenade.
Nancy Cohen—Black Star

to the Second Empire, was designed in the neo-Baroque style by Charles Garnier. It is known especially for its decorative embellishments, chief among them the Grand Staircase. Just behind the Opera House are various large department stores.

The next *place* along the rue de Rivoli is the Place des Pyramides. The gilded equestrian statue of Joan of Arc stands not far from where she was wounded at the Saint-Honoré Gate (Porte Saint-Honoré) in her unsuccessful attack on Paris (at that time held by the English), on Sept. 8, 1429.

Farther along toward the Place de la Concorde the rue de Castiglione leads to the Place Vendôme, an elegant octagonal *place*, little changed from the 1698 designs of Jules Hardouin-Mansart. In the centre, the Vendôme Column bears a statue of Napoleon. It was pulled down during the Commune and put back up under the Third Republic (1871–1940). The *place* and the rue de la Paix have lost none of their discreet distinction, nor have their shops.

The "Triumphal Way." From the Arc de Triomphe du Carrousel, in the courtyard between the open arms of the Louvre, extends one of the most remarkable perspectives to be seen in any modern city. It is sometimes called la Voie Triomphale ("the Triumphal Way"). From the middle of the Carrousel arch the line of sight runs the length of the Tuileries Gardens, lines up on the obelisk in the Place de la Concorde, and goes up the Champs-Élysées to the centre of the Arc de Triomphe and beyond to the skyscrapers of La Défense, in the western suburbs.

The Louvre's modest triumphal arch stands in the open space where costumed nobles performed in an equestrian display—*carrousel*—to celebrate the Dauphin's birth in 1662. The design of the arch, an imitation of that of the Arch of Septimius Severus in Rome, was conceived in 1808 by Charles Percier and Pierre Fontaine. The flanks of the Carrousel arch are incised with a record of Napoleon's victories.

The Tuileries Gardens

The Tuileries Gardens, which fronted the Tuileries Palace (looted and burned in 1871 during the Commune), have not altered much since André Le Nôtre redesigned them in 1664. Le Nôtre was born and died in the gardener's cottage in the Tuileries; he succeeded his father there as master gardener. His design carried the line of the central allée beyond the gardens and out into the countryside by tracing a path straight along the wooded hill west of the palace. On this hilltop, 170 years later, the Arc de Triomphe was erected.

At the western edge of the gardens, Napoleon III erected a hothouse, known as the Orangery, and the Jeu de Paume, an indoor court for court tennis. Both eventually were adapted as museums: the Orangery had a small permanent collection, including a group of 19 of Claude Monet's paintings of waterlilies displayed as panoramas. The Jeu de Paume housed the Louvre's collection of paintings by the Impressionists and their forerunners. The collections of the two museums—with the exception of the Monet panoramas—were moved to the Orsay Museum, which opened in 1986, and the Jeu de Paume and the Orangery were then reserved for occasional exhibitions.

The formal exit gate from the Tuileries is flanked by two winged horses (17th century), and the entrance to the Champs-Élysées across the square is similarly adorned, only by earthbound horses (18th century), both pairs having been removed in turn from the water trough at the Château de Marly.

The Place de la Concorde was designed as a moated octagon in 1755 by Jacques-Ange Gabriel. The river end was left open, and on the inland side two matching buildings were planned. The ground floor was arcaded and the facade was nimbly adapted from the Louvre colonnade, all with a refinement typical of the era. Although Gabriel built eight giant pedestals around the periphery of his *place*, they remained untenanted until Louis-Philippe gave them statues representing provincial capitals. Viewed clockwise starting from the Navy Ministry (Ministère de la Marine), the statues are Lille, Strasbourg, Lyon, Marseille, Bordeaux, Nantes, Brest, and Rouen. Louis-Philippe also had the Luxor Obelisk, a gift from Egypt, installed in the centre and flanked by two fountains. Later, the surrounding moat was filled in.

Louis XVI was decapitated near the statue of Brest on Jan. 21, 1793. Four months later the guillotine was erected

near the gates of the Tuileries, and the executions continued there for nearly three years, accounting for a total of 1,343 deaths.

The Madeleine

Between the twin buildings on the northeastern side of the *place,* the broad rue Royale mounts to the Church of Sainte-Marie-Madeleine, consecrated in 1842. The church is a stern oblong, fenced with columns 60 feet high. Its design, supposedly that of a Greek temple, is actually closer to the Roman notion of Greek.

The Place de la Madeleine is the western terminus of the Grands Boulevards, which imitate the arch of the river from there north and east to the Place de la République and the Bastille. The glittering chic of the Grands Boulevards, which flavoured Paris life from 1750 to the 1880s, disappeared with the *boulevardiers,* but many of the boulevards' theatres and other entertainments survive. The state-run Opéra-Comique ("Comic Opera") stands fast just off the boulevard des Italiens, the wax museum survives on the boulevard Montmartre, and, a few doors away, the Théâtre des Variétés, founded under the Second Empire by the composer Jacques Offenbach, still operates. The Théâtre de la Renaissance, where the actor Benoît-Constant Coquelin created the role of Cyrano de Bergerac in 1897, remains on the boulevard Saint-Martin, as does the Théâtre de l'Ambigu, where Frédéric Lemaître, the celebrated actor in boulevard melodrama, thrilled all Paris in the mid-19th century. Some of the movie palaces of the 1930s also remain along the boulevards.

To the west off the rue Royale runs the rue du Faubourg Saint-Honoré, which, in addition to the British Embassy and the Élysée Palace (residence of the French president), has on its shop windows some of the most prestigious names in the Paris fashion trade.

Along the first 2,500 feet of the Champs-Élysées, between Concorde and the Rond-Point des Champs-Élysées (a roundabout, or traffic circle), little has changed for a century: the avenue is bordered with chestnut trees, behind which on both sides are gardens, usually full of children at play. The pavilions in the gardens are used as tearooms, restaurants, and theatres.

The Champs-Élysées today

From the Rond-Point up to the Arc de Triomphe, however, the avenue has changed with the times. Under the Second Empire this was a street of luxurious town houses. They were supplanted by cafés, nightclubs, luxury shops, and cinemas, but the street retained its feeling of luxury, and the tree-shaded sidewalks (wide as a normal street) offered promenades that were the pride of Paris. Since the 1950s, however, banks, automobile showrooms, airline offices, and fast-food eateries have taken over much of the space. Automobiles park on the sidewalks and in many places one rank of the double file of trees has been felled.

At the top of the Champs-Élysées is a circular *place* from which 12 imposing avenues radiate to form a star (*étoile*). It was called Place de l'Étoile from 1753 until 1970, when it was renamed Place Charles de Gaulle. In the centre of the *place* is the Arc de Triomphe, commissioned by Napoleon in 1806. It is twice as high and as wide as the Arch of Constantine, in Rome, which inspired it. Jean Chalgrin was the architect and François Rude sculpted the frieze and the spirited group, "The Departure of the Volunteers of 1792" (called "La Marseillaise"). On Armistice Day in 1920, the Unknown Soldier was buried under the centre of the arch, and each evening the flame of remembrance is rekindled by a different patriotic group.

In the 1970s the largest concentration of tall buildings in Europe arose some two miles beyond the arch, on the far side of the suburban wedge of Neuilly. The site, called the Quartier de la Défense, was formerly just a place on the road adjoined by the suburban municipalities of Puteaux, Courbevoie, and Nanterre. Today tall office buildings, heated and air conditioned from a central plant, are the hub of the complex. The "ground level" between buildings is a raised platform reserved to pedestrians, with roads and parking below. There are shops, restaurants, cafés, a shopping centre, hotels, and apartment houses. Before the project was begun, the state had already constructed at La Défense its National Centre for Industries and Technologies, a large exhibition hall. Nanterre became the site of a campus of the University of Paris in the 1960s, and specialized schools at the university level have been installed in the new centre: the National School of Architecture, the National School of Decorative Arts, the National Conservatory of Music, and the Institute of Advanced Cinematographic Studies. The three municipalities have benefited by the acquisition of low-rise public housing in park settings, a large park, day-care centres for children, and new schools.

Urban development at La Défense

Modern business quarters. As a counterbalance to the westward march of office buildings and as part of the effort to limit the obliteration of residential quarters around the business centre, five "poles of attraction" were instituted in other parts of Paris, beginning in the late 1960s. Two of these are directly on the waterfront at each end of Paris: Front de Seine ("Seine Waterfront") in the southwest corner, and Austerlitz-Bercy-Lyon in the southeast corner.

The Front de Seine is located on the Left Bank, between

Josef Muench



Fountain and the Luxor Obelisk in the Place de la Concorde.

the Eiffel Tower and the southern city limits. A neighbourhood of factories and substandard housing has been replaced by a spread of 16 towers of 32 stories and four buildings 15 stories high. One-quarter of the space is used for offices and the rest for apartments, some of which are low-cost housing.

Modernization at Bercy

The project at the opposite end of the river straddles the Seine, with office buildings around the Austerlitz (Left Bank) and the Lyon (Right Bank) railroad stations. Bercy, which lies directly on the river on the Right Bank, was until this development one of the "secret cities" of Paris. This was the village of vintages, where merchants stored and sold their stocks of wine. Fenced and guarded, its chalets lined cobbled lanes named for the great vineyard districts of France. The great oaks, it was said, flourished because their roots were soaked in wine. The Bercy area is being redeveloped on a large scale, with many large office blocks and a new sports and leisure centre.

The "Italian sector," near the Place d'Italie in the south of Paris, between the centre and Orly Airport, underwent an extensive renovation during the 1970s. Some 9,000 time-worn dwellings were demolished and replaced by 13,000 new ones capable of housing 45,000 persons. A cultural centre was built, and new green spaces and a sports ground were provided.

The planning achievement that overshadows the others is Maine–Montparnasse, where the centrepiece of the composition is a 56-story office tower on the site of the old Montparnasse railway station. A more compact station has been built one street away on the avenue du Maine, where the rails are hidden on three sides by three buildings 15 to 18 stories high. The units are joined by a raised platform that serves as a "ground level" above the street.

The ministry quarter. The government, both local and national, occupies a vast amount of Paris' office space. The seventh *arrondissement,* a district on the Left Bank that runs along the river from the Eiffel Tower to the Carrousel Bridge, is known as the ministry quarter. Most of the ministries are located there along with the headquarters of the Île-de-France Region and the National Assembly. The *arrondissement* is the old Faubourg Saint-Germain, an impeccable address since the early 18th century. As such, it was subject to heavy expropriation during the Revolution, and ministries are lodged mostly in splendid old mansions and convents. Although imposing, these are never large enough and are always difficult to adapt to the needs of modern administration. When it proves impractical to spread into adjacent buildings or to construct annexes in the garden, branches are installed wherever space can be found. Some of the ministries occupy as many as 25 separate buildings.

Quai d'Orsay

Probably the best known of all ministries is the low-built, ornate Ministry of Foreign Affairs on the Quai d'Orsay between the Invalides Esplanade and the National Assembly. The address "Quai d'Orsay" has become a synonym for the ministry.

The National Assembly is housed in the Bourbon Palace (1722–28), which was seized during the Revolution. Succeeding regimes have added bits and pieces onto the old palace, including the Greek peristyle facing the river as ordered in 1807 by Napoleon.

The old, disused Orsay railway station near the river was renovated and in 1986 was reopened as a museum of 19th-century art and civilization. It contains, among other collections, the Impressionist and Postimpressionist paintings—by Cézanne, Manet, Monet, Renoir, van Gogh, and others—that were formerly in the Jeu de Paume.

The Latin Quarter (Quartier Latin). At the Concorde Bridge the boulevard Saint-Germain begins, curving eastward to join the river again at the Sully Bridge. A little less than halfway along the boulevard is the pre-Gothic church of Saint-Germain-des-Prés. The old church, which belonged to a Benedictine abbey founded in the 8th century, was sacked four times by Vikings and was rebuilt between 990 and 1201. Parts of the present church date from that time.

The area has long been a gathering place for practitioners of the arts. Racine died there in 1699, Delacroix had his studio in the Place Fürstemberg, publishing houses moved in during the 19th century, and the principal cafès have been meeting places for artists, writers, and publishers ever since. From 1945 to about 1955 it was the hub of the existentialist movement and an associated revival of bohemianism. It is still a lively centre for literature, food, and conversation.

Straight north from the Saint-Germain-des-Prés crossroads is the School of Fine Arts (École des Beaux-Arts), the state school of painting and sculpture, on the Quai Malaquais. Two streets south of the crossroads is the church of Saint-Sulpice (1646–1780), the work of six successive architects. The street alongside the church is sprinkled with shops specializing in devotional statuary, much of it on the aesthetic level of tourist souvenirs and known in France as "Saint Sulpicerie." Fastward to the boulevard Saint-Michel, the area toward the river from the boulevard Saint-Germain is a tangle of narrow, animated streets, which typify the tourist's idea of a vivacious and noisy Paris.

Beyond the boulevard Saint-Michel is the university precinct, self-governing under the kings, where, in class and out, students and teachers spoke Latin until 1789. At the junction of the boulevards Saint-Germain and Saint-Michel are the remains of one of the three baths of the Roman city. These are in the grounds of the Cluny Museum, a Gothic mansion built 1485–1500, which now houses a collection of medieval works of art, including the renowned six-panel unicorn tapestry "La Dame à la licorne."

The wide, straight boulevard Saint-Michel is the main street of the student quarter. It is lined with bookshops, cafès, cafeterias, and movie houses. The buildings of the university are found on smaller streets. The university was built up of colleges, each founded and supported by a donor, often a prelate or a religious order. In about 1257 Robert de Sorbon, chaplain to Louis IX, established a college, known as the Sorbonne, that eventually became the centre of theological study in France. The oldest part of the Sorbonne is the chapel (1635–42), the gift of Richelieu, who is buried there. Designed by Lemercier, it was one of a number of new domed Jesuit-style churches of the period.

The Sorbonne

The Sorbonne served for centuries as the administrative seat of the University of Paris. In 1968–71 the university was divided into a number of entirely separate universities, and the Sorbonne building proper continues to serve as premises for some of these. Other faculties, schools, and institutes have moved out of Paris to more spacious sites in the suburbs in an effort to ease the overcrowding of the Paris student milieu.

The independent College of France was set up a few steps from the university by Francis I in 1529 to offer a more liberal, modern curriculum than the narrow theology and Latin of the Sorbonne. Bestowing no degrees, it has always had a superb faculty of well-known specialists, especially in philosophy, literature, and the sciences.

At the top of the hill rising from the river the boulevard skirts the Luxembourg Gardens, the remains of the park of Marie de Médicis' Luxembourg Palace (1616–21), which now houses the French Senate. The gardens are planted with chestnuts and are enhanced with a pond for toy sailboats, a marionette theatre, and statuary.

The Panthéon

Across the boulevard at the end of the rue Soufflot stands the Panthéon (1755–92), designed by Jacques-Germain Soufflot. It was commissioned by Louis XV, after his recovery from an illness, as a votive offering to St. Geneviève and was to replace the mouldering 5th-century abbey in her name. Though intended as the principal church in Paris, it was renamed the Panthéon by the Revolutionary authorities, who made it the last resting place for heroes of the Revolution. The walling up of a number of its windows and removal of much interior decoration replaced the intended effect of light interior space with a gloomy dignity. Among those buried under the inscription "Aux grands hommes, la Patrie reconnaissante" ("To great men, [from] their grateful land") are the authors Victor Hugo, Voltaire, Rousseau, and Zola and Jean Moulin, chief of the Resistance in World War II.

Behind the Panthéon is a steep street named the rue

de la Montagne Sainte-Geneviève. It was the paved road to Italy in Roman times. The hill leads down to the lively market square of Place Maubert and a tangle of ancient, picturesque riverside streets. The best known of these is the medieval rue de la Huchette, from which the rue du Chat-qui-Pêche ("Street of the Fishing Cat") leads to the Quai Saint-Michel. Two churches, Saint-Séverin (1489–94), Gothic and humble, and Saint-Julien-le-Pauvre (1165–1220), which belongs to the transitional period between the Romanesque and the Gothic, rise from the tangle of streets. The square in front of the church offers one of the finest views of Notre-Dame de Paris.

The Buttes. The river valley of Paris is almost entirely circled by high ground. Upon the heights of Passy, on the Right Bank between the western city limits and the Arc de Triomphe, perch the wealthy neighbourhoods of the 16th *arrondissement.* The Butte-Montmartre (18th *arrondissement*) and the Buttes-Chaumont (19th *arrondissement*), which rise along the northern rim of the city, are still working-class. The 18th *arrondissement* has broad avenues, but it also has winding lanes, some of which become stairways on the steeper hills. From the early 19th century until the migration in the 1920s to Montparnasse, Montmartre was the major art colony of Paris. Some sections are highly commercialized for the tourist trade; others, however, are unself-consciously picturesque. Montmartre is known for its nightclubs and entertainment.

Mont-martre

The most noted landmark of Montmartre was built only in 1919: the Sacred Heart Basilica (Basilique du Sacré-Coeur), paid for by national subscription after the French defeat by the Prussians in 1870. The work began in 1876 but was delayed by the death of the architect, Paul Abadie, who took inspiration from the 12th-century five-domed Romanesque church of Saint-Front in Périgueux, itself inspired by either Venetian or Byzantine churches. Alongside the monumental terraced stairway of the garden-planted Square Willette below the church entrance runs the only funicular railway in Paris.

On the Buttes-Chaumont just to the east of Montmartre is the Buttes-Chaumont Park, which was created under Haussmann in 1864–67. A bare hill, half hollowed out by abandoned tunnel quarries and filled with the refuse of generations, was turned into a romantic landscape with a lake, a waterfall, a grotto, winding woodland paths, and picturesque bridges. It is the largest park within the walls of Paris.

This portion of the 19th *arrondissement* is known as Belleville, and just below on the slope toward the river is Ménilmontant, in the 20th *arrondissement.* Père-Lachaise Cemetery, on hilly ground in the 20th *arrondissement,* is the site of the Federalists' Wall (Mur des Fédérés), against which the last of the Communard fighters were shot and to which pilgrimages are still made. Chopin, Marshal Ney, Baron Haussmann, Alfred de Musset, Balzac, Delacroix, Bizet, Rossini, Sarah Bernhardt, Isadora Duncan, Colette, and Edith Piaf are among the well-known people buried there. On the northeastern edge of the city proper, at La Villette, the giant structure of the old city abattoirs, out of use since 1974, was reopened in 1986 as a science museum. Nearby are a spherical panoramic cinema building and a large leisure park and culture complex.

Parks and open spaces. The appearance of Paris as a city well-endowed with parks is fostered in part by the garden effect of the Seine's open waters and its tree-lined banks. Some 87,000 trees (mostly plane trees, with a scattering of chestnuts) line the streets as well. But in fact parks and open spaces cover only 11 percent of the area of the City of Paris. Under Mayor Jacques Chirac the municipal government initiated efforts to create new parks, but the results have been modest.

Abundance of trees

Most of the parks and gardens of the modern central city are on land that formerly was reserved for the kings on the old city's outskirts. Under Napoleon III, who had been impressed by London's parks while living in Britain, two ancient royal military preserves at the approaches to Paris were made into "English" parks—the Bois de Boulogne on the west, and the Bois de Vincennes to the east. Moreover, during his reign a large area of land was laid out in promenades and garden squares. An assortment of municipal departments care for the public squares, street trees, riverbank trees, and various gardens.

(B.E./J.A.C.A.)

THE PEOPLE

In 1850 Paris had only 600,000 inhabitants. But it then grew rapidly, as industrial expansion attracted a constant stream of people from the provinces. By 1870 the population had surpassed 1,000,000; and by 1931 the conurbation contained some 5,000,000 people, more than half of them living in the City of Paris, the administrative city within the old gates. Since World War II this growth has continued, and Greater Paris by the late 1980s had close to 9,000,000 inhabitants. The population of the City of Paris, however, has steadily declined, from a peak of 2,900,000 in 1931 to 2,200,000 by 1982, so that more than three out of four Parisians are now suburbanites. The shift has taken place in part because massive rehousing has reduced the City's high density, though it remains well above the north European average. Many families have moved out to newer and more spacious homes in the suburbs, leaving the City of Paris with an aging population and one that is curiously solitary: almost half of the households consist of just one person.

Paris-born Parisians are outnumbered by immigrants who keep their provincial ties; hence many shops, restaurants, and neighbourhoods have a regional flavour. For example, there are said to be more Aveyronnais living in Paris than in the Aveyron *département* of central France. Most of the population is nominally Roman Catholic, though only a small percentage attend Mass regularly.

The foreign element has been increasing and now accounts for nearly one-fifth of the total population. The majority are Muslim Arabs from Algeria, Morocco, and Tunisia. In general, the North Africans are badly housed in the poorer quarters and are employed at menial jobs; in the 1980s their presence gave rise to racial tensions and conflicts. The sizable population of blacks is made up mainly of immigrants from the French Caribbean territories of Martinique and Guadeloupe: they tend to have better jobs and living conditions and are better accepted than are the Muslims. The Jewish community, which has long been settled in Paris, is centred on the Rue des Rosiers quarter of the Marais, where there are numerous synagogues, kosher stores, and Hebrew bookshops. In the 1980s encounters with Sephardic Jewish immigrants from North Africa provoked a mild revival of the anti-Semitism rife in prewar Paris.

In the earlier part of the 20th century Paris was favoured by expatriate writers and artists, including Ernest Hemingway from the United States, James Joyce from Ireland, Pablo Picasso from Spain, and Amedeo Modigliani from Italy. The foreign population from Europe and North America is now fairly small, however, consisting mainly of business people and the staffs of the large Paris-based international agencies, notably UNESCO and the Organisation for Economic Co-operation and Development.

THE ECONOMY

Paris is not only the political and cultural capital of France but also its major financial and commercial centre. Even if large French firms have their manufacturing plants in the provinces, they nearly all keep their headquarters in Paris, conveniently close to the major banks and key ministries. As an industrial centre the Paris region is less dominant in France than it was in its heyday in the 1930s, for since World War II most industrial growth has been in the provinces; but it still contains nearly a quarter of French industry. As a financial centre Paris is the base for many large international concerns in commerce and banking, and despite some pockets of poverty it is a very wealthy city, home of many vast private fortunes both French and foreign. After industry and commerce, the main activity is government administration, which employs nearly 700,000 people.

Industry. It was largely because it was already the political capital, with firms thus attracted to it, that Paris became an actively industrial city in the 19th century. Unlike other older French industrial areas, such as Lor-

raine and the Nord, it was not near mineral resources. But it did have some natural assets of its own, notably the Seine River, which is still used for heavy barge traffic moving downstream to the sea at Le Havre. Traditional industries were devoted mainly to handicrafts and luxury goods, but, when the growth of railways and canals in the 19th century made the northern coalfields more accessible, heavier industries such as engineering and chemicals began to develop. These soon spread beyond the city into the new industrial suburbs.

Since the 1950s the policy of successive French governments has been to limit the industrial growth of the Paris region in favour of the provinces, many of them still underdeveloped. Paris, which is in many ways overcongested, has instead been encouraged in developing commerce, finance, and services, a policy that has achieved some success. Industrial firms were forbidden to expand in the Paris area or were given various cash incentives to transfer their plants to the regions, and many did so. From 1962 to 1973 the Paris region had a net loss of 77,000 industrial jobs (about 5 percent of the total), while the rest of France gained 670,000.

The region, however, still employs well over a million people in industry. The City of Paris itself is still the home of many small-scale but typically Parisian activities: haute couture, notably in the Faubourg Saint-Honoré; the clothing industry, in the Sentier quarter; jewelry, in the Place Vendôme and rue de la Paix; furniture making, in the Faubourg Saint-Antoine. Newer and larger industries have developed in the northern and western suburbs, especially along the Seine valley toward Rouen: located there are several large automobile and aircraft factories. To the northwest, along the Seine's loop from Suresnes to Gennevilliers, are armaments factories, heavy engineering works, and chemical plants. Other firms are located to the northeast, within reach of Charles de Gaulle Airport. To the south and southwest along the valley of the Bièvre River are tanneries, cement works, breweries, tobacco factories, and other traditional concerns. Many light industries have settled in the new towns that now ring the Paris conurbation.

Commerce and finance. The major French banks, insurance companies, and other financial bodies are all centred on Paris, predominantly in the main financial quarter on the Right Bank around the Stock Exchange (Palais de la Bourse) and the state-owned Bank of France. Scores of foreign multiservice banks have branches in Paris, but the city is not a dynamic centre for specialist activities such as merchant banking and venture capital investment. The Paris Stock Exchange, though expanded and modernized in the 1980s, still handles little more than one-fifth of the volume of the London Stock Exchange.

Paris has developed greatly since World War II as a centre for international commerce, especially in the new skyscraper quarter of La Défense, where many large companies are situated. The city is one of the world's most popular sites for international business conferences, often hosting more than 200 a year. It has several major modern convention centres, notably the Palais des Congrès at the Maillot Gate.

It was the French who invented the modern department store, with the opening of the Bon Marché on the Left Bank in 1852. In the 1960s the major food and wine wholesale markets were transferred from their central locations at the Halles on the Right Bank and the Halle aux Vins (the wine market) on the Left Bank to new and more spacious homes in the suburbs.

Transportation. The Paris public transport system, operated by a body that is largely state-controlled, has been superbly modernized and extended since the early 1970s, and it is now regarded by many as the finest of any major world city. Trains on the 16 principal lines of the Métropolitain (Métro) subway system are fast and frequent. The Réseau Express Régional (RER), a high-speed express subway system, extends far into the suburbs, and at some points its lines have been integrated with the mainline railway network. The hub of the system is Châtelet–Les-Halles, said to be the world's largest and busiest underground station. By the 1980s Paris had become a commuter city, its Métro carrying 5,000,000 passengers daily. The cost of the Métro is subsidized by the government; passengers pay only about half of the actual cost. The city's bus system also has been modernized.

Such improvements have been part of an official campaign to ease traffic problems by discouraging the daily use of automobiles for commuting. But the campaign has not been entirely successful. The volume of traffic has swelled inexorably, and, although it often moves briskly along the many broad avenues and boulevards, it also can get clogged for hours in the honeycomb of narrow older streets. In a further attempt to decrease congestion the city has created many new underpasses and riverside expressways and has developed an extensive network of one-way streets. Underground garages have been built to provide parking space.

The city of Paris is encircled by an expressway, the boulevard Péripherique. The expressway is linked with suburban highways and with the national highway network, of which Paris is the hub. Similarly, the large Paris railway terminals serve the French railway network, first built in the 19th century. The network has been modernized and operates high-speed trains to Lyon and elsewhere. The main international airport is Charles de Gaulle, to the northeast; the older Orly Airport, to the south, is used mainly for domestic and charter flights. The Seine River carries barges and pleasure traffic; there are commercial ports both upstream and downstream from the city.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. A sharp distinction is drawn between city administration and suburban administration. The City of Paris is a single political unit—a *commune*—governed by an elected mayor and council, like any other French *commune* down to the smallest village. The suburbs consist of more than 1,200 separate *communes,* large and small, which together with the City of Paris form the administrative region of Île-de-France. The Île-de-France region, with an area of about 4,640 square miles (12,000 square kilometres), extends far beyond the Paris conurbation. The urban area of Greater Paris therefore is not a political unit, and coordination is frequently poor between Paris and its inner suburbs. Because of the fierce rivalries between left-wing and right-wing *communes,* it has never been possible to follow the pattern of other major world cities and create a federated urban district.

The Île-de-France is the most populous of France's 22 regions. The region consists of eight *départements:* Hauts-de-Seine, Seine-Saint-Denis, Val-de-Marne, Essonne, Yvelines, Val-d'Oise, Seine-et-Marne, and Paris. Under the socialist government's devolutionary reforms of 1982–86, the Île-de-France, like the other regions, was given a certain degree of autonomy. It has a directly elected assembly with a chairman and executive; it can raise its own taxes; and it has responsibility for adult education and for some aspects of culture, tourism, road building, planning, and aid for industrial development. The directly elected representatives of the eight *départements* have also been given increased responsibilities: they run the welfare and social services, involving large budgets, as well as controlling some matters concerning the infrastructure. The *communes* in turn look after their own town planning and building. Each *département* is supervised by a state-appointed prefect and the Île-de-France by a regional prefect, but the office of prefect has much less power than before the reforms.

The City of Paris itself has a curious history of local government. The Municipal Council of Paris is elected by the people every six years. From 1871 to 1977 the council had no mayor and was controlled directly by the departmental prefect, so that Paris had less autonomy than any village. The national government, worried by memories of the uprisings of 1789, 1848, and 1871, wanted to keep power from the Paris populace. A statute passed in 1975, however, permitted the councillors once again to elect their own mayor. The mayor now has the same status and powers as in other French towns. The first mayoral election was held in 1977. In 1982 a ward system was introduced, whereby each of the 20 *arrondissements* was

given its own mayor and local council. In practice, however, these have not proved very effective, and real control remains in the hands of Paris' mayor.

Services. Electricity and gas utilities in Paris are run by the state, as are the fire departments, the telephone service and, unlike all other French cities, the police, who have wider powers than in other towns. In addition to dealing with crime, traffic, and public order, the Paris police register vehicles and drivers; issue passports, identity cards, and aliens' residence permits; and conduct political surveillance. One of their main challenges has been the growing wave of crime, particularly terrorism. Special police agencies include detective and counterespionage services and the Republican Security Force, used for dispersing demonstrations. The Republican Guard, a mounted squadron of spurred, helmeted, plumed, and breast-plated guardsmen armed with sabres, are used only for ceremonial occasions such as visits by foreign heads of state.

Health. The public hospitals and hospital groups in Paris are run jointly by the city and the national Ministry of Health and are financed largely by a social welfare system. Some other hospitals are run by churches and by private organizations, and there are numerous private clinics. Of the city's many medical research bodies the best known is the Pasteur Institute, founded in 1887.

Pasteur Institute

Education. As in the rest of France, schools are largely in the hands of the state and are of three main kinds: primary, junior secondary (*collèges*), and senior secondary (*lycées*). Almost one-fifth of all pupils are in private, non-state schools, most of them run by the Roman Catholic Church. Some of the best-known *lycées* in central Paris are the Lycée Henri IV, the Lycée Louis-le-Grand, and the Lycée Janson de Sailly; such schools traditionally have educated a large number of the nation's intelligentsia.

Paris has long been an important world centre of higher education. The building known as the Sorbonne housed the arts and science faculties of the University of Paris, one of Europe's oldest universities, until it was split into 13 autonomous universities (1968–71), some of which still occupy the Sorbonne. The 13 universities, overcrowded and underfunded, with some 250,000 students, have lost prestige to a number of much smaller bodies of higher education that are more specialized and more elitist. These include the engineering, technical, and business colleges known as the Grand Schools (Grandes Écoles), of which the two best known in the Paris area are the School of Higher Business Studies (École des Hautes Études Commerciales; HEC) and the Polytechnical School (École Polytechnique), both of them now transferred to spacious new homes in the suburbs. The Normal Superior School (École Normale Supérieure) serves mainly to prepare future university and *lycée* teachers. The Institute of Political Sciences (Institut des Sciences Politiques) nurtures many of those who go on to the influential postgraduate Normal School for Administration (École Normale d'Administration; ENA) to train as senior civil servants. This selective structure has served France well in many ways but has come under heavy criticism and undergone sporadic reform.

CULTURAL LIFE

Paris has for centuries been regarded as the main cultural powerhouse of the Western world, a magnet for artists and intellectuals, a place where new ideas originate and art reigns supreme. While some people feel its cultural life is now more a matter of show and dazzle than of true creativity, the cultural life is still highly active and distinctive. Parisians love novelty, they have an abounding intellectual curiosity, they know how to dress up the simplest cultural event with flair and elegance, and they are avid patrons of the arts, so that the theatres and concert halls, museums, art galleries, and art cinemas are always well-attended.

The principal state-run theatres are the Comédie-Française, the Odéon Theatre, and the National Theatre of Chaillot, which offer a repertoire of French classics, serious modern plays, and foreign imports. Lighter fare is provided by the many privately owned "boulevard" theatres, which struggle to survive. There are more than 150 smaller theatres, many of them state-supported, which present a mixed program of experimental "fringe" shows, cabaret, and the like. In addition to many multiscreen commercial cinemas, there are scores of little "art" houses that show a wide variety of movies, many of them with subtitles. France's main film studios are in the suburbs of Paris.

Theatre

While the newly remodeled Louvre is the greatest of the classic art museums, Paris has built major new museums in the past years, including the National Museum of Modern Art in the Pompidou Centre, the Orsay Museum of 19th-century art and civilization (opened in 1986), and the new science museum at La Villette. A speciality of Paris is the staging of large and lavishly mounted exhibitions, usually retrospectives of an individual artist or historical period. The city's musical life, once moribund, has become much livelier since the early 1970s, in part because the state has provided much-needed funding. The quality of the Paris Opera has been uneven and the building ill-suited to modern productions, but the city renewed its commitment to opera by opening a new opera house in 1989. Major annual festivals emphasizing music and drama are the Marais Festival, held in June and July, and the Autumn Festival, held from mid-September through December.

The main publishing houses and bookshops are located in the Latin Quarter and Saint-Germain-des-Prés. The best-known daily newspaper is the evening *Le Monde*, followed by the daily *Le Figaro* and *Liberation*. While the Paris daily press is relatively weak, accounting for less than one-third of total daily circulation in France, weekly newsmagazines flourish, notably *L'Express* and *Le Point*. All of the main French national radio and television networks are centred in Paris; some are owned by the state and some are privately owned. (J.A.C.A.)

History

FOUNDATION AND MEDIEVAL GROWTH

Paris was in existence by the end of the 3rd century BC as a settlement on an island, the modern Île de la Cité, in the Seine River and was inhabited by a Gallic tribe known as the Parisii. The first recorded name for the settlement was Lutetia (Latin: "Midwater-Dwelling"). When the Romans arrived, the Parisii were sufficiently organized and wealthy to have their own gold coinage. Julius Caesar wrote in his *Commentaries* (52 BC) that the inhabitants burned their town rather than surrender it to the Romans. In the 1st century AD Lutetia grew as a Roman town and spread to the left bank of the Seine. The straight streets and the public buildings in this locale were characteristically Roman, including a forum, several baths, and an amphitheatre.

The Roman town

A series of barbarian invasions began in the late 2nd century. The town on the left bank was destroyed by the mid-3rd century, and the inhabitants took refuge on the island, around which they built a thick stone wall. From the early 4th century the place became known as Paris.

By this time, Christianity seems to have spread to Paris. A 10th-century sacramentary cites St. Denis (Latin Dionysius) as having been the first bishop of Paris, in about AD 250. A graveyard excavated near the Carrefour des Gobelins shows that there was a Christian community in very early times on the banks of the Bièvre (a left-bank tributary of the Seine); but it was probably under St. Marcel, the ninth bishop (*c.* 360–436), that the first Christian church, a wooden structure, was built on the island.

By the end of the 5th century, the Salian Franks, under Clovis, had captured Paris from the Gauls, making it their own capital. It remained the capital until the end of Chilperic's reign in 584, but succeeding Merovingians carried the crown elsewhere. Charlemagne's dynasty, the Carolingians, tended to leave the city in the charge of the counts of Paris, who in many cases had less control over administration than did the bishops. After the election of Hugh Capet, a count of Paris, to the throne in 987, Paris, as a Capetian capital, became more important.

The population and commerce of Paris increased with the gradual return of political stability and public order under the Capetian kings. The maintenance of order was entrusted to a representative of the king, the provost of Paris (*prévôt de Paris*), first mentioned in 1050. In the 11th century the first guilds were formed, among them the butchers' guild and the river-merchants' guild, or *marchan-*

dise de l'eau. In 1141 the crown sold the principal port (near the Hôtel de Ville) to the *marchandise,* whose ship-blazoned arms eventually were adopted as those of Paris. In 1171 Louis VII gave the *marchandise* a charter confirming its "ancient right" to a monopoly of river trade.

Expansion under Philip II

During the reign of Philip II (1179–1223), Paris was extensively improved. Streets were paved, the city wall was enlarged, and a number of new towns were enfranchised. In 1190, when Philip II went on a crusade for a year, he entrusted the city's administration not to the provost but to the guild. In 1220 the crown ceded one of its own precious rights to the townsmen—the right to collect duty on incoming goods. The merchants were also made responsible for maintaining fair weights and measures. The King's formal recognition of the University of Paris in 1200 was also a recognition of the natural division of Paris into three parts. On the Right Bank were the mercantile quarters, on the island was the *cité,* and the Left Bank contained the university and academic quarters. Numerous colleges were also founded, including the Sorbonne (about 1257).

In the 14th century the development of Paris was hindered not only by the Black Death (1348–49) but also by the Hundred Years' War (1337–1453) and by internal disturbances resulting from it. The provost of the merchants in 1356 was Étienne Marcel, who wanted a Paris as rich and free as the independent cities of the Low Countries. He gave the House of Pillars to the municipal government, and he slew the Dauphin's counselors in the palace throne room and took over the city. Marcel showed great executive skill and equally great political stupidity and allied himself with the revolting peasants (the Jacquerie), with the invading English, and with Charles the Bad, the ambitious king of Navarre. While going to open the city gates to the Navarrese in 1358, Marcel was slain by the citizens.

In 1382 a tax riot grew into a revolt called the "Maillotin uprising." The rioters, armed with mauls (*maillets*), were ruthlessly put down, and the municipal function was suspended for the next 79 years. It was not until 1533, when Francis I ordered the teetering House of Pillars replaced by a new building, that a monarch manifested an encouraging interest in municipal government.

The dynastic and political vendetta between the Burgundian and the Armagnac faction (1407–35) had continual repercussions in Paris, where the butchers and skinners, led by Simon Caboche, momentarily seized power (1413). The resumption of the Hundred Years' War by the English in 1415 made matters worse. After a revolt of the Parisians (1418), the Burgundians occupied Paris; the Anglo-Burgundian Alliance (1419) was followed by the installation of John Plantagenet, duke of Bedford, as regent of France for the English king Henry VI (1422). Whereas Charles VI had lived in his father's Hôtel Saint-Paul, Bedford lived in the Hôtel des Tournelles, on the southeastern edge of the Marais, which was to be the Paris residence of later kings until 1559. During the reign of Charles VI, construction began on the Notre-Dame Bridge (1413).

In 1429 Joan of Arc failed to capture Paris. Only in 1436 did it fall to the legitimists, who welcomed Charles VII in person in 1437. Successive disturbances had reduced the population, but the Anglo-French truce of 1444 allowed Charles to begin restoring prosperity.

In 1469, during Louis XI's reign (1461–83), the Sorbonne installed the first printing press in Paris. Otherwise this was a period of intellectual stagnation. Churches were rehabilitated and new houses were built, however; from 1480 splendid private mansions began to appear, such as the Hôtel de Sens and the Hôtel de Cluny.

FROM THE RENAISSANCE TO THE REVOLUTION

The influence of the Italian Renaissance on town architecture appeared in the new building for the accounting office and in the reconstruction of the Notre-Dame Bridge (1500–10) in Louis XII's reign. Under Francis I (1515–47) this influence grew stronger, finding notable expression in the new Hôtel de Ville. Furthermore, whereas from Charles VII's time the kings of France had preferred to reside in Touraine, Francis returned the chief seat of royalty to Paris. With this in mind he had extensive alterations made to the Louvre from 1528 onward. The new splendour of the monarchy, which was well on its way toward absolute rule, was reflected in the way Paris developed as the capital of an increasingly centralized state. The population increased and the town expanded again. Rigorous measures were taken to stamp out Protestantism, which first appeared in Paris during Francis I's reign.

The Renaissance in Paris culminated with Henry II, who made his solemn entry into the capital in 1549. The new impulse given to building mansions for the nobility and bourgeoisie began to transform Paris from a medieval to a modern city. In 1548 the Brothers of the Passion began performing secular plays at the Hôtel de Bourgogne, in the rue Française, thus inaugurating the first theatre in Paris.

The transfer of the royal residence from the Hôtel des Tournelles to the Louvre, signaling the development of the neglected western outskirts of Paris, was completed after Henry II's death in 1559. Catherine de Médicis began to build the Tuileries Palace, the gardens of which became a meeting place for elegant society. Classical taste was brilliantly exemplified by the Pont-Neuf, begun in 1577.

The Wars of Religion

In the mid-16th century the Wars of Religion broke out in France between Roman Catholics and Huguenots, which in Paris brought about the Massacre of St. Bartholomew's Day (1572); the Day of the Barricades (1588), when the Catholic League rose against Henry III; and the long resistance of the Parisians to the Protestant Henry of Navarre, who succeeded as Henry IV in 1589. Henry IV's siege in 1590 was unsuccessful, and only after his conversion to Catholicism did Paris submit to him (1594).

In Louis XIII's reign (1610–43) Paris expanded farther. On the Left Bank, outside the wall, the queen mother, Marie de Médicis, built the Luxembourg Palace, with its spacious gardens; along the Right Bank, west of the Tuileries, she laid out the Cours-la-Reine as a promenade for carriages. While the Marais north of the Place Royale was being reclaimed and developed, two uninhabited islets east of the *cité* were united to form the Île Saint-Louis. On the western fringe of the town, a quarter with straight streets was laid out north of Richelieu's new palace, the Palais-Cardinal (1624–36; later the Palais-Royal), which also had a magnificent garden; west of this there was more building and a new fortification was erected.

The war of the Fronde (1648–53) was the major event of the first two decades of Louis XIV's reign. From 1661, when Cardinal Mazarin died and Louis started his personal rule, Paris was dedicated to reflecting the glory of the monarch, even though he was early resolved to establish himself and the seat of his government outside of Paris (he chose Versailles). For the planning of the new splendours of Paris, the greatest part of the credit must go to Jean-Baptiste Colbert, the king's superintendent of buildings.

Work on the Louvre had been resumed in 1624 and was completed by Claude Perrault's magnificent colonnade (1667–74). The Tuileries Palace was altered and sumptuously decorated. Beyond its gardens to the west, outside the walls of Paris, the tree-planted avenues of the Champs-Élysées were laid out (1667); these were complemented, at the opposite end of Paris, by the Cours de Vincennes.

In 1702 the Marquis d'Argenson (Marc René de Voyer), who as lieutenant general of police succeeded the provosts of Paris, raised the number of districts from 16 to 20 (15 on the Right Bank, five on the Left). Paris had nearly 600,000 people, and from the Left Bank new suburbs were advancing toward the villages on the surrounding hills.

During the 18th century a great deal was done to improve and beautify Paris. Louis XV's temporary residence in the Tuileries during his younger days encouraged development nearby, so that the Faubourg Saint-Honoré expanded and became, like the Faubourg Saint-Germain, an aristocratic quarter. The garden of the Palais-Royal became a centre of elegant society. The Grands Boulevards began to be bordered with houses, including some fine mansions, and the eastern stretch became a fashionable promenade with little theatres and cafés. Villas built by nobles and financiers were scattered around this outlying sector. On the Left Bank the southern course of boulevards was laid out and the routes were lined with trees and houses. Some of the houses that had been built earlier

on the bridges were razed in 1786–88; others remained until 1808. Water was supplied to both banks by two fire pumps, developed by Jacques-Constantin Périer and his father, Auguste-Charles. The wall of the farmers-general, built in the 1780s to facilitate the levying of duties on imports, represented the extension and the unity of Paris.

EVOLUTION OF THE MODERN CITY

The Revolution and Napoleon I. The French Revolution of 1789 destroyed those vestiges of the seigneurial systems that had remained in Paris and consolidated the status of Paris as the capital of a centralized France. The major events of the Revolution took place in Paris, including the storming of the Bastille (July 14, 1789); the conveying of the King and the National Constituent Assembly from Versailles to Paris (October 1789); the establishment of the numerous clubs in the convents of the old religious orders, Jacobins, Cordeliers, and Feuillants; the insurrection that heralded the abolition of the monarchy (Aug. 10, 1792); the execution of the King (Jan. 21, 1793) in the Place de la Révolution, not yet named Place de la Concorde; the most prolonged manifestation of the Terror (1793–94); and the series of coups d'état, from that of 9 Thermidor, year II (1794), to that of 18 Brumaire, year VIII (1799), which preceded the ascendancy of Napoleon Bonaparte.

Under the Thermidorians and the Directory the boulevard des Italiens became a resort of the fashionable and the frivolous, whereas the populace favoured the boulevard du Temple. After the inauguration of the First Empire, Napoleon in 1806 ordered the triumphal arches of the Carrousel and of the Étoile to be erected. While the Neoclassical style recalled imperial Rome, great works of public utility served to modernize Paris: the Bourse; new quays and bridges (the Arts, Jena, Austerlitz, and Saint-Louis bridges); the Ourcq and Saint-Martin canals; numerous fountains (such as the Palmier Fountain, on the site of the Châtelet); as well as slaughterhouses, marketplaces, the wine market, and the warehouses of Bercy.

The Restoration and the July Monarchy

Industrialization, in progress in the Napoleonic period, advanced rapidly under the Restoration (1814–30) and the July Monarchy (1830–48). Gas lighting was introduced; omnibus services began in 1828; and Paris got its first railway, which ran to Le Pecq, near Saint-Germain-en-Laye, in 1837. New districts grew up on the outskirts of Paris. Although the wall of the farmers-general remained the administrative boundary of Paris until 1859, it was decided in 1840 to refortify the capital with a longer military wall.

Napoleon III and Haussmann. Even by the mid-19th century, some areas of Paris had not been improved substantially for hundreds of years. Access from one centre to another and to the railway stations (which had become in effect the gateways of Paris) was difficult; moreover, overpopulation and rapid industrialization had brought squalor and misery, which account in part for the dominant role of Paris in the revolutions of both 1830 and 1848. Napoleon III, emperor from 1852 to 1870, enjoined his prefect of the Seine, Baron Haussmann, to remedy these problems.

Haussmann was the creator of modern Paris. A planner on the grand scale, he advocated straight arterial thoroughfares, symmetry, and advantageous vistas. He slashed the boulevards through the tangles of slums, began the modern sewer and water systems, gutted the Île de la Cité, rebuilt the ancient market of the Halles, and added four new Seine bridges and rebuilt three old ones. The brilliance and prosperity of Paris under Napoleon III were exemplified in the exhibitions held there in 1855 and 1867.

The Third Republic and after. The Franco-German War (1870–71), which brought the fall of the Second Empire and the siege of Paris, was followed by the Commune (1871). Under the Third Republic, Haussmann's projects were continued. Further international exhibitions (1878, 1889, 1900, and 1937) were the occasions for the building of monuments such as the Trocadéro (1878), the Eiffel Tower (1889), and the Grand Palais and Petit Palais, with the Alexandre III Bridge (1900), and for the reconstruction of the Trocadéro as the Chaillot Palace (1937). The Métro was constructed, commerce and industry annexed formerly residential districts, and the ever-expanding population overflowed the old limits of Paris. Louis-Philippe's fortifications were abolished by a law of April 1919.

During the German occupation of Paris in World War II, the city was only slightly damaged. It was a centre for the activities of the Resistance movement, which culminated in the liberation of the city in August 1944.

The immediate postwar years were a time of eager intellectual activity but also of poverty and social tension. The housing shortage was grave, the psychological scars of the German occupation were slow to heal, and colonial wars and political instability lowered morale. Conditions gradually improved, especially after Charles de Gaulle returned to power in 1958. The city's economy improved, old buildings and neighbourhoods were cleaned up and renovated, and housing and commercial space were built.

Great student uprising

In May 1968 Paris was rocked by a great student uprising, which swelled from scattered unrest among students in the Latin Quarter to a nationwide outbreak of labour strikes and protests. Attention was focused on Paris' economic and social problems, and the uprising was later seen as useful in hastening the modernization of French society. (B.E./J.A.C.A.)

During the last decades of the 20th century, several new developments bolstered the cultural and economic position of Paris. These included the architecturally innovative Pompidou Centre (or Beaubourg) and the Orsay Museum. Commercial projects included the office complexes at Bercy, at La Villette, and at La Défense, a high-rise business district on the periphery of the city. The latter's Grande Arche de la Défense, a monumental rectangular arch completed in 1989, echoes the Arc de Triomphe to its east. This arch was one of the architectural *grands projets* ("great projects") promoted by President François Mitterrand to symbolize French cultural and economic leadership. Other such projects included the glass pyramid at the Louvre and the French National Library (Bibliothèque nationale de France), which was completed in 1995.

During the late 20th century, Paris attracted thousands of migrants from Africa, the Middle East, the Caribbean region, and other parts of Europe. While their presence aroused social and political tensions, they contributed to the city's economy and cultural diversity. The increasingly cosmopolitan character of Paris reinforced its place as one of the great metropolises of the world. (Ed.)

BIBLIOGRAPHY

General description: JOHN RUSSELL, *Paris* (1983); and ANTHONY GLYN, *The Companion Guide to Paris* (1985). Guide books include *Paris,* 3rd ed. (1996), from the Michelin Tourist Guides series; and IAN LITTLEWOOD, *Paris: A Literary Companion* (1987), for literary landmarks. JEAN BASTIÉ, *Géographie du Grand Paris* (1984), describes the geography of Greater Paris; and FRANÇOIS CHASLIN, *Les Paris de François Mitterrand: histoire des grands projets architecturaux* (1985), describes architectural projects under President Mitterrand. For information on churches, see AMÉDÉE C.L. BOINET, *Les Églises parisiennes,* 3 vol. (1958–64). Of sentimental interest are LÉON-PAUL FARGUE *et al., Dans les rues de Paris au temps des fiacres* (1950); and R. HÊRON DE VILLEFOSSE, *Histoire et géographie galantes de Paris* (1957).

History: General histories include ROBERT LAFFONT (ed.), *The Illustrated History of Paris and the Parisians* (1958; originally published in French, 1958; also published as *Paris and Its People*); and PHILIPPE LEFRANÇOIS, *Paris à travers les siècles,* 10 vol. (1948–56). Studies of special periods include PAUL M. DUVAL, *Paris antique: des origines au troisième siècle* (1961); and BERNARD CHAMPIGNEULLE, *Paris de Napoléon à nos jours* (1969). Among analytical works are DANIEL ROCHE, *The People of Paris: An Essay in Popular Culture in the 18th Century* (1987; originally published in French, 1981); DAVID GARRIOCH, *Neighbourhood and Community in Paris, 1740–1790* (1986); LENARD R. BERLANSTEIN, *The Working People of Paris, 1871–1914* (1984); ANN-LOUISE SHAPIRO, *Housing the Poor of Paris, 1850–1902* (1985); and JERROLD SEIGEL, *Bohemian Paris: Culture, Politics, and the Boundaries of Bourgeois Life, 1830–1930* (1986). HERBERT R. LOTTMAN, *The Left Bank: Writers, Artists, and Politics from the Popular Front to the Cold War* (1982), details the intellectual life in Paris during the 20th century. NORMA EVENSON, *Paris: A Century of Change, 1878–1978* (1979), presents the city's evolution from a planner's point of view. (B.E./J.A.C.A.)

Particle Accelerators

A particle accelerator is a device that produces a beam of fast-moving, electrically charged atomic or subatomic particles. Physicists use accelerators in fundamental research on the structure of nuclei, the nature of nuclear forces, and the properties of nuclei not found in nature, such as the transuranic (heavier than uranium) elements and other unstable elements. Accelerators are also used for radioisotope production, industrial radiography, cancer therapy, sterilization of biological materials, polymerization of plastics, and a certain form of radiocarbon dating. The largest accelerators are used in research on the fundamental interactions of the elementary subatomic particles.

This article reviews the development of accelerators and delineates the various types and their distinguishing features. For specific information about the particles accelerated by these devices, see ATOMS and SUBATOMIC PARTICLES.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 111, 112, 127, and 723, and the *Index*.

The article is divided into the following sections:

General considerations 449
History 449
Constant-voltage accelerators 450
  Voltage multipliers (cascade generators)
  Van de Graaff generators
Betatrons 451
Cyclotrons 451
  Classical cyclotrons
  Synchrocyclotrons
  Sector-focused cyclotrons
Linear resonance accelerators 452
  Linear electron accelerators
  Linear proton accelerators
Synchrotrons 453
  Electron synchrotrons
  Proton synchrotrons
Other types of accelerators 456
  Colliding-beam storage rings
  Impulse accelerators
Bibliography 457

GENERAL CONSIDERATIONS

The effectiveness of an accelerator is usually characterized by the kinetic energy, rather than the speed, of the particles. The unit of energy commonly used is the electron volt (eV), which is the energy acquired by a particle that has a charge of the same magnitude as that of the electron when it passes between electrodes that differ in potential by one volt; it is equivalent to 1.602×10^{-19} joules. Related larger units are the kiloelectron volt (keV; equivalent to 1,000 eV), the megaelectron volt (MeV; 1,000,000 eV), the gigaelectron volt (GeV; 1,000,000,000 eV), and the teraelectron volt (TeV; 1,000,000,000,000 eV). Compared with the quantities of energy encountered in everyday experience, even the teraelectron volt is a very small amount, about that of a mosquito in flight. The masses of the particles accelerated are so small, however, that kinetic energies in this range correspond to very high speeds: the particles in the smallest ion accelerators travel about 8,000 kilometres (5,000 miles) per second, about 3 percent of the speed of light.

The particles that are accelerated most often are electrons or protons (ionized hydrogen), and their antiparticles, or heavier ionized atoms. Sometimes the primary beam is used; in other cases, the primary beam is directed onto a target to produce a beam of secondary particles, such as X rays, neutrons, mesons, hyperons, or neutrinos. A few accelerators are operated as sources of the intense radiation, called synchrotron radiation, emitted by electrons moving at almost the speed of light along curved paths.

Basic components

Every accelerator has three essential parts: a source of the particles to be accelerated, a vacuum chamber in which to accelerate them, and a source of the electric fields needed to effect the acceleration. Thermionic emission (the emission of electrons from the surface of a heated solid) provides electrons. Positive and negative ions are produced by electric arc or glow discharges in a gas at low pressure in a chamber; a high-voltage electrode extracts the ions from the gas through a hole in the chamber. The region in which the particles are accelerated must be highly evacuated to keep the particles from being scattered out of the beam, or even stopped, by collisions with molecules of air.

Accelerators are differentiated by the arrangements of the accelerating electric fields. In a linear accelerator, the path of the particles is a straight line, and the final energy of the particles is proportional to the sum of the voltages produced by the accelerating devices along that line. In a cyclic accelerator, the path of the particles is bent by the action of a magnetic field into a spiral or a closed curve that is approximately circular. In this case, the particles pass many times through the accelerating devices; the final energy depends on the magnitude of the voltages multiplied by the number of times the particles pass through. Because the total distance traveled by the particles in a cyclic accelerator may be more than a million kilometres, the cumulative effect of minute deviations from the desired trajectory would be dissipation of the beam. Therefore, the beam must be continually focused by the magnetic fields, which are precisely shaped by powerful magnets.

Since the late 1920s, when the first accelerators were built, the highest energies accessible have risen from around 1 MeV to 1 TeV. Several specific developments have allowed this progress to successively higher energies, but the basic principles of particle acceleration have remained essentially the same. For example, superconductivity has been employed to extend the reach of the highest-energy machines, but the machines themselves are still direct descendants of the first accelerators. At the same time, modern versions of the earlier low-energy machines have become valuable tools in medicine and industry, as well as in various areas of scientific research.

HISTORY

The motivation for most of the development of the various types of particle accelerators has been their application to research into the properties of atomic nuclei and subatomic particles. Starting with Ernest Rutherford's discovery in 1919 of a reaction between a nitrogen nucleus and an alpha particle, all research in nuclear physics until 1932 was performed with alpha particles from naturally radioactive elements. The natural alpha particles have kinetic energies as high as 8 MeV. Rutherford believed that the way forward lay in accelerating ions artificially to even higher energies, but there seemed little hope of generating laboratory voltages sufficient to accelerate ions to the desired energies. A calculation made by George Gamow (then at the University of Göttingen, in Germany) in 1928, however, indicated that considerably less energetic ions could be useful, and this stimulated attempts to build an accelerator that could provide a beam of particles suitable for nuclear research. The first successful experiments with artificially accelerated ions were performed at the University of Cambridge, Eng., by John Douglas Cockcroft

First proton accelerator

and E.T.S. Walton in 1932. Using a voltage multiplier, they accelerated protons to energies as high as 710 keV and showed that these react with the lithium nucleus, the products being two energetic alpha particles. Other developments of that period demonstrated principles still employed in the design of particle accelerators. For example, Robert J. Van de Graaff had constructed the first belt-charged electrostatic high-voltage generator at Princeton University, in New Jersey, U.S., by 1931.

The principle of the resonance linear accelerator was demonstrated by Rolf Widerŏe in 1928. At the Rhenish-Westphalian Technical University in Aachen, Ger., he had used alternating high voltage to accelerate ions of sodium and potassium to energies twice as high as those imparted by one application of the peak voltage. In 1931 in the United States Ernest O. Lawrence and his assistant David H. Sloan, at the University of California, Berkeley, employed high-frequency fields to accelerate mercury ions to more than 1.2 MeV. This work augmented Widerŏe's achievement in accelerating heavy ions, but the ion beams were not useful in nuclear research.

The magnetic resonance accelerator, or cyclotron, was conceived by Lawrence as a modification of Widerŏe's linear resonance accelerator. Lawrence's student M.S. Livingston demonstrated the principle of the cyclotron in 1931, producing 80-keV ions; in 1932 Lawrence and Livingston announced the acceleration of protons to more than 1 MeV. Later in the 1930s, cyclotron energies reached about 25 MeV and Van de Graaff generators about 4 MeV. In 1940 Donald W. Kerst, applying the results of careful orbit calculations to the design of magnets, constructed the first betatron, a magnetic-induction accelerator of electrons, at the University of Illinois.

The rapid advance in the science of accelerating particles to high energies that has occurred since the end of World War II was initiated by Edwin Mattison McMillan, at Berkeley, and Vladimir Iosifovich Veksler, at Moscow, who in 1945 independently described the principle of phase stability, which removed an apparent limitation on the energy of resonance accelerators for protons (see below *Cyclotrons: Classical cyclotrons*) and made possible the construction of magnetic-resonance accelerators (called synchrotrons) for electrons. The principle was promptly demonstrated by the construction of a small synchrocyclotron at the University of California and an electron synchrotron in England. The first proton linear resonance accelerator was constructed soon thereafter. The large proton synchrotrons that have been built since then all depend on this principle.

Phase stability

In 1947 William W. Hansen, at Stanford University, Calif., constructed the first traveling-wave linear accelerator of electrons, exploiting microwave technology that had been developed for radar during World War II.

The progress in research made possible by raising the energies of protons led to the building of successively larger accelerators; the trend was ended only by the cost of fabricating the huge magnet rings required—the largest weighs approximately 40,000 tons. A means of increasing the energy without increasing the scale of the machines was provided by the demonstration in 1952 by Livingston, Ernest D. Courant, and H.S. Snyder of the technique of alternating-gradient focusing (sometimes called strong focusing). Synchrotrons incorporating this principle needed magnets only $^1/_{100}$ the size that would be required otherwise. All recently constructed accelerators make use of alternating-gradient focusing.

In 1956 Kerst realized that, if two sets of particles could be maintained in intersecting orbits, it should be possible to observe interactions in which one particle collided with another moving in the opposite direction. Application of this idea requires the accumulation of accelerated particles in loops called storage rings (see below *Other types of accelerators*). The highest reaction energies now obtainable have been produced by the use of this technique.

Particle storage rings

CONSTANT-VOLTAGE ACCELERATORS

The simplest type of particle accelerator is constructed by mounting a particle source on one end of an insulated, evacuated tube and creating a high voltage between the ends, with the polarity such that the particles are impelled from the source toward the far end of the tube. Such an accelerator is necessarily linear, and the electrostatic field can be applied to a given particle only once (unless, as in the tandem accelerator described below, the charge of the particle undergoes a change in sign). The simplicity of concept becomes complex in execution when the electric potential exceeds 1,000,000 volts (1 megavolt, or 1 MV): these high voltages produce corona discharges and lightninglike sparks outside the accelerator, which dissipate the potential needed to accelerate the particles. Even more difficult to control are sparks within the equipment and, in positive-ion accelerators, unwanted secondary beams produced when the accelerated ions strike the end of the tube.

Voltage multipliers (cascade generators). The source of the high voltage for Cockcroft and Walton's pioneering experiments was a four-stage voltage multiplier assembled from four large rectifiers and high-voltage capacitors. Their circuit, in effect, combined four rectifier-type direct-voltage power supplies in series. The alternating voltage supplied by a high-voltage transformer was transmitted to the higher stages through an array of capacitors; a second group of capacitors kept the direct voltage constant. The final direct voltage would have been four times the peak voltage available from the transformer (200,000 volts) if corona discharge had not drained away considerable power. Nevertheless, the apparatus did accelerate protons to energies of 710 keV, sufficient to bring about the hoped-for result, a reaction with lithium nuclei. Their achievement, the first nuclear reaction effected by artificially accelerated particles, was recognized by the award of the Nobel Prize for Physics in 1951.

Cockcroft-Walton generator

Cockcroft and Walton's system for building up high direct voltages can be extended to multiplication factors many times that originally demonstrated. Commercially available accelerators that reach 4 MeV are based on this circuitry.

Van de Graaff generators. In Van de Graaff generators, electric charge is transported to the high-voltage terminal on a rapidly moving belt of insulating material driven by a pulley mounted on the grounded end of the structure; a second pulley is enclosed within the large, spherical high-voltage terminal, as shown in Figure 1. The belt is charged by a comb of sharp needles with the points close to the belt a short distance from the place at which it moves clear of the grounded pulley. The comb is connected to a power supply that raises its potential to a few tens of kilovolts. The gas near the needle points is ionized by the intense electric field, and in the resulting corona discharge the ions are driven to the surface of the belt. The motion of the belt carries the charge into the high-voltage terminal and transfers it to another comb of needles, from which it passes to the outer surface of the terminal. A carefully

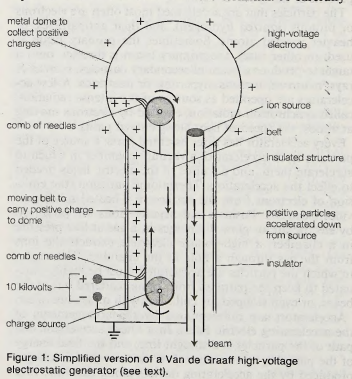


Figure 1: Simplified version of a Van de Graaff high-voltage electrostatic generator (see text).

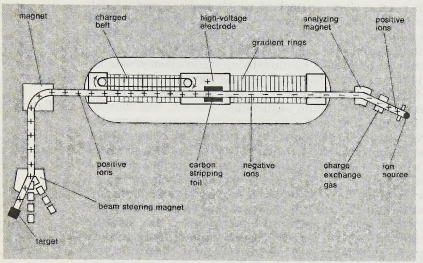


Figure 2: Two-stage tandem accelerator (see text).
Adapted from R. J. Van de Graaff, *Nuclear Instruments and Methods*, vol. 8, p. 195 (1960); North-Holland Publishing Co.

designed Van de Graaff generator insulated by pressurized gas can be charged to a potential of about 20 megavolts. An ion source within the terminal then produces positive particles that are accelerated as they move downward to ground potential through an evacuated tube.

Tandem Van de Graaff accelerators

In most constant-voltage accelerators, Van de Graaff generators are the source of high voltage, and most of the electrostatic proton accelerators still in use are two-stage tandem accelerators. These devices provide a beam with twice the energy that could be achieved by one application of the high voltage. Figure 2 is a diagram of a tandem accelerator. An ion source yields a beam of protons, which are accelerated to a low energy by an auxiliary high-voltage supply. This beam passes through a region containing a gas at low pressure, where some of the protons are converted to negative hydrogen ions by the addition of two electrons. As the mixture of charged particles moves through a magnetic field, those with negative charge are deflected into the accelerator tube, and those with positive charge are deflected away. The beam of negative ions is then accelerated toward the positive high-voltage terminal. In this terminal, the particles pass through a thin carbon foil that strips off the two electrons, changing many of the negative ions back into positive ions (protons). These, now repelled by the positive terminal, are further accelerated through the second part of the tube. At the output end of the accelerator, the protons are magnetically separated, as before, from other particles in the beam and directed to the target. In three- or four-stage tandem accelerators, two Van de Graaff generators are combined with the necessary additional provisions for changing the charge of the ions.

Van de Graaff and Cockcroft-Walton generators also are utilized for accelerating electrons. The rates at which charge is transported in electron beams correspond to currents of several milliamperes; the beams deliver energy at rates best expressed in terms of kilowatts. These intense beams are used for sterilization, industrial radiography, cancer therapy, and polymerization of plastics.

BETATRONS

A betatron is a type of accelerator that is useful only for electrons, which sometimes are called beta particles—hence the name. The electrons in a betatron move in a circle under the influence of a magnetic field that is caused to grow stronger as the energy of the electrons is increased. The magnet that produces the field on the electron orbit also produces a field in the interior of the orbit. The increase in the strength of this field with time produces an electric field that accelerates the electrons. If the average magnetic field inside the orbit is always twice as strong as the magnetic field on the orbit, the radius of the orbit remains constant, so that the acceleration chamber can be made in the shape of a torus, or doughnut, as shown in Figure 3. The poles of the magnet are tapered to cause the field near the orbit to weaken with increasing radius. This focuses the beam by causing any particle that strays from the orbit to be subjected to forces that restore it toward its proper path. The theory of this focusing was first worked out for the betatron; by analogy, the oscillations of particles about their equilibrium orbits in all cyclic accelerators are called betatron oscillations.

Betatron oscillations

Just after the sinusoidally varying strength of the magnetic field has passed through zero and starts increasing in the direction proper to guide the electrons in their circular orbit, a burst of electrons is sent into the doughnut, where—in a 20-MeV betatron—they gain about 100 eV per revolution and traverse the orbit about 200,000 times during the acceleration. The acceleration lasts for one-quarter of the magnet cycle until the magnetic field has reached its greatest strength, whereupon the orbit is caused to shrink, deflecting the electrons onto a target to produce a beam of intense X rays.

The practical limit on the energy imparted by a betatron is set by the emission of electromagnetic energy by electrons moving in curved paths. The intensity of this radiation, commonly called synchrotron radiation (see below *Synchrotrons: Electron synchrotrons*), rises rapidly as the speed of the electrons increases. The largest betatron accelerates electrons to 300 MeV, sufficient to produce pimesons in its target; the energy loss by its electrons through radiation (a few percent) is compensated by changing the relation between the field on the orbit and the average field inside the orbit. At higher energies this compensation would not be feasible.

Betatrons are now commercially manufactured, principally for use as sources of highly energetic ("hard") X rays for industrial and medical radiography.

CYCLOTRONS

The magnetic resonance accelerator, or cyclotron, was the first cyclic accelerator and the first resonance accelerator that produced particles energetic enough to be useful for nuclear research. For many years the highest particle energies were those imparted by cyclotrons modeled upon Lawrence's archetype. In these devices, commonly called classical cyclotrons, the accelerating electric field oscillates at a fixed frequency and the guiding magnetic field has a fixed intensity.

Classical cyclotrons. The key to the operation of a cyclotron is the fact that the orbits of ions in a uniform magnetic field are isochronous; that is, the time taken by a particle of a given mass to make one complete circuit is the same at any speed or energy as long as the speed is much less than that of light. (As the speed of a particle approaches that of light, its mass undergoes the increase—predicted by the theory of relativity—that is called relativistic increase.) This isochronicity makes it possible for a high voltage, reversing in polarity at a constant frequency, to accelerate a particle many times. As shown in Figure 4, an ion source is located at the centre of an evacuated chamber that has the shape of a short cylinder, like a pillbox, between the poles of an electromagnet that creates a uniform field perpendicular to the flat faces. The accelerating voltage is applied by electrodes, called dees from their shape: each is a D-shaped half of a pillbox. The source of the voltage is an oscillator—similar to a radio transmitter—that operates at a frequency equal to the

Isochronous ion orbits

From *Jaderna Energie* (1958)

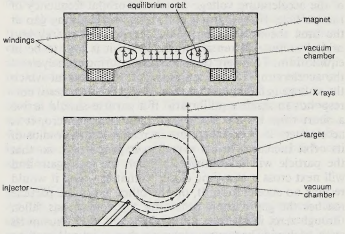


Figure 3: (Top) Cross section and (bottom) plan view of a betatron (see text).

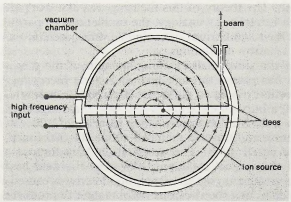


Figure 4: Plan view of the classical cyclotron.

frequency of revolution of the particles in the magnetic field. The electric fields caused by this accelerating voltage are concentrated in the gap between the dees; there is no electric field inside the dees. The path of the particle inside the dees is therefore circular. Each time the particle crosses the gap between the dees it is accelerated, because in the time between these crossings the direction of the field reverses. The path of the particle is thus a spirallike series of semicircles of continually increasing radius.

Some means of focusing is required, as otherwise a particle that starts out in a direction making a small angle with the orbital plane will spiral into the dees and be lost. While the energy of the particle is still low, this focusing is supplied by the accelerating electric fields; after the particle has gained significant energy, focusing is a consequence of a slight weakening of the magnetic field toward the peripheries of the dees, as in the betatron.

The energy gained by a particle in a classical cyclotron is limited by the relativistic increase in the mass of the particle, a phenomenon that causes the orbital frequency to decrease and the particles to get out of phase with the alternating voltage. This effect can be reduced by applying higher accelerating voltages to shorten the overall acceleration time. The highest energy imparted to protons in a classical cyclotron is less than 25 MeV, and this achievement requires the imposition of hundreds of kilovolts to the dees. The beam current in a classical cyclotron operated at high voltages can be as high as five milliamperes; intensities of this magnitude are very useful in the synthesis of radioisotopes.

Synchrocyclotrons. Cyclotrons in which the frequency of the accelerating voltage is changed as the particles are accelerated are called synchrocyclotrons, frequency-modulated (FM) cyclotrons, or phasotrons. Because of the modulation, the particles do not get out of phase with the accelerating voltage, so that the relativistic mass increase does not impose a limit on the energy. Moreover, the magnetic focusing can be made stronger, so that the magnetic field need not be so precisely shaped.

Evasion of relativistic constraint

Because of the phenomenon of phase stability, it is unnecessary to program the frequency of the accelerating voltage precisely to follow the decreasing frequency of revolution of the particles as they are accelerated. To see how phase stability affects the operation of a cyclotron, consider a particle moving in an orbit. Let the frequency of the accelerating voltage match the orbital frequency of this particle. If the particle crosses the accelerating gap at the time the accelerating voltage is zero, its energy and orbital radius will remain unchanged—it is said to be in equilibrium. There are two such times during each cycle of the accelerating voltage; only one of these (that at which the voltage is falling, rather than rising, through zero) corresponds to stable equilibrium. If a particle should arrive a short time before the voltage has fallen to zero, it is accelerated. Its speed therefore increases, but the radius of its orbit increases by an even larger proportion, so that the particle will take longer to reach the gap again and will next cross it at a time closer to that at which it would receive no acceleration. If, on the other hand, the particle reaches the gap a short time after the voltage has fallen through zero, its speed is diminished, and the radius of its orbit is diminished even more, so that it takes less time to reach the gap again, arriving—like the other particle—at a time closer to that at which it receives no acceleration. This phenomenon, by which the trajectories of errant particles are continually corrected, confers stability on the entire beam and makes it possible to accelerate the particles uniformly, by modulating the frequency, without dispersing them. The small periodic variations of the particles about the equilibrium values of phase and energy are called synchrotron oscillations.

Synchrotron oscillations

In the operation of a synchrocyclotron, particles are accelerated from the ion source when the frequency of the accelerating voltage is equal to the orbital frequency of the particles in the central field. As the frequency of the voltage falls, the particles, on the average, encounter an accelerating field. They oscillate in phase but around a value that corresponds to the average acceleration. The particles reach the maximum energy in bunches, one for each time the accelerating frequency goes through its program. The intensity of the beam is a few microamperes, much lower than that of a classical cyclotron.

Large synchrocyclotrons have been constructed in many countries. They are used primarily for research with secondary beams of pi-mesons. The practical upper limit of the energy of a synchrocyclotron, set by the cost of the huge magnets required, is about 1 GeV.

Sector-focused cyclotrons. The sector-focused cyclotron is another modification of the classical cyclotron that also evades relativistic constraint on its maximum energy. Its advantage over the synchrocyclotron is that the beam is not pulsed and is more intense. The frequency of the accelerating voltage is constant, and the orbital frequency of the particles is kept constant as they are accelerated by causing the average magnetic field on the orbit to increase with orbit radius. This ordinarily would cause the beam to spread out in the direction of the magnetic field, but in sector-focused cyclotrons the magnetic field varies with the angular position as well as with the radius; this produces the equivalent of alternating-gradient focusing (see below *Synchrotrons*). This principle was discovered in 1938 by Llewellyn H. Thomas, then at Ohio State University, but was not applied until the alternating-gradient synchrotron was invented in 1952. Several of these devices, sometimes called azimuthally varying field (AVF) cyclotrons, have been built for use in nuclear and medical research.

LINEAR RESONANCE ACCELERATORS

The technology required to design a useful linear resonance accelerator was not developed until after 1940. These accelerators require very powerful sources of radio-frequency accelerating voltage. Further, a practical linear accelerator for heavy particles, such as protons, must make use of the principle of phase stability.

Linear accelerators fall into two distinct types: standing-wave linear accelerators (used for heavy particles) and traveling-wave linear accelerators (used to accelerate electrons). The reason for the difference is that after electrons have been accelerated to a few megaelectron volts in the first few metres of a typical accelerator, they have speeds very close to that of light. Therefore, if the accelerating wave also moves at the speed of light, the particles do not get out of phase, as their speeds do not change. Protons, on the other hand, must reach much higher energies before their speeds can be taken as constant, so that the accelerator design must allow for the prolonged increase in speed.

Types of linear accelerators

Linear electron accelerators. The force that acts on electrons in a traveling-wave accelerator is provided by an electromagnetic field with a frequency near 3,000 MHz (1 MHz = 1,000,000 Hertz, or 1,000,000 cycles per sec-

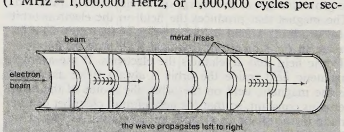


Figure 5: Waveguide acceleration chamber of a linear electron accelerator showing the irises, which decrease the phase velocity of the wave (see text).

ond)—a microwave. The acceleration chamber is an evacuated cylindrical pipe that serves as a waveguide for the accelerating field. The phase velocity of an electromagnetic wave in a cylindrical pipe is greater than the velocity of light in free space, so the wave must be slowed down by the insertion of metal irises a few centimetres apart in the pipe, as shown in Figure 5. In the intense field, the electrons gain about 2 MeV every 30 centimetres or so. The microwaves are produced by large klystrons (high-frequency vacuum-tube amplifiers) with power outputs of 20–30 megawatts. Because sources of radio-frequency power of this magnitude must be operated intermittently (they will not survive continuous service), the beams from these accelerators are delivered in short bursts.

Pulses of electrons are injected at energies of a few hundred keV (that is, speeds about half that of light). The accelerator is so designed that, during the first part of the acceleration, the electrons are caused to gather into bunches, which then are accelerated nearly to the speed of light. Subsequently, the electrons move with the crest of the electromagnetic wave.

Linear electron accelerators are manufactured commercially. They are used for radiography, for cancer treatment, and as injectors for electron synchrotrons.

SLAC

The 3.2-kilometre- (2-mile-) long linear electron accelerator at Stanford University (called SLAC, an acronym for *S*tanford *l*inear *a*ccelerator *c*entre) is the source of very energetic beams of electrons and positrons (Figure 6), up to a maximum of 50 GeV. The positrons are produced as secondary particles when the electron beam is allowed to strike a target one-third of the distance along the accelerator, and they are later fed back into the machine, alternately with electrons, for acceleration along its full length. At the far end of the accelerator, the electrons and positrons can be directed into the Stanford Linear Collider (SLC), which consists of two separate arcs of magnets forming a loop to bring the two beams into head-on collision at a total energy of about 100 GeV.

Linear electron accelerators constructed of superconducting materials have been developed. Such structures dissipate far less energy than conventional metal structures, allowing a continuous electron beam, rather than a pulsed beam, to be accelerated. This principle is being exploited to good effect at the Continuous Electron Beam Accelerator Facility (CEBAF) in Newport News, Va. This consists of two 250-metre linear accelerators joined at each end by semicircular arcs to form an oval "racetrack." Electrons are injected at 45 MeV and can be accelerated to energies of 4 GeV or more, the highest energies being reached after the beams have completed five circuits of the machine.

Linear proton accelerators. The design principle applied in linear accelerators for protons was originated by Luis Alvarez at Berkeley in 1946. It is based on the formation of standing electromagnetic waves in a long cylindrical metal tank, or cavity, as it is sometimes called. In the design that has been adopted, the electric field is parallel to the axis of the tank. Most of these accelerators operate at frequencies of about 200 MHz.

During the time required for a proton to traverse one of these tanks, the accelerating electric fields undergo many reversals of direction. In Alvarez' design, the decelerating effect of the field during the intervals when it opposes the motion of the particles is prevented by installing on the axis of the tank a number of "drift tubes," as shown in Figure 7. The electric field is zero inside the drift tubes, and, if their lengths are properly chosen, the protons cross the gap between adjacent drift tubes when the direction of the field produces acceleration and are shielded by the drift tubes when the field in the tank would decelerate them. The lengths of the drift tubes are proportional to the speeds of the particles that pass through them.

Use of drift tubes

It would appear that any error in the magnitude of the accelerating voltages would cause the particles to lose the synchronism with the fields needed for proper operation of the device, but the principle of phase stability reduces to a manageable magnitude the need for precision in construction. It also makes possible an intense beam because protons can be accelerated in a stable manner even if they do not cross the gaps at exactly the intended times.

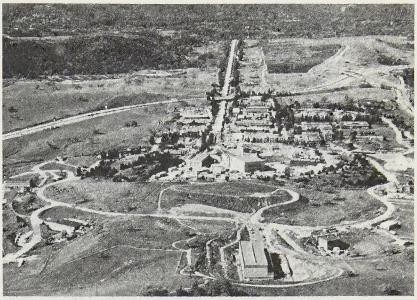


Figure 6: The Stanford Linear Accelerator Center has a 3.2-kilometre linear accelerator that produces the highest-energy electron beams in the world.
Stanford Linear Accelerator Center and the U.S. Department of Energy

The principle is the same as that of a synchrotron, except that the gap-crossing time for stable phase oscillations coincides with the rise, rather than the fall, of the voltage wave. If a proton arrives at the accelerating gap late, it receives a larger-than-normal increment of energy, enabling it to "catch up."

A very large amount of radio-frequency power is required to produce the accelerating voltages. This makes it necessary for linear proton accelerators to be operated in a pulsed mode. They are supplied with protons accelerated to about 750 keV by a Cockcroft-Walton generator. The entering beam passes through an accelerating radio-frequency cavity a short distance upbeam from the main linear accelerator, so that, as the particles pass through the first drift tubes, they are already bunched.

Injectors for proton synchrotrons

The intense pulses of protons emerging from linear accelerators make these devices ideal as injectors for proton synchrotrons. Their high cost has precluded their construction for other uses except as meson factories. The largest linear accelerators used as injectors are located at the Brookhaven National Laboratory, Upton, N.Y., and at the Fermi National Accelerator Laboratory ("Fermilab"), Batavia, Ill. These accelerators are very similar in construction. The beam energy is 200 MeV, and the peak beam current is more than 100 milliamperes. They are needed as injectors only for a short time every few seconds; most of the beams are used for radioisotope production and medical applications.

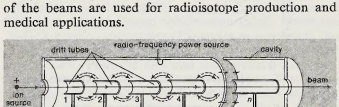


Figure 7: Linear proton resonance accelerator containing *n* metallic drift tubes (see text).

At the Clinton P. Anderson Meson Physics Facility (LAMPF) in Los Alamos, N.M., a linear proton accelerator has been constructed for nuclear research and as a meson factory. The protons are accelerated to 100 MeV in Alvarez-type tanks and then to 800 MeV in a standing-wave linear accelerator of the coupled-cavity type operated at a frequency of 805 MHz. The apparatus produces a beam carrying a current in excess of one milliampere, which delivers a power of more than 800 kilowatts. Its intense secondary beam of low-energy pi-mesons has been applied in experimental cancer therapy.

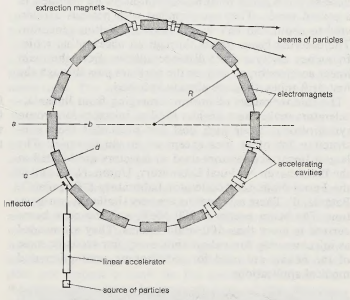
SYNCHROTRONS

As the particles in a synchrotron are accelerated, the strength of the magnetic field is increased to keep the radius of the orbit approximately constant. This technique has the advantage that the magnet required to form the

particle orbits is much smaller than that needed in a cyclotron to produce the same particle energies. The acceleration is effected by radio-frequency voltages, while the synchronism is maintained by the principle of phase stability. The rate of increase of the energy of the particles is set by the rate of the strengthening of the magnetic field. The peak accelerating voltage is ordinarily about twice as large as the average energy gain per turn would require, to provide the margin for phase stability. Particles can be stably accelerated with a range of energies and phases with respect to the accelerating voltage, and very intense beams can be produced.

The magnetic field must be shaped so as to focus the beam of particles. In early synchrotrons the field was caused to decrease slightly with increasing radius, as in a betatron. This arrangement resulted in a weak focusing effect that was adequate for machines in which the dimensions of the magnet gap could be appreciable in comparison with the radius of the orbit. The magnitude of the magnetic fields that may be used is limited by the saturation of the iron components that shape the field and provide a path for the magnetic flux. Therefore, if the energy of accelerators is to be increased, their radius must be increased correspondingly. For relativistic particles, the radius is proportional to the kinetic energy. The magnet of a synchrotron with weak focusing, designed to have a reasonable intensity, would have a mass proportional to the cube of the radius. It is clear that increasing the energy beyond some point—in practice, about 10 GeV—would be very expensive.

The introduction of alternating-gradient focusing provided the solution to this problem and made possible the development of synchrotrons with much higher energies.



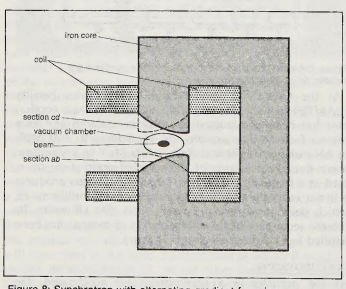


Figure 8: Synchrotron with alternating-gradient focusing, showing the placement of the two types of magnets. Shown at the bottom is a cross section of a typical magnet with the two different types of pole-tips used (see text).

The idea was promptly incorporated in the design of the 33-GeV proton synchrotron at the Brookhaven National Laboratory and the 28-GeV machine at the European Organization for Nuclear Research (CERN), near Geneva. Alternating-gradient focusing

The magnetic fields in an alternating-gradient synchrotron vary much more strongly with radius than those used for weak focusing. A magnet with pole-tips shaped as shown in cross section *ab* in Figure 8 produces a magnetic field that sharply decreases with increasing radius. To the particle beam, this magnetic field acts like a lens with a very short focal length. In the vertical direction (the orbital plane is horizontal) it focuses the beam, but in the radial direction it is almost equally defocusing. A magnet with the pole-tip shapes shown in cross section *cd* in Figure 8 produces a field that strongly increases with increasing radius. This field is defocusing in the vertical direction and focusing in the radial direction. Although pairing such magnetic fields results in partial cancellation, the overall effect is to provide focusing in both directions. The ring of magnetic field is created by a large number of magnets, with the two types of pole-tips alternating, as shown at the top of Figure 8. The beam, in effect, passes through a succession of lenses as the particles move around the ring, producing a large beam current in a vacuum chamber of small cross section.

Particles accelerated in a large synchrotron are commonly injected by a linear accelerator and are steered into the ring by a device called an inflector. They begin their acceleration in the ring when the magnetic field is small. As the field created by the ring magnets increases, the injection pulse is timed so that the field and the energy of the particles from the linear accelerator are properly matched. The radio-frequency accelerating devices, usually called cavities, operate on the same principle as a short section of a linear accelerator. The useful beam may be either the accelerated particles that have been extracted from the ring by special magnets or secondary particles ejected from a target that is introduced into the beam.

Electron synchrotrons. The invention of the synchrotron immediately solved the problem of the limit on the acceleration of electrons that had been imposed by the radiation of electrons moving in circular orbits. This radiation has been named synchrotron radiation because it was first observed during the operation of a 70-MeV electron synchrotron built at the General Electric Company research laboratory, Schenectady, N.Y. A betatron can accelerate electrons to 300 MeV only if the radiation is carefully compensated, but a synchrotron needs only a modest increase in the radio-frequency accelerating voltage. As the particles lose energy by radiation, their average phase with respect to the accelerating voltage simply shifts slightly so as to increase their average energy gain per revolution. Synchrotron radiation

Electron synchrotrons with energies near 300 MeV have been constructed in several countries, the first being the one built in 1949 at Berkeley under McMillan's direction. In these accelerators, the electrons were injected by a pulsed electron gun, and the initial acceleration from 50–100 keV to 2–3 MeV was induced as in a betatron. The magnets were specifically designed to provide the accelerating flux in the initial part of the magnet cycle; during this time, the speed of the electrons increased from about 50 percent of the speed of light to more than 95 percent. At this point, acceleration by the radio-frequency cavity supervened, and the small further change in speed was accommodated by a 5-percent change in the radius of the orbit.

Strong focusing was first applied to the electron synchrotron in the 1.2-GeV device built in 1954 at Cornell University in New York. All large electron synchrotrons now are equipped with linear accelerators as injectors. The practical limit on the energy of an electron synchrotron is set by the cost of the radio-frequency system needed to restore the energy the electrons lose by radiation. To minimize this energy loss, the acceleration time is made as short as possible (a few milliseconds) and the magnetic fields are kept weak. The weak fields keep down the energy loss by guiding the electrons on gently curved paths. However, because synchrotron radiation losses increase as

the fourth power of the energy, small increases in energy lead to large increases in radius.

The largest electron synchrotrons, used in research in particle physics, operate as colliding-beam storage rings (see below). At CERN the Large Electron-Positron collider (LEP) accelerates electrons and positrons to 50 GeV in a ring with a circumference of 27 kilometres, which is probably the practical limit for such machines.

Another way to reduce the energy used in an electron synchrotron is to employ superconducting radio-frequency accelerating cavities, which have smaller losses. These are used, for example, to accelerate electrons in the 6.3-kilometre ring of the electron-proton collider at the DESY (German Electronic Synchrotron) laboratory in Hamburg (see below). Superconducting cavities are also being used to double the energy of the beams in LEP.

Proton synchrotrons. The mode of operation of a proton synchrotron is very similar to that of an electron synchrotron, but there are two important differences. First, because the speed of a proton does not approach the speed of light until its energy is well above 1 GeV, the frequency of the accelerating voltage must be modulated to keep it proportional to the speed of the particle during the initial stage of the acceleration. Second, protons do not lose a significant amount of energy by radiation at energies attainable by present-day techniques. The limit on the energy of a proton synchrotron is therefore set by the cost of the magnet ring, which increases only as the first power of the energy or even more slowly. The highest-energy particle accelerators yet built are proton synchrotrons.

The first proton synchrotron to operate (1952) was the 3-GeV Cosmotron at Brookhaven. It, and other accelerators that soon followed, had weakly focusing magnets. The 28-GeV proton synchrotron at CERN and the 33-GeV machine at Brookhaven made use of the principle of alternating-gradient focusing but not without complications. Such focusing is so strong that the time required for a particle to complete one orbit does not depend strongly on the energy of the particle. Therefore, for the energy range (which may extend to several GeV) within which acceleration appreciably affects the speed of the particle, phase stability operates as it does in a linear accelerator: the region of stable phase is on the rising side of the time curve of the accelerating voltage. At higher energies, however, the speed of the proton is substantially constant, and the region of stable phase is on the falling side of the voltage curve, as it is in a synchrocyclotron. At the point that divides these regions, called the transition energy, there is no phase stability. At Brookhaven a model electron accelerator was built to demonstrate that the beam could be accelerated through the transition energy in a stable manner.

Fermilab Tevatron

A large proton synchrotron called the Tevatron at Fermilab was designed to operate at 1,000 GeV or 1 TeV, the energy that gives the device its name. The intense magnetic fields needed to guide and focus such an energetic proton beam are provided by magnets with windings made of a superconducting alloy, and the whole ring is kept at 4.5 kelvins by liquid helium. The Tevatron was built in the same tunnel as the main ring of conventional magnets of the 400-GeV accelerator that went into operation in 1972 and now serves as the injector for the Tevatron (Figure 9). The proton beam originates in a 750-keV Cockcroft-Walton generator and is accelerated to 200 MeV in a linear accelerator. The protons are then injected into a "booster" synchrotron in which they are accelerated to 8 GeV (99.4 percent of the speed of light). While the magnetic field of the main accelerator ring is held at a strength of 400 gauss, 12 pulses of protons are injected into the ring from the booster. The fields in the bending and focusing magnets are then strengthened as the protons are accelerated to 150 GeV in 1.5 seconds. The protons are then transferred to the ring of superconducting magnets for final acceleration. Figure 10 shows the sequence of processes occurring during one cycle in the operation of the main ring and the superconducting ring to produce and extract a beam of 800-GeV protons. The device completes one cycle per minute, including the extraction of the beam, which occupies about 20 seconds. Prolonging the beam

Ring of superconducting magnets

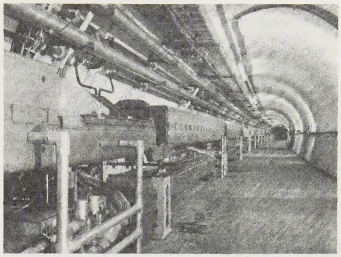


Figure 9: *The tunnel of the main 400-GeV accelerator at Fermilab.*
The upper ring of conventional magnets accelerates particles to 150 GeV. The particle beam can then be injected into the Tevatron, which is composed of superconducting magnets (lower ring) designed to accelerate protons and antiprotons to energies approaching 1 TeV (see text).
By courtesy of Fermi National Accelerator Laboratory

extraction over this extended interval is important for the proper interpretation of the experiments. In a typical experiment, several particles (called secondary particles) are ejected from a target when an accelerated particle (the primary particle) strikes it. The goal of the experiment is the study of either this production process or the properties of the secondary particles that result from it. These particles produce signals as they pass through detectors, and it is assumed that all the signals produced at the same time arise from particles coming from the interaction of a single primary particle with the target. If many accelerated particles arrived at the target within too short an interval, numerous interactions would occur at practically the same time, resulting in accidental coincidences that would obscure the effects being searched for.

The two large rings at Fermilab are 1,000 metres in radius; their circumference is 6.3 kilometres. In these rings the functions of bending the beam around its orbit and focusing the beam are separated. Each ring has 774 bending magnets; the main ring has 180 focusing magnets, the superconducting ring has 216. Injection, acceleration, and extraction of the beam take place in six straight sections free from magnets. The beam from this accelerator is divided to serve many experimental areas. Figure 11 shows a plan of the accelerator and experimental areas.

In 1976 a similar accelerator went into operation at CERN: the Super Proton Synchrotron (SPS), which the 28-GeV proton synchrotron (PS) serves as a booster. The SPS began operating at 400 GeV and was later upgraded to 450 GeV. Both the SPS and the Tevatron can operate as proton-antiproton colliders (see below). The CERN machine can also accelerate heavy ions (such as sulfur

By courtesy of John Crawford, Fermilab, Batavia, Ill.

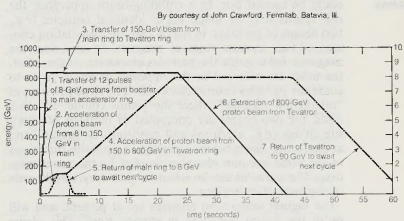


Figure 10: Successive stages comprising one cycle in the operation of the two large accelerating rings at the Fermi National Accelerator Laboratory. The solid line (right-hand scale) shows the number of protons circulating in the rings. The broken lines show the magnet currents, which indicate the energy of the protons (left-hand scale).

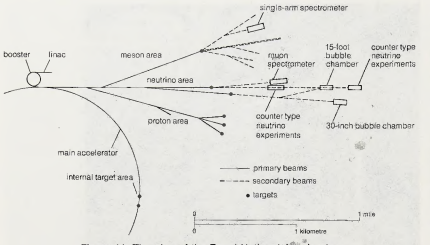


Figure 11: The plan of the Fermi National Accelerator Laboratory accelerator and experimental areas. The angles between the branches of the beam are exaggerated for clarity.
Adapted from *CERN Courier* with permission of Fermilab, Batavia, Ill.

ions), as well as electrons and positrons for injection into the LEP collider.

Proton synchrotrons are in operation in laboratories in several countries; they all are used for research into the properties of subatomic particles.

OTHER TYPES OF ACCELERATORS

Interactions of particle beams

Colliding-beam storage rings. Although particles are sometimes accelerated in storage rings, the main purpose of these rings is to make possible energetic interactions between beams of particles moving in opposite directions. When a moving object strikes an identical object that is at rest, at most half of the kinetic energy of the moving object is available to produce heat or to deform the objects; the remainder is accounted for by the motions of the objects after the encounter. If, however, the two objects are in motion in opposite directions with equal speeds, then all the kinetic energy is available to produce heat or deformation at the instant of collision. If the objects stick together, the combination is at rest after the collision. For particles with speeds close to that of light, the effect is accentuated. If a 400-GeV proton strikes a proton at rest, only 27.4 GeV are available for the interaction; the remainder produces motion of the particles. On the other hand, if two 31.4-GeV protons collide, 62.3 GeV are available for the interaction (the collision is not quite "head on").

In a target of liquid or solid matter, the number of particles per unit volume accessible to an accelerated beam is large, but, when the target of one beam is another beam, the number of particles interacting is much smaller: the rate of interactions is proportional to the product of the currents in the two beams. Kerst realized in 1956 that, though the beam current in a high-energy accelerator is small, the currents circulating in the magnet rings are effectively much larger because of the high orbital frequency of the particles. Thus, if the colliding beams are circulating in such rings, useful experiments on the interactions could be carried out. In a colliding-beam apparatus, the two beams may be made up of identical particles (*e.g.*, two beams of protons), in which case the installation consists of two separate rings of magnets. In one ring, the magnetic fields guide the particles clockwise; in the other, the fields are oriented in the opposite direction so as to guide the particles counterclockwise. The rings intersect at "interaction regions," where the beams collide. In other cases, the two beams are composed of particles of opposite charge (*e.g.*, electrons and positrons, or protons and antiprotons). Such beams circulate in opposite directions in the same vacuum chamber, guided by the same magnets. The particles are bunched so that they collide only in the interaction regions.

Colliding-beam systems

The highest interaction energies are at present, and will be in the future, achieved in storage rings. This places the research with them at the very forefront of the quest for knowledge, even though many types of experiments cannot be conducted with storage rings. This is true partly because the number of interactions in a storage ring is a small fraction of that occurring in a stationary target and partly because storage beams do not produce intense beams of secondary particles.

Electron storage rings. Many storage rings have been constructed to study the interactions of electrons with positrons. The principal centres of this research are Stanford University; CERN; Tsukuba, Japan; Orsay, Fr.; Hamburg, Ger.; Frascati, Italy; and Novosibirsk, Russia.

The manner of operation of a typical electron-positron colliding-beam device is shown in Figure 12. Since the signs of their electric charges are opposite, electrons and positrons circulate in opposite directions in a magnet ring. The figure shows a linear electron accelerator for filling the ring; a synchrotron can be used instead. The deflecting magnet sends electrons to an injection point on the ring so that they circulate clockwise. Positrons are created by high-energy electrons in a target, called a converter, and then accelerated by the second stage of the linear accelerator. The deflecting magnet sends them to an injection point so that they circulate counterclockwise. Because of the radiation of energy by the particles, it is necessary to have accelerating cavities if the energy of the particles is to be kept fixed. The accelerating system causes the particles to circulate in bunches so that the collisions take place in only a few places in the ring. This sparsity of intersections simplifies the operation by minimizing the disruption of each beam caused by interactions with the other, allowing

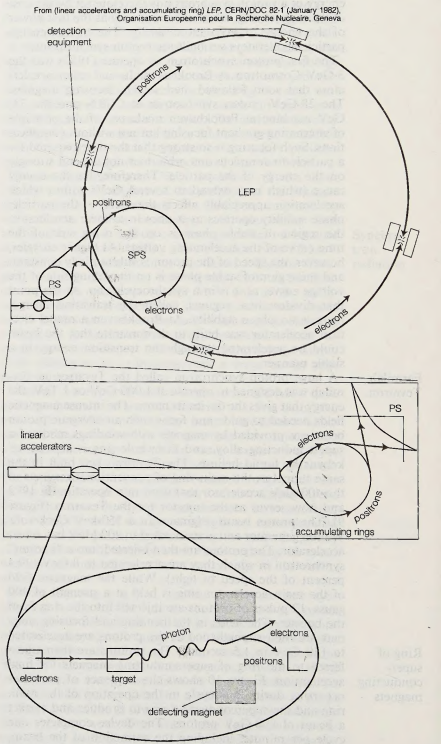


Figure 12: The injection scheme for the LEP collider (see text).

more intense beams to be collected and used in the ring. The detection equipment is positioned near the points where the beams intersect.

The highest-energy electron-positron collider is the LEP machine at CERN, which began operation in 1989. LEP reaches a maximum of 50 GeV per beam in a magnet ring that is 27 kilometres in circumference and that occupies a 4-metre-wide tunnel lying, on average, 100 metres underground. As shown in Figure 12, other accelerators built earlier at CERN act as injectors to LEP in a complex interlinked system. A purpose-built linear accelerator produces bunches of electrons and positrons at 600 MeV, as described above, and feeds them into the 28-GeV proton synchrotron where they are accelerated to 3.5 GeV. They are then transferred to the SPS for acceleration to 20 GeV before injection into LEP. In the final stage, LEP accelerates the counterrotating beams of electrons and positrons to a maximum energy of 50 GeV. The beams are then collided at four points around the ring where detectors are located.

The electrons and positrons in a storage ring emit synchrotron radiation at very great rates—more than a megawatt in some installations. From a high-energy storage ring, the wavelength of this radiation extends into the X-ray region. These storage rings now constitute the brightest sources of electromagnetic radiation available in the ultraviolet and X-ray regions. This radiation is proving to be increasingly useful for research in solid-state physics, biophysics, and chemical physics; a few electron storage rings of relatively low energy are operated solely for this purpose.

Source of X rays and ultraviolet radiation

Proton storage rings. In 1971 CERN pioneered the storage of protons with the Intersecting Storage Rings (ISR), in which two interlaced rings each stored protons at 31 GeV. The two beams collided at eight crossing points, giving a total collision energy of 62 GeV. This was equivalent to a stationary target being struck by a beam of 2 TeV. Such a high energy was possible because protons, unlike electrons, do not dissipate significant energy in the form of synchrotron radiation.

A decade later CERN reached much higher energies with a radical new technique, colliding protons with antiprotons that were accelerated and stored together in the ring of the 450-GeV SPS. Protons and antiprotons have opposite electric charge, so like electrons and positrons they can circulate in opposite directions around the same synchrotron ring. In this case, the creation of an intense beam of antiprotons required a technique known as "stochastic cooling," developed by Simon Van der Meer at CERN. Antiprotons at 3.5 GeV are produced at CERN by directing protons from the 28-GeV synchrotron onto a copper target. This process is very inefficient: about one million protons must strike the target to produce each antiproton. Furthermore, the antiprotons emerge from the target with random energies and directions. They are collected by magnetic fields into a diffuse beam that is injected into a small storage ring with a large cross-sectional area. As the antiprotons circulate in this ring, electronic devices sense the deviations from the desired orbit and apply corrective forces that focus the particles into a narrow beam of uniform energy. Although the SPS has collided protons and antiprotons at 450 GeV per beam, it has operated mainly at 315 GeV per beam, providing a total interaction energy of 630 GeV. Since 1987 the Tevatron has also operated as a proton-antiproton collider, in this case with beam energies as high as 1 TeV.

Electron-proton storage rings. The Hadron-Electron Ring Accelerator (HERA) at the DESY laboratory stores both electrons and protons. It is the only machine that operates in this way with particles of different masses. To do so requires two interlaced rings: one to accelerate and store the electrons, the other to accelerate and store the protons. The machine, which began operation in 1992, occupies a tunnel 6.3 kilometres in circumference. With high fields generated by superconducting magnets, the proton ring can reach energies up to 820 GeV. The electron energy, however, is limited by synchrotron radiation losses but reaches a maximum 30 GeV, with the aid of low-loss superconducting accelerating cavities.

Impulse accelerators. Primarily for use in research on thermonuclear fusion of hydrogen isotopes, several high-intensity electron accelerators have been constructed. One type resembles a string of beads in which each bead is a torus of laminated iron and the string is the vacuum tube. The iron toruses constitute the cores of pulse transformers, and the beam of electrons, in effect, forms the secondary windings of all of the transformers, which are connected in series. The primaries are all connected in parallel and are powered by the discharge of a large bank of capacitors. These accelerators produce beams with energies between 1 and 9 MeV and currents between 200 and 200,000 amperes. The pulses are very brief, lasting about 50 nanoseconds. Besides their application to thermonuclear fusion, such accelerators are utilized for flash radiography, research on collective ion acceleration, microwave production, and laser excitation.

Use in fusion research

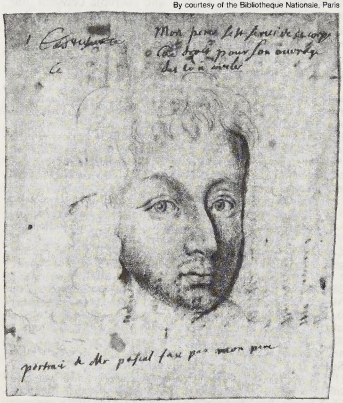
BIBLIOGRAPHY. The physics background is dealt with in FRANK CLOSE, MICHAEL MARTEN, and CHRISTINE SUTTON, *The Particle Explosion* (1987), at the general level, and at a deeper level by W.S.C. WILLIAMS, *Nuclear and Particle Physics* (1991); and EMILIO SEGRÈ, *Nuclei and Particles: An Introduction to Nuclear and Subnuclear Physics,* 2nd ed., completely rev. and enlarged (1977). An extensive survey for nonspecialists is available in the article "Particle Accelerator," in *McGraw-Hill Encyclopedia of Science & Technology,* 7th ed., vol. 13, pp. 114–153 (1992). Early research papers of historical interest include J.D. COCKCROFT and E.T.S. WALTON, "Experiments with High Velocity Positive Ions," *Proceedings of the Royal Society of London, Series A,* vol. 137, pp. 229–242 (1932), on the cascade generator; and a selection of articles in *Physical Review:* ROBERT J. VAN DE GRAAFF, "A 1,500,000 Volt Electrostatic Generator," 38:1919–20 (1931); D.W. KERST, "Acceleration of Electrons by Magnetic Induction," 60:47–53 (1941), on the betatron; ERNEST O. LAWRENCE and M. STANLEY LIVINGSTON, "The Production of High Speed Light Ions Without the Use of High Voltages," 40:19–35 (1932), on the classical cyclotron; DAVID H. SLOAN and ERNEST O. LAWRENCE, "The Production of Heavy High Speed Ions Without the Use of High Voltages," 38:2021–32 (1931), on the development of the resonance linear accelerator idea; EDWIN M. MCMILLAN, "The Synchrotron: A Proposed High Energy Particle Accelerator," 68:143–144 (1945); and D.W. KERST *et al.,* "Attainment of Very High Energy by Means of Intersecting Beams of Particles," 102:590–591 (1956). Discussions of some modern accelerators include MAURICE GOLDSMITH and EDWIN SHAW, *Europe's Giant Accelerator* (1977), a nontechnical history of the CERN 450-GeV proton synchrotron; ROBERT R. WILSON, "The Next Generation of Particle Accelerators," *Scientific American,* 242(1):42–57 (January 1980); M.C. CROWLEY-MILLING, "High-Energy Particle Accelerators," *Reports on Progress in Physics,* 46(1):51–95 (January 1983); three articles from *Scientific American:* JOHN R. REES, "The Stanford Linear Collider, "261(4):58–65 (October 1989); STEPHEN MYERS and EMILIO PICASSO, "The LEP Collider," 263(1):54–61 (July 1990); and LEON M. LEDERMAN, "The Tevatron," 264(3):48–55 (March 1991); and CHRISTINE SUTTON, "Particle Accelerators," a special 4-page insert in *New Scientist,* vol. 134, no. 1825 (June 13, 1992). More technical works include M. STANLEY LIVINGSTON and JOHN P. BLEWETT, *Particle Accelerators* (1962); JOHN J. LIVINGOOD, *Principles of Cyclic Particle Accelerators* (1961); A.A. KOLOMENSKY and A.N. LEBEDEV, *Theory of Cyclic Accelerators* (1966; originally published in Russian, 1962); PHILIP J. BRYANT and KJELL JOHNSEN, *The Principles of Circular Accelerators and Storage Rings* (1993); D.A. EDWARDS and M.J. SYPHERS, *An Introduction to the Physics of High Energy Accelerators* (1993); *AIP Conference Proceedings,* papers from seminars and courses published by the American Institute of Physics; and *IEEE Transactions on Nuclear Science* (bimonthly), for new developments in accelerator technology.

(Y.M.A./F.C.S./Ch.Su.)

Pascal

Blaise Pascal, a mathematician, physicist, religious philosopher, and master of French prose, laid the foundation for the modern theory of probabilities, formulated what came to be known as Pascal's law of pressure, and propagated a religious doctrine that taught the experience of God through the heart rather than through reason. The establishment of his principle of intuitionism had an impact on such later philosophers as Jean-Jacques Rousseau and Henri Bergson and also on the Existentialists.

By courtesy of the Bibliotheque Nationale, Paris



Pascal, red crayon drawing by Jean Domat, *c.* 1649. In the Bibliothèque Nationale, Paris.

Pascal's life to the Port-Royal years. Pascal was born on June 19, 1623, at Clermont-Ferrand, France, where his father, Étienne Pascal, was presiding judge of the tax court. His mother died in 1626, and in 1631 the family moved to Paris. Étienne, who was respected as a mathematician, devoted himself henceforth to the education of his children. While his sister Jacqueline (born in 1625) figured as an infant prodigy in literary circles, Blaise proved himself no less precocious in mathematics. In 1640 he wrote an essay on conic sections, *Essai pour les coniques,* based on his study of the now classical work of Girard Desargues on synthetic projective geometry. The young man's work, which was highly successful in the world of mathematics, aroused the envy of no less a personage than the great French Rationalist and mathematician René Descartes. Between 1642 and 1644, Pascal conceived and constructed a calculating device to help his father—who in 1639 had been appointed intendant (local administrator) at Rouen—in his tax computations. The machine was regarded by Pascal's contemporaries as his main claim to fame, and with reason, for in a sense it was the first digital calculator since it operated by counting integers. The significance of this contribution explains the youthful pride that appears in his dedication of the machine to the chancellor of France, Pierre Seguier, in 1644.

Until 1646 the Pascal family held strictly Roman Catholic principles, though they often substituted *l'honnêteté* ("polite respectability") for inward religion. An illness of his father, however, brought Blaise into contact with a more profound expression of religion, for he met two disciples of the abbé de Saint-Cyran, who, as director of the convent of Port-Royal, had brought the austere moral and theological conceptions of Jansenism into the life and thought of the convent. Jansenism was a 17th-century form of Augustinianism in the Roman Catholic Church. It repudiated free will, accepted predestination, and taught that divine grace, rather than good works, was the key to salvation. The convent at Port-Royal had become the centre for the dissemination of the doctrine. Pascal himself was the first to feel the necessity of entirely turning away from the world to God, and he won his family over to the spiritual life in 1646. His letters indicate that for several years he was his family's spiritual adviser, but the conflict within himself—between the world and ascetic life—was not yet resolved. Absorbed again in his scientific interests, he tested the theories of Galileo and Evangelista Torricelli (an Italian physicist who discovered the principle of the barometer). To do so, he reproduced and amplified experiments on atmospheric pressure by constructing mercury barometers and measuring air pressure, both in Paris and on the top of a mountain overlooking Clermont-Ferrand. These tests paved the way for further studies in hydrodynamics and hydrostatics. While experimenting, Pascal invented the syringe and created the hydraulic press, an instrument based upon the principle that became known as Pascal's law: pressure applied to a confined liquid is transmitted undiminished through the liquid in all directions regardless of the area to which the pressure is applied. His publications on the problem of the vacuum (1647–48) added to his reputation. When he fell ill from overwork, his doctors advised him to seek distractions; but what has been described as Pascal's "worldly period" (1651–54) was, in fact, primarily a period of intense scientific work, during which he composed treatises on the equilibrium of liquid solutions, on the weight and density of air, and on the arithmetic triangle: *Traité de l'équilibre des liqueurs et de la pesanteur de la masse de l'air* (Eng. trans., *The Physical Treatises of Pascal,* 1937) and also his *Traité du triangle arithmétique.* In the last treatise, a fragment of the *De Alea Geometriae,* he laid the foundations for the calculus of probabilities. By the end of 1653, however, he had begun to feel religious scruples; and the "night of fire," an intense, perhaps mystical "conversion" that he experienced on November 23, 1654, he believed to be the beginning of a new life. He entered Port-Royal in January 1655, and though he never became one of the solitaires, he thereafter wrote only at their request and never again published in his own name. The two works for which he is chiefly known, *Les Provinciales* and the *Pensées,* date from the years of his life spent at Port-Royal.

Contributions to physics and mathematics

Association with the Convent of Port-Royal

"Les Provinciales." Written in defense of Antoine Arnauld, an opponent of the Jesuits and a defender of Jansenism who was on trial before the faculty of theology in Paris for his controversial religious works, Pascal's 18 *Lettres écrites par Louis de Montalte à un provincial* deal with divine grace and the ethical code of the Jesuits. They are better known as *Les Provinciales.* They included a blow against the relaxed morality that the Jesuits were said to teach and that was the weak point in their controversy with Port-Royal; Pascal quotes freely Jesuit dialogues and discrediting quotations from their own works, sometimes in a spirit of derision, sometimes with indignation. In the two last letters, dealing with the question of grace, Pascal proposed a conciliatory position that was later to make it possible for Port-Royal to subscribe to the "Peace of the Church," a temporary cessation of the conflict over Jansenism, in 1668.

The *Provinciales* were an immediate success, and their popularity has remained undiminished. This they owe primarily to their form, in which for the first time bombast and tedious rhetoric are replaced by variety, brevity, taut-

ness, and precision of style; as Nicolas Boileau, the founder of French literary criticism, recognized, they marked the beginning of modern French prose. Something of their popularity, moreover, in fashionable, Protestant, or skeptical circles, must be attributed to the violence of their attack on the Jesuits. In England they have been most widely read when Roman Catholicism has seemed a threat to the Church of England. Yet they have also helped Catholicism to rid itself of laxity; and, in 1678, Pope Innocent XI himself condemned half of the propositions that Pascal had denounced earlier. Thus, the *Provinciales* played a decisive part in promoting a return to inner religion and helped to secure the eventual triumph of the ideas set forth in Antoine Arnauld's treatise *De la fréquente communion* (1643), in which he protested against the idea that the profligate could atone for continued sin by frequent communion without repentance, a thesis that thereafter remained almost unchallengeable until the French church felt the repercussion of the revocation of the Edict of Nantes (which had granted religious freedom to French Protestants) in 1685. Whereas the Jesuits seemed to represent a Counter-Reformation predominantly concerned with orthodoxy and obedience to ecclesiastical authority, the *Provinciales* advocated a more spiritual approach, emphasizing the soul's union with the Mystical Body of Christ through charity.

Morality and spirituality

Further, by rejecting any double standard of morality and the distinction between counsel and precept, Pascal aligned himself with those who believe the ideal of evangelical perfection to be inseparable from the Christian life. Although there was nothing original in these opinions, Pascal nevertheless stamped them with the passionate conviction of a man in love with the absolute, of a man who saw no salvation apart from a heartfelt desire for the truth, together with a love of God that works continually toward destroying all self-love. For Pascal, morality cannot be separated from spirituality. Moreover, his own spiritual development can be traced in the *Provinciales.* The religious sense in them becomes progressively refined after the first letters, in which the tone of ridicule is smart rather than charitable.

"Pensées." Pascal finally decided to write his work of Christian apologetics, *Apologie de la religion chrétienne,* as a consequence of his meditations on miracles and other proofs of Christianity. The work remained unfinished at his death. Between the summers of 1657 and 1658, he put together most of the notes and fragments that editors have published under the inappropriate title *Pensées* ("Thoughts"; Eng. trans., *Pensées,* 1962). In the *Apologie,* Pascal shows the man without grace to be an incomprehensible mixture of greatness and abjectness, incapable of truth or of reaching the supreme good to which his nature nevertheless aspires. A religion that accounts for these contradictions, which he believed philosophy and worldliness fail to do, is for that very reason "to be venerated and loved." The indifference of the skeptic, Pascal wrote, is to be overcome by means of the "wager": if God does not exist, the skeptic loses nothing by believing in him; but if he does exist, the skeptic gains eternal life by believing in him. Pascal insists that men must be brought to God through Jesus Christ alone, because a creature could never know the infinite if Jesus had not descended to assume the proportions of man's fallen state.

The second part of the work applies the Augustinian theory of allegorical interpretation to the biblical types (*figuratifs*); reviews the rabbinical texts, the persistence of true religion, the work of Moses, and the proofs concerning Jesus Christ's God-like role; and, finally, gives a picture of the primitive church and the fulfillment of the prophecies. The *Apologie (Pensées)* is a treatise on spirituality. Pascal was not interested in making converts if they were not going to be saints.

Pascal's apologetics

Pascal's apologetics, though it has stood the test of time, is primarily addressed to individuals of his own acquaintance. To convert his libertine friends, he looked for arguments in their favourite authors: in Michel de Montaigne, in the Skeptic Pierre Charron, in the Epicurean Pierre Gassendi, and in Thomas Hobbes, an English political philosopher. For Pascal, Skepticism was but a stage. Modernist theologians in particular have tried to make use of his main contention, that "man is infinitely more than man," in isolation from his other contention, that man's wretchedness is explicable only as the effect of a Fall, about which a man can learn what he needs to know from history. In so doing, they sacrifice the second part of the *Apologie* to the first, keeping the philosophy while losing the exegesis. For Pascal as for St. Paul, Jesus Christ is the second Adam, inconceivable without the first.

Finally, too, Pascal expressly admitted that his psychological analyses were not by themselves sufficient to exclude a "philosophy of the absurd"; to do so, it is necessary to have recourse to the convergence of these analyses with the "lines of fact" concerning revelation, this convergence being too extraordinary not to appear as the work of providence to an anguished seeker after truth (*qui cherche en gémissant*).

He was next again involved in scientific work. First, the "Messieurs de Port-Royal" themselves asked for his help in composing the *Élemenis de géométrie;* and second, it was suggested that he should publish what he had discovered about cycloid curves, a subject on which the greatest mathematicians of the time had been working. Once more fame aroused in him feelings of self-esteem; but from February 1659, illness brought him back to his former frame of mind, and he composed the "prayer for conversion" that the English clergymen Charles and John Wesley, who founded the Methodist Church, were later to regard so highly. Scarcely capable of regular work, he henceforth gave himself over to helping the poor and to the ascetic and devotional life. He took part intermittently, however, in the disputes to which the "Formulary"—a document condemning five propositions of Jansenism that, at the demand of the church authorities, had to be signed before a person could receive the sacraments—gave rise. Finally a difference of opinion with the theologians of Port-Royal led him to withdraw from controversy, though he did not sever his relations with them.

Pascal died on August 19, 1662, after suffering terrible pain, probably from carcinomatous meningitis following a malignant ulcer of the stomach. He was assisted by a non-Jansenist parish priest.

Assessment. At once a physicist, a mathematician, an eloquent publicist in the *Provinciales,* and an inspired artist in the *Apologie* and in his private notes, Pascal was embarrassed by the very abundance of his talents. It has been suggested that it was his too concrete turn of mind that prevented his discovering the infinitesimal calculus; and in some of the *Provinciales* the mysterious relations of human beings with God are treated as if they were a geometrical problem. But these considerations are far outweighed by the profit that he drew from the multiplicity of his gifts; his religious writings are rigorous because of his scientific training; and his love of the concrete emerges no less from the stream of quotations in the *Provinciales* than from his determination to reject the vigorous method of attack that he had used so effectively in his *Apologie.*

MAJOR WORKS

MATHEMATICS, LOGIC, AND THE FOUNDATIONS OF SCIENCE: *Essai pour les coniques* (1640); *Lettre sur le sujet de la machine inventée par le sieur B.P. pour faire toutes sortes d'opération d'arithmétique* (1645); *De l'autorité en matière de philosophie,* the first editor's title of Pascal's preface to a projected *Traité du vuide,* consisting of a general statement of the principle of scientific research (written 1647, printed 1779); *Traité du triangle arithmétique avec quelques autres petits traités sur la même matière* (written 1654, printed 1665); *De l'esprit géométrique—de l'art de persuader* (written *c.* 1658, first printed in two parts 1728 and 1776); *Histoire de la roulette, appellée autrement trochoïde ou cycloïde* (1658); *Lettres de A. Dettonville, contenant quelques-unes de ses inventions de géométrie* (1658–59).

PHYSICS: *Expériences nouvelles touchant le vuide* (1647); *Récit de la grande expérience de l'équilibre des liqueurs* (1648); *Traites de l'équilibre des liqueurs et de la pesanteur de la masse de l'air* (written 1651, printed 1663; *The Physical Treatises of Pascal: The Equilibrium of Liquids and the Weight of the Mass of the Air,* 1937).

RELIGIOUS PHILOSOPHY AND CONTROVERSY: *Abrégé de la vie de Jésus-Christ* (written 1654 or 1655, printed 1846); *Lettre escritte à un Provincial par un de ses amis sur le sujet des disputes présentes de la Sorbonne* (January 1656, followed by

17 more pamphlets on the same themes down to March 1657), all 18 being subsequently republished together as *Les Provinciales* (1657; augmented edition, with supplementary pamphlets, 1659); *Projet de mandement contre l'Apologie pour les casuistes* (written 1658, printed 1779); *Ecrits sur la grâce* (four treatises drafted between 1656 and 1658 and printed from 1779 onward); *Pensées de M. Pascal sur la religion et sur quelques autres sujets* (1670; the first Eng. trans. was *Monsieur Pascal's Thoughts, Meditations, and Prayers*, 1688; numerous new translations and versions have appeared since then); *Prière pur demander à Dieu le bon usages des maladies* (written 1659, printed 1666 and 1670).

OTHER WORKS: *Trois discours sur la condition des grands* (comprised 1660, printed 1670). Collected editions of Pascal's works include letters of mathematical or spiritual interest.

COLLECTED EDITIONS: Considerable editions of Pascal's *Oeuvres complètes* were undertaken by the Abbé Charles Bossut, 5 vol. (1779); and by Léon Brunschvicg, Pierre Boutroux, and Félix Gazier, 14 vol. (1908–25). The handiest edition is by LOUIS LAFUMA, *l'Intégrale* (1960). Collections in English include *The Miscellaneous Writings of Pascal* (1849); *Thoughts, Letters, Minor Works* (1910); and *Provincial Letters, Pensées, Scientific Treatises*, Encyclopædia Britannica "Great Books of the Western World" (1952).

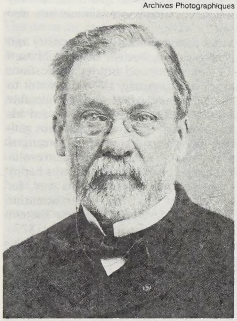
BIBLIOGRAPHY. A standard edition of Pascal's works is that by LÉON BRUNSCHVICG, PIERRE BOUTROUX, and FÉLIX GAZIER, 14 vol. (1908–25); in part replaced by the edition of JEAN MESNARD, 2 vol. (1964–72); and by *Pensées sur la religion*, ed. by LOUIS LAFUMA, 3rd ed., 3 vol. (1960). See also the editions by HUGH F. STEWART of *Pensées* (1950), containing an English translation, and of *Provinciales* (1920, reissued 1951). Biographical and critical studies include: JEAN MESNARD, *Pascal: His Life and Works* (1952; originally published in French, 1951), an excellent biography, with bibliography, *Pascal* (1965), and *Pascal et les Roannez*, 2 vol. (1965); LUCIEN JERPHAGNON, *Le Caractère de Pascal* (1962); PHILIPPE SELLIER, *Pascal et saint Augustin* (1970); ROBERT J. NELSON, *Pascal: Adversary and Advocate* (1982), a biographical study concentrating on linguistics, theology, and personal development; ALBAN J. KRAILSHEIMER, *Pascal* (1980), an introductory biography for the general reader; JAN MIEL, *Pascal and Theology* (1970), a study of his theology of grace; HUGH M. DAVIDSON, *The Origins of Certainty* (1979), an introduction to the *Pensées*; ÉMILE CAILLIET, *The Clue to Pascal* (1943, reprinted 1973); and *Blaise Pascal, l'homme et l'oeuvre* (1956), a collection of articles and discussions from the proceedings of a congress on Pascal held at Royaumont, France, in 1954.

(J.Or./L.J./Ed.)

Pasteur

The scientific contributions of Louis Pasteur, French chemist and microbiologist, were among the most varied and valuable in the history of science and industry. It was he who proved that microorganisms cause fermentation and disease; he who originated and was the first to use vaccines for rabies, anthrax, and chicken cholera; he who saved the beer, wine, and silk industries of France and other countries; he who performed important pioneer work in stereochemistry; and he who originated the process known as pasteurization.

Archives Photographiques



Pasteur.

Born December 27, 1822, at Dole in eastern France, Pasteur was the descendant of generations of tanners. His great-grandfather had been an indentured labourer who had purchased his freedom. In his youth Pasteur showed little interest in anything but drawing and produced a number of pastels, portraits of his parents and friends. After attending primary and secondary schools in Arbois, where his family had moved, and then in Besançon, Pasteur earned his *bachelier ès lettres* (bachelor of arts) in 1840 and *bachelier ès sciences* (bachelor of science) at the Royal College in Besançon in 1842. The following year he was admitted to the École Normale Supérieure, the famous teachers' college in Paris. He became *licencié ès sciences* (master of science) in 1845, and, after acquiring an advanced degree in physical sciences, he won his *docteur ès sciences* (doctor of philosophy) in 1847. On May 22, 1848, at the age of 26, he presented before the Paris Academy of Sciences a paper reporting a remarkable discovery he had just made—that certain chemical compounds were capable of splitting into a "right" component and a "left" component, one component being the mirror image of the other. His discoveries arose out of a crystallographic investigation of tartaric acid, an acid formed in grape fermentation that is widely used commercially, and racemic acid—a new, hitherto unknown acid that had been discovered in certain industrial processes in the Alsace region. Both acids not only had identical chemical compositions but also had the same structure; yet they showed marked differences in properties. The German chemist Eilhardt Mitscherlich (1794–1863) had shown that while ordinary commercial tartaric acid affects the rotation of plane polarized light, the unknown acid had no such effect. With the help of his own chemical methods Pasteur supplied the clue to this enigma by showing that the salts of the racemic acid consisted of two types of crystals that were mirror images of one another (like right- and left-hand gloves). When separated the two types of crystals rotated plane polarized light to the same degree but in opposite directions (one to the right, or clockwise, and the other to the left, or counterclockwise). One of the two crystal forms of racemic acid proved to be identical with the tartaric acid of fermentation. As Pasteur showed further, one component of the racemic acid (that identical with the tartaric acid from fermentation) could be utilized for nutrition by micro-organisms, whereas the other, which is now termed its optical antipode, was not assimilable by living organisms. On the basis of these experiments, Pasteur elaborated his theory of molecular asymmetry, showing that the biological properties of chemical substances depend not only on the nature of the atoms constituting their molecules but also on the manner in which these atoms are arranged in space.

Stereochemical investigations

In 1848 Pasteur was appointed professor of physics at the Dijon Lycée (secondary school) but was shortly called to the University of Strasbourg as professor of chemistry. There, on May 29, 1849, he married the daughter of the rector of the university, Marie Laurent, by whom he was to have five children, only two of whom survived childhood.

In 1854 Pasteur became dean of the new science faculty at the University of Lille, where he initiated a highly modern educational concept: by instituting evening classes for the many young workmen of the industrial city, conducting his regular students around large factories in the area, and organizing supervised practical courses, he demonstrated the relationship that he believed should exist between theory and practice, between university and industry. At Lille, after receiving a query from an industrialist on the production of alcohol from grain and beet sugar, Pasteur began his studies on fermentation. In the course of his analysis he once again encountered—though in liquid form—new "right" and "left" compounds. From studying the fermentation of alcohol he went on to the problem of lactic fermentation, showing yeast to be an organism capable of reproducing itself, even in artificial media, without free oxygen—a concept that became known as the Pasteur effect.

Fermentation research

In 1857 he was named Director of Scientific Studies at the École Normale Supérieure. He continued his researches and announced that fermentation was the result of the activity of minute organisms and that when fermentation failed, either the necessary organism was absent or was unable to grow properly. Before this discovery, all explanations of fermentation had lacked experimental foundation; Pasteur showed that milk could be soured by injecting a number of organisms from buttermilk or beer but could be kept unchanged if such organisms were excluded.

He was elected to the Academy of Sciences in 1862, and the following year a chair at the École des Beaux-Arts was established for him for a new and original program of instruction in geology, physics, and chemistry applied to the fine arts.

As a scholar engaged in research, Pasteur eventually found his administrative duties as Director of Scientific Studies at the École Supérieure too irksome. He gave up the post in 1867, and, thanks to the support of Emperor Napoleon III, a laboratory of physiological chemistry was created for him at the same institution. As a logical sequel to his work on fermentation, he began research on spontaneous generation (the concept that bacterial life arose spontaneously), a question which at that time divided scientists into two opposing camps. Pasteur's recognition of the fact that both lactic and alcohol fermentations were hastened by exposure to air led him to wonder whether his invisible organisms were always present in the atmosphere or whether they were spontaneously generated. By means of simple and precise experiments, including the filtration of air and the exposure of unfermented liquids to the air

of the high Alps, he proved that food decomposes when placed in contact with germs present in the air, which cause its putrefaction, and that it does not undergo transformation or putrefy in such a way as to spontaneously generate new organisms within itself.

After laying the theoretical groundwork, Pasteur proceeded to apply his findings to the study of vinegar and wine, two commodities of great importance in the economy of France; his pasteurization process, the destruction of harmful germs by heat, made it possible to produce, preserve, and transport these products without their undergoing deterioration.

Pasteurization

In 1865 he undertook a government mission to investigate the diseases of the silkworm, which were about to put an end to the production of silk at a time when it comprised a major section of France's economy. To carry out the investigation, he moved to the south of France, the centre of silkworm breeding. Three years later he announced that he had isolated the bacilli of two distinct diseases and had found methods of preventing contagion and of detecting diseased stock.

In 1870 Pasteur devoted himself to the problem of beer. Following an investigation conducted both in France and among the brewers in London, he devised, as he had done for vinegar and wine, a procedure for manufacturing beer that would prevent its deterioration with time. British exporters, whose ships had to sail entirely around the African continent, were thus able to send British beer as far as India without fear of its deteriorating.

Although Pasteur was partially paralyzed in 1868 and applied for retirement from the university, he continued his researches. In 1873 he was elected a member of the Academy of Medicine, and in 1874 the French Parliament provided him with an award that would ensure his material security while he pursued his work.

When in 1881 he had perfected a technique for reducing the virulence of various disease-producing microorganisms, he succeeded in vaccinating a herd of sheep against the disease known as anthrax. Likewise, he was able to protect fowl from chicken cholera, for he had observed that once animals stricken with certain diseases had recovered they were later immune to a fresh attack. Thus, by isolating the germ of the disease and by cultivating an attenuated, or weakened, form of the germ and inoculating fowl with the culture, he could immunize the animals against the malady. In this he was following the example of the English physician Edward Jenner in his method for vaccinating animals against cowpox.

On April 27, 1882, Pasteur was elected a member of the Académie Française, at which point he undertook research that proved to be the most spectacular of all—the preventive treatment of rabies. After experimenting with inoculations of saliva from infected animals, he came to the conclusion that the virus was also present in the nerve centres, and he demonstrated that a portion of the medulla oblongata of a rabid dog, when injected into the body of a healthy animal, produced symptoms of rabies. By further work on the dried tissues of infected animals and the effect of time and temperature on these tissues, he was able to obtain a weakened form of the virus that could be used for inoculation. Having detected the rabies virus by its effects on the nervous system and attenuated its virulence, he applied his procedure to man; on July 6, 1885, he saved the life of a nine-year-old boy, Joseph Meister, who had been bitten by a rabid dog. The experiment was an outstanding success, opening the road to protection from a terrible disease. In 1888 the Pasteur Institute was inaugurated in Paris for the purpose of undertaking fundamental research, prevention, and treatment of rabies. Pasteur, although in failing health, headed the institute until his death on September 28, 1895.

Assessment

Louis Pasteur brought about a veritable revolution in the 19th-century scientific method. By abandoning his laboratory and by tackling the agents of disease in their natural environments, he was able through his investigations to supply the complete solution to a given question, not only identifying the agent responsible for a disease but also indicating the remedy.

A skillful experimenter endowed with great curiosity and a remarkable gift of observation, Pasteur devoted himself with immense enthusiasm to science and its applications to medicine, agriculture, and industry. He was prompt to defend his ideas with courage and often with considerable harshness—in writings as well as in speech—toward his opponents. It was chiefly in his work on spontaneous generation and on rabies that he encountered the strongest opposition to his ideas (which were, for the time, revolutionary) from medical circles and the press. He was happy to accept the glory and honours that came his way, for he was well aware of his own value and of his scientific successes. A great friendship developed between Pasteur and the renowned British surgeon Sir Joseph Lister (1827–1912), who was quick to apply to his own discipline the discoveries of his French colleague.

BIBLIOGRAPHY. Studies of Pasteur's life and work include ÉMILE DUCLAUX, *Pasteur: The History of a Mind* (1920, reissued 1973; originally published in French, 1896), a scientific and philosophical work written by a collaborator of Pasteur; FRANÇOIS DAGOGNET, *Méthodes et doctrine dans l'œuvre de Pasteur* (1967), primarily a detailed work of methodology; RENÉ J. DUBOS, *Louis Pasteur, Free Lance of Science* (1950, reissued 1986), more philosophical than scientific; ELIE METCHNIKOFF (LI. MECHNIKOV), *The Founders of Modern Medicine: Pasteur, Koch, Lister* (1939, reprinted 1971; originally published in French, 1933), written by an important scholar who worked with Pasteur; JACQUES NICOLLE, *Louis Pasteur: A Master of Scientific Enquiry* (1961; originally published in French, 1953), and *Louis Pasteur: The Story of His Major Discoveries* (1961), both works giving a complete authoritative review of Pasteur's discoveries, and *Pasteur: sa vie, sa méthode, ses découvertes* (1969), an account of Pasteur's life and work; and RENÉ VALLERY-RADOT, *The Life of Pasteur*, 2 vol. (1902, reissued 1960; originally published in French, 1900), written by Pasteur's son-in-law, who was also his secretary, a fundamental work on the life of Pasteur but weak from the scientific point of view.

(J.-M.-R.N./Ed.)

The Apostle Paul

The Apostle Paul is an outstanding figure in the history of Christianity. Converted only a few years after the death of Jesus, he became the leading Apostle (missionary) of the new movement and played a decisive part in extending it beyond the limits of Judaism to become a worldwide religion. His surviving letters are the earliest extant Christian writings. They reveal both theological skill and pastoral understanding and have had lasting importance for Christian life and thought.

Sources. There are no reliable sources for Paul's life outside the New Testament. The primary source is his own letters. Of these, Romans, I and II Corinthians, and Galatians are indisputably genuine. Most scholars also accept Philippians, I Thessalonians, and Philemon. Opinion is divided about Ephesians, Colossians, and II Thessalonians. The Pastoral Letters (I and II Timothy and Titus) are held by many scholars to have been written considerably later than the time of Paul. The story of Paul's conversion and missionary career is given in Acts, probably written many years after his death. Some sections dealing with sea journeys may be derived from the diary of a companion of Paul. Traditionally this was thought to be Luke, the evangelist and author of Acts, a view still held by a number of scholars. For further information on the sources for Paul's life, see BIBLICAL LITERATURE AND ITS CRITICAL INTERPRETATION: *New Testament literature.*

Alinari/Art Resource, NY



St. Paul preaching the gospel, detail of a 12th-century mosaic in the Cappella Palatina, Palermo, Sicily.

Life. *Early life.* Paul was a Jew, born, perhaps in AD 10, at Tarsus, a city in Cilicia on the main trade route between East and West, and the home of famous Stoic philosophers. Like many of the Jews there he inherited Roman citizenship, probably granted by the Romans as a reward for mercenary service in the previous century. This fact explains his two names. He used his Jewish name, Saul, within the Jewish community and his Roman surname, Paul, when speaking Greek. Though he had a strict Jewish upbringing, he also grew up with a good command of idiomatic Greek and the experience of a cosmopolitan city, which fitted him for his special vocation to bring the gospel to the Gentiles (non-Jews). At some stage he became an enthusiastic member of the Pharisees, a Jewish sect that promoted purity and fidelity to the Law of Moses. According to Acts, he received training as a rabbi in Jerusalem under Gamaliel I. His knowledge of the Law and of rabbinic methods of interpreting it is evident in his letters. Like most rabbis he supported himself with a manual trade—tent making—probably learned from his father. It is clear that he never met Jesus while in Jerusalem, if, indeed, he was there before the Crucifixion. He learned enough about Jesus and his followers, however, to regard the Christian movement as a threat to the Pharisaic Judaism that he had embraced so eagerly. Thus he first appears on the scene of history as a persecutor of the newly founded church.

Training as rabbi

Serious persecution of Christians first arose in connection with converts among the Hellenists (Greek-speaking Jews) in Jerusalem. When one of them, Stephen, was stoned to death, the murderers "laid down their garments at the feet of a young man named Saul" (Acts 7:58). At that time Paul shared the sense of outrage aroused by the Hellenist converts. They had not only proclaimed Jesus as the Messiah and heavenly Lord, a man who had been crucified and therefore accursed by God (Deut. 21:23), but they also claimed that the temple and its sacrifices were superseded by the sacrificial death of Jesus and that therefore the Law could be disregarded (the subject of another curse, Deut. 27:26). Paul thus joined in the effort to stamp out the Christian movement. The Hellenist converts fled to the foreign cities where they had family connections, while the original Aramaic-speaking group in Jerusalem kept a low profile to avoid giving provocation.

The stoning of Stephen

Conversion. Paul, in Galatians, bears out the impression given in Acts that he was converted as a result of a vision on the road to Damascus, on his way to apprehend some of the scattered converts. His own account is tantalizingly brief: "he who had set me apart before I was born, and had called me through his grace, was pleased to reveal his Son to me, in order that I might preach him among the Gentiles" (Gal. 1:15–16). The longer description in Acts, given three times, dramatizes what may have been essentially an inward experience. It was certainly a moment of revelation, changing Paul from bitter enmity to lifelong dedication to the Christian cause.

Paul's conversion has often been explained psychologically as the resolution of an inner conflict. But the notion that Paul was tormented by scruples rests on a misunderstanding of Rom. 7. This chapter is concerned not with autobiography but with universal experience seen in the light of mature Christian understanding. Paul would not have spoken in these terms before his conversion. In fact, it is clear from other passages that his early life was free from such struggle. He excelled in zeal for the Law, and by its standards his life was blameless.

Explanations for Paul's conversion

Paul's own account is much more in keeping with Old Testament callings of a prophet. Though it is impossible to state exactly what happened, the central feature was certainly his vision of Jesus in glory. It convinced him that Jesus was risen from the dead and exalted as Lord in heaven, as the Christians claimed. It also was proof that Jesus had been crucified wrongfully. Hence the curse did not apply, and his death could be understood as a sacrifice on behalf of others.

To Paul this had universal significance. Believing, like many Jews of his time, that God's final Day of Judgment, on which he would come to free the world from evil and to establish lasting peace and righteousness, was imminent, Paul then saw his vocation to be a missionary to people of

every nation to prepare them for God's coming. The new feature of this expectation was the place accorded to Jesus Christ. In agreement with the earliest apostolic preaching, Paul believed that Jesus, having died for the sins of mankind, was now reserved in heaven as God's agent for the judgment. Those that believed in him and acknowledged him as Lord would have him as their deliverer on that day. Thus faith in Christ became the foundation of Paul's preaching. Along with this he proclaimed the love of God shown in the sacrificial death of Christ, who "loved me and gave himself for me" (Gal. 2:20). All his devotion was transferred to this new centre. Formerly his energy had been directed to preparing people for God's Kingdom by imposing on them strict Pharisaic interpretation of the Law. Now all that seemed useless in the light of what God himself had done for humanity through Jesus. Henceforth his one aim was to proclaim the faith of Jesus as Lord everywhere.

The cornerstone of Paul's message

Immediately after his conversion Paul spent a period of solitude in Arabia. He then took up residence in Damascus. There presumably he established contact with the Christians he had originally planned to harm and received from them information about Jesus and his teaching as well as experience of Christian fellowship. Damascus was the base for his first missionary work, but nothing is known of the effects of his mission in the region.

Paul in Antioch. After three years his work in Damascus came to an abrupt end. Somehow he had fallen foul of the ethnarch (governor) of the region of Nabataean Arabia. The ethnarch set a watch on the gates of Damascus, but Paul escaped over the wall in a basket and made his way to Jerusalem. There he met Peter, the Apostle, and James, the Lord's brother. This was an important meeting, for it established Paul as a recognized Apostle alongside the founders of the church at Jerusalem. The visit was brief, and Paul did not meet the Christian communities in the vicinity. Most likely this was due to the danger of reprisals from the Pharisees, who regarded Paul as a renegade. Therefore, after only two weeks, he set out on a new mission to Cilicia and Syria, with a base in his native city of Tarsus. About this mission, again, there is no information.

At some point Paul moved to Antioch, the capital of Syria, to assist Barnabas in his successful mission there. The converts included a large number of Gentiles. This eventually led to a serious crisis, in which Paul emerged as the champion of the Gentiles. The controversy, which lasted several years, stimulated Paul's most important contribution to Christian theology. His stand on behalf of the Gentiles ensured that Christianity became not just a Jewish sect but a universal religion. The point at issue was the relationship between Jewish and Gentile Christians. Primitive Christianity was a closely knit fellowship with the common meal and the Eucharist (thanksgiving for the sacrificial death of Christ) at the heart of it. But the Jewish purity rules made Jews reluctant to eat with Gentiles for fear of transgressing the Law. Jesus had taught that purity of heart was more important than attention to rules, but this did not lead his followers to abandon them. But at Antioch the accession of Gentile converts created a mixed congregation, in which the Jewish members were content to eat with the Gentiles for the sake of Christian fellowship. In Jerusalem, however, since the death of Stephen, the Christians had had to take great care not to offend Jewish susceptibilities, and the prospect of making headway in the mission there depended on their being seen as faithful to the Law. Thus reports of the liberal attitude of the Christians in Antioch were bound to be extremely damaging. Some of the Jerusalem Christians who were converted Pharisees even held the view that Gentile converts should be required to accept circumcision and the obligations of the Law.

Champion of the Gentiles

Paul states in Galatians that he did not revisit Jerusalem for 14 years, and, when he finally did so, it was to deal with the problem of Gentile membership of the church. This conflicts with the information in Acts, which tells of a visit by Paul and Barnabas to bring relief during a famine at some time in AD 47–49. Acts then describes a further visit to deal with the Gentile issue. Most scholars today identify the latter visit with that described in Galatians. This means that Luke, in writing Acts on the basis of various sources, either presented twice what was actually one visit or wrongly included Paul's name in the earlier relief visit.

Antioch continued to be Paul's base for further pioneering work. Acts records three itineraries, generally referred to as missionary journeys, spanning a number of years. The second visit to Jerusalem probably took place at the end of the first of these.

First missionary journey. Acts describes how Paul and Barnabas, accompanied by Barnabas' cousin John Mark, set out for Cyprus, visiting Salamis and Paphos. They then crossed to the mainland (modern Turkey), landing at Perga (near modern Murtana), but Mark left them and returned to Jerusalem. They worked in Pisidia and Pamphylia, which formed the southern part of the Roman province of Galatia, beginning in Pisidian Antioch (near modern Yalvaç). Acts records a sermon that Paul preached in the synagogue, which is a fine specimen of the presentation of the faith to a Jewish audience in New Testament times. After further stops at Iconium (modern Konya), Lystra (near modern Hatunsaray), and Derbe (unidentified), they retraced their steps to Perga and the port of Attalia (modern Antalya) and then sailed back to Antioch.

It is unclear from this account how many of the new converts were drawn from local Jewish communities and how many were Gentiles. The monotheism and strong morality of the Jews always attracted to the synagogues Gentiles who proved to be receptive to the Christian mission, especially as Paul did not require circumcision and observance of the Law for Christian fellowship. In some places the new congregations may have been entirely composed of Gentiles.

At this time Greek and Roman traditional religion was losing its hold, and a deputation had come from Jerusalem to Antioch to insist that the Gentile converts should be circumcised. This led to Paul's second visit to Jerusalem. Paul says that he and Barnabas went "by revelation," perhaps meaning as a result of a message from a prophet, not in response to a summons from Jerusalem as stated in Acts. The party from Antioch included Titus, a Gentile whom Paul had taken into his mission team.

Second visit to Jerusalem

It is almost impossible to harmonize the information in Acts 15 and Gal. 2, but it is best to regard them as accounts of the same occasion. In Jerusalem there seem to have been three main actions. First, Paul and Barnabas had a private consultation with James, Peter, and John, in which they compared the content of their mission preaching and established that they were in basic agreement. This confirmed Paul's contention that the gospel message did not require the circumcision of Gentile converts. A campaign by the hard-line party to have Titus circumcised was firmly resisted. Second, a larger conference was convened in order to inform all about the Gentile mission so that they should have no doubt that the power of the Holy Spirit had been at work. This resulted in the decision that the Gentile mission should continue without pressure to Judaize converts. Paul would carry this on from Antioch, while Peter would continue the mission among Jews from the base at Jerusalem. Paul, however, was urged to bear in mind the precarious position of the Jerusalem church. Third, a letter was sent to Antioch with minimum rules for Gentile converts: to abstain from meat used in pagan sacrifices, to use only kosher meat according to Jewish custom, and to observe Jewish restrictions on sexual relationships. Later events show that the contents of this letter were unknown to Paul, and it is conjectured that it belongs to a later attempt to regulate relationships with the numerous Jewish Christian congregations of Judaea and Syria after Paul had ceased to have close contact with Antioch.

Paul's view had been endorsed by Peter, who subsequently visited the church in Antioch. Apparently he had no difficulty in sharing in the life of the mixed congregation. Yet when some hard-liners came from Jerusalem, Peter felt compelled to withdraw from meals with Gentile members. Other Jewish members also yielded to the pressure, including even Barnabas. Paul, however, was



The missionary journeys of St. Paul.

adamant in his conviction that this was fundamentally wrong. This crisis could never have arisen if the letter from Jerusalem had already been sent; it must have been due to differing views of the implications of what had been agreed. Not only Paul but also Peter and the main body in Jerusalem had assumed that the purity rules would not be allowed to interfere with table fellowship in mixed congregations. But it is clear from the trouble over Titus that the hard-liners would demand separation into two groups and then claim that the unity of the congregation would require Judaizing of the Gentile converts. Paul insisted on his own understanding of the agreement, and the visitors left.

Demands to Judaize Gentile converts

Second missionary journey. Paul then planned to revisit the churches of south Galatia. Barnabas wished to take Mark, but Paul refused in view of his previous failure. Barnabas and Mark went to Cyprus, and nothing more is said about them in Acts. The subsequent account is entirely concentrated on Paul, who took with him Silas, also a Roman citizen (Roman name Silvanus). They went overland to Galatia. At Lystra Paul took into his team Timothy, a Gentile with a Jewish mother, who is mentioned with Silas in Paul's letters. The claim of Acts that Paul circumcised him seems improbable in view of the earlier decisions but is not impossible if the work was mainly among Jewish communities.

Because Paul hoped to establish the church in large centres of influence, he planned to go to Ephesus, the principal city of the province of Asia and a port on the Aegean coast. He was, however, prevented from doing so "by the Holy Spirit" (perhaps another reference to Christian prophecy). Instead he turned toward the large cities of Bithynia in the north. Possibly the Gentile churches of north Galatia, to which the letter to the Galatians is addressed, were founded on the way. Once more his plans were prevented, and so he moved northwest to Troas. From there, in response to a vision, he sailed to Macedonia and founded churches at Philippi, Thessalonica (modern Thessaloníki, Greece), and Beroea. Philippi, a Roman colony on the Via Egnatia, the major route across Greece, produced a loyal group of Gentile converts, who frequently contributed funds to Paul in later years. Acts tells how Paul and Silas were imprisoned there but released when they revealed their Roman citizenship. At Thessalonica and Beroea trouble from hostile Jews compelled Paul to move on to Athens. After a short stay there, during which he is said to have addressed the council of the Areopagus, he went on to Corinth. The speech, as given in Acts, was an attempt to meet the needs of a philosophically trained audience. No church was founded in Athens.

The events of that time are reflected in I Thessalonians, perhaps the earliest of Paul's letters, written after Silas and Timothy had joined Paul at Corinth. The letter expressed his great anxiety for this newly founded church in Thessalonica, which he had had to leave hurriedly, having been accused of treason for proclaiming Christ as a rival emperor. It emerges from the letter that he had taught the Gentile audience to turn "to God from idols, to serve a living and true God, and to wait for his Son from heaven, whom he raised from the dead, Jesus who delivers us from the wrath to come" (1:9–10). This can be taken as a good example of Paul's basic mission preaching. Timothy had reported that the converts were anxious about their fate because some of them had already died. Paul explained that the time of Christ's coming (Parousia) for judgment was unknown, but both living and dead who had faith in him would be claimed by him as his own and saved for the everlasting kingdom. II Thessalonians is regarded by some as a supplementary letter, written shortly afterward, but there is doubt about its authenticity. It contains details of the events that are to precede the Parousia (unfortunately these details are by no means easy to understand).

Success in Corinth

Paul was in low spirits when he reached Corinth after the failure at Athens. At Corinth he met a Jewish couple, Aquila and Priscilla, tentmakers like himself, who became his lifelong friends. They had recently come from Rome, following an edict of the emperor Claudius expelling all Jews from the capital. Possibly they had already become Christians in Rome. In Corinth Paul at last was able to exercise a long and fruitful ministry in a great trading centre. Acts records an incident in which Paul was brought before the proconsul Gallio. This is important for dating Paul's career because an inscription discovered at Delphi proves that Gallio began his year of office in AD 51. Paul had probably arrived in the previous year. When he left Corinth, Aquila and Priscilla accompanied him to Ephesus, but he went on alone by sea to Caesarea for Jerusalem and from there to Antioch.

Third missionary journey. Paul had by then established churches in Asia Minor and Greece, with a major centre at Corinth, and had begun work in the equally important Ephesus. Then followed a period of consolidation.

He went overland to Ephesus, which became his base for the next three years. Acts gives little detail, but he must have founded the churches at Colossae, Hierapolis, and Laodicea in the Lycus Valley during this period. A group of followers of John the Baptist at Ephesus is mentioned, and there were probably other Christian missionaries working in the same region. References in his letters to fighting wild beasts at Ephesus and to imprisonment show that he faced great hazards.

Active period of correspondence

This was the period of Paul's most important letters. His correspondence with Corinth shows the grave difficulties that were liable to arise. I Corinthians refers to a previous letter urging the Christians not to associate with immoral persons, but it has not survived. In I Corinthians Paul tackles a whole array of problems. Rival groups were claiming the authority of different teachers (Peter, Apollos, and Paul himself). A case of incest had gone unrebuked. Paul's teaching on freedom from the Law had been twisted to justify licentiousness. There were problems of marriage and divorce. The question of which foods a Gentile Christian might eat was causing problems of conscience. There was disorderly conduct at the Eucharist (Lord's Supper). In dealing with these matters Paul showed knowledge of Jesus' teaching on marriage, and he gave the account of the Last Supper in its oldest known form. A section on the gifts of the Holy Spirit includes his famous chapter on love (chapter 13) and regulates the practice of speaking with tongues. A long section on resurrection shows that, while teaching that Christian life was already participation in the risen Christ, Paul still thought that the Parousia was near and that the full experience of eternal life lay beyond this event.

Before long, however, there were fresh troubles at Corinth. Intruders from another church were trying to undermine Paul's authority. He dashed to Corinth but failed to restore confidence. He returned to Ephesus and wrote a severe letter (possibly partly preserved in II Cor. 10–13), which he regretted as soon as Titus had left with it. Paul had intended to work at Troas but was so anxious about Corinth that he went on to Macedonia instead in the hope of meeting Titus on his return. Titus returned with the good news that the severe letter had accomplished its purpose. With tremendous relief Paul wrote II Corinthians (perhaps only chapters 1–9), which is full of the theme of reconciliation: "God was in Christ reconciling the world to himself, not counting their trespasses against them, and entrusting to us the message of reconciliation" (5:19). Paul also gave further teaching on the resurrection of the body in terms of renewal and transformation into the state of glory.

Another theme of II Corinthians is a collection for the poor church of Jerusalem, a gift that Paul intended to symbolize the unity between the Jewish and Gentile churches. Behind this project was the continuing problem of the Judaizing party. This comes to the fore in Galatians, probably written during this period. The letter is concerned with the attempt of some Jewish Christians to persuade the Gentile Christians of Galatia to be circumcised and keep the Law. Here Paul lays out his doctrine of justification by faith, generally reckoned his most important contribution to Christian theology, which was to reach its classic expression in Romans.

Justification by faith

From Macedonia Paul went to Corinth, and it was during his three months there that he wrote to the Christians in Rome. The letter was written ostensibly to seek their help in his plan to evangelize the far west (Spain is mentioned) after taking the collection to Jerusalem. In fact, he clearly felt the need to win their support for his position on the Judaizing issue, and he presented the case at length. God's plan, he argued, is for universal salvation. This is God's gift available through faith in the sacrificial death of Christ. By itself the Law cannot bring salvation. It can show the nature of human sin but is powerless to make people righteous. Paul's opponents feared that without the Law the Gentile converts would be liable to libertine behaviour (as had happened at Corinth). Paul replied that faith in Christ opens the believer to the sanctifying power of the Holy Spirit. Then the opponents complained that Paul's argument left no room for the privileged position of the Jews as God's chosen people. Paul replied that, though many Jews had failed to respond to the gospel, the success of the mission to the Gentiles would prompt them to seek salvation at the end of time, "and so all Israel [would] be saved" (Rom. 11:26). Then the universe would reach the fulfillment of its purpose, and the final transformation could begin.

Arrest and imprisonment. At the end of the letter Paul expressed his fear of danger from the Jews in Jerusalem and even hinted that the church there might not feel able to accept the collection. It seems that both these fears were realized. Acts tells that Paul was accompanied by delegates from the Gentile churches but does not mention the collection. This omission is best explained on the assumption that Luke did not wish to say that the church in Jerusalem did not dare to accept it. If so, Paul's hope that it would symbolize the gathering of the Gentiles into the one family of God was disappointed. In Jerusalem Paul was mistakenly accused of bringing one of the Gentile delegates into the inner courts of the Temple, beyond the barrier excluding Gentiles. He was arrested, partly to save his life from the mob, but given good treatment on account of his Roman citizenship. When a plot against his life came to light, he was removed to Caesarea, the Roman military headquarters. The governor Felix kept him in prison to avoid antagonizing the Jewish authorities. Two years later Felix's successor, Festus, wanted to send him to Jerusalem for trial, but Paul refused to go and appealed to Caesar.

The journey to Rome began in late autumn, but a shipwreck delayed the travelers for three months at Malta, so that they arrived in Rome in the spring of AD 60. There Paul was kept under house arrest for two years awaiting trial. At this point the narrative of Acts closes, and it is left to the reader to guess what happened. As long as the Pastoral Letters were accepted as genuine, their evidence demanded the hypothesis of acquittal, further work in Greece, Asia Minor, and even Crete, before a second arrest, return to Rome, and sentence to death. Now that these letters are recognized to be pseudonymous, there is no reason to suppose that Paul was acquitted at all.

Paul wrote several letters during captivity. These might have been written during an earlier imprisonment in Ephesus or, perhaps, while he was at Caesarea, but Rome seems most likely. Of the four captivity letters, Philippians and Philemon are generally accepted as genuine; Colossians and Ephesians are questioned. The letter to Philemon, a Christian of Colossae, concerns his runaway slave whom Paul has converted in prison and now sends back to him "no longer as a slave but more than a slave, as a beloved brother" (verse 16). This letter, with its sensitive handling of a delicate situation, is a gem among the Pauline writings. Philippians is a serene acknowledgement of the generosity of the Christians at Philippi. Colossians is concerned with trouble from false teachers at Colossae, conjectured to be an unorthodox fringe sect of Judaism. In response, Christ is presented as the true wisdom of God, embodying his whole plan of salvation. Ephesians is an eloquent, perhaps overly rhetorical, statement of the privilege of the Gentiles, who in Christ enjoy the status of God's chosen people. Through his death Christ "has made us both one, and has broken down the dividing wall of hostility" (2:14).

Letter to Philemon

Achievement and influence. Paul's lasting monument is the worldwide Christian Church. Though he was not the first to preach to the Gentiles, his resolute stand against the Judaizing party was decisive for future progress. It can be justly claimed that it was due to Paul more than anyone else that Christianity grew from being a small sect within Judaism to become a world religion.

Paul's influence continued after his death. The Pastoral Letters to Timothy and Titus were written in Paul's name to promote fidelity to his teaching, probably around the end of the 1st century. At the same time, Paul's surviving letters were collected for general circulation. They quickly became a standard of reference for Christian teaching. In particular, theories of atonement (the reconciliation of mankind to God through the sacrificial death of Christ) have always relied heavily on Paul.

In the Western (Latin) half of Christendom Paul had a profound effect upon the history of the church through the

writings of St. Augustine. The Pelagian controversy concerning grace and free will turned on the interpretation of passages in Paul's letter to the Romans. In arguing for the necessity of divine grace for salvation, Augustine built on Paul's idea of predestination, correctly interpreting Paul's idea as a reference to God's predestined plan of universal salvation and as a concept that did not necessarily conflict with the exercise of free will.

The reformers of the 16th century were also deeply indebted to Paul. Martin Luther seized on the doctrine of justification by faith and made the distinction between faith and works the basis of his attack on the late medieval church. John Calvin drew from Paul his concept of the church as the company of the elect, using the idea of predestination and adding that predestination to salvation belongs only to the elect. Thus Paul's teaching came through the influence of Augustine to dominate the Reformation and its legacy in the Lutheran and Calvinist churches of modern Protestantism. These issues, however, never had the same prominence in the Eastern Orthodox churches.

Modern study of Paul has tried to reach behind these controversies and to see Paul in his true context of the rise of Christianity. Once the basis of Paul's thought in the context of Jewish concepts of his time is understood in the light of modern scholarship, uncompromising predestinarian views of some of Calvin's followers can be seen to be an overly rigid interpretation of Paul's meaning. Attempts to derive Paul's ideas from Greek or Gnostic influences have been largely abandoned. Paul stands out more clearly as a Christian Jew, whose conversion experience convinced him that Christ was the universal Lord under God, the agent and leader of God's kingdom. Paul thus maintained that through Christ every barrier is broken down: "There is neither Jew nor Greek, there is neither slave nor free, there is neither male nor female; for you are all one in Christ Jesus" (Gal. 3:28).

BIBLIOGRAPHY. From the immense Pauline bibliography, which includes many commentaries on the Acts and letters, only a representative selection can be chosen. GUSTAV ADOLF DEISSMANN, *St. Paul*, trans. from the German (1912); T.R. GLOVER, *Paul of Tarsus* (1925, reissued 1938); and JAMES S. STEWART, *A Man in Christ* (1935, reissued 1975), are valuable biographical studies. H.J. SCHOEPS, *Paul* (1961, reissued 1979; originally published in German, 1959), is a modern Jewish view of Paul. MICHAEL GRANT, *Saint Paul* (1976, reissued 1982), is an introduction to his life and his writings for the general reader. GÜNTHER BORNKAMM, *Paul* (1971; originally published in German, 1969); and FREDERICK F. BRUCE, *Paul: Apostle of the Heart Set Free* (1978, reprinted 1984), are a critical and a conservative account, respectively, of both his life and thought. WAYNE A. MEEKS, *The First Urban Christians* (1983), is a reinterpretation of Pauline thought in light of new information on its social and political contexts. Other specialized studies include WILLIAM M. RAMSAY, *St. Paul the Traveller and the Roman Citizen* (1896, reissued 1982, with an essay on the author), a valuable contribution to historical geography; ROBERT JEWETT, *A Chronology of Paul's Life* (1979); JOHANNES MUNCK, *Paul and the Salvation of Mankind* (1959; originally published in German, 1954), dealing with various aspects of Paul's relationship with the other Apostles; ALBERT SCHWEITZER, *The Mysticism of Paul the Apostle* (1931, reissued 1968; originally published in German, 1930); W.D. DAVIES, *Paul and Rabbinic Judaism*, 4th ed. (1980); E.P. SANDERS, *Paul and Palestinian Judaism* (1977), and *Paul, the Law, and the Jewish People* (1983), providing a balanced discussion of modern assessments of Paul's relationship to Judaism. D.E.H. WHITELEY, *The Theology of St. Paul*, 2nd ed. (1974); and J. CHRISTIAAN BEKER, *Paul the Apostle* (1980), are both thorough works on Paul's teaching.

(B.L.)

Peking (Beijing)

Peking (Wade–Giles romanization: Pei-ching, Pinyin: Beijing), the capital of the People's Republic of China, is one of the world's great cities. Few cities have served for so long as the political headquarters and cultural centre of a country as immense as China. The city has been an integral part of China's history over the past eight centuries, and there is scarcely a major building of any age in Peking that has not made its contribution to the evolution of the country as a whole. The importance of Peking thus makes it impossible to understand China without a knowledge of this city as well.

More than 2,000 years ago, a site north of present-day Peking was already an important military and trading centre for the northeastern frontier of China. In 1267, during the Yüan (Mongol) dynasty (1206–1368), a new city built northeast of the old—called Ta-tu—became the administrative capital of China. During the reigns of the first two emperors of the Ming dynasty (1368–1644), Nanking was the capital, and the old Mongol capital was renamed Pei-p'ing ("Northern Peace"); the third Ming emperor, however, restored it as the imperial seat of the dynasty and gave it a new name, Peking ("Northern Capital"). Peking has remained the capital of China except for a brief period (1928–49) when the Nationalist government again made Nanking the capital (although the capital was removed to Chungking during World War II); during that time Peking once again resumed the old name Pei-p'ing, which is still used by the Nationalist government in Taiwan.

The city remained the most flourishing cultural centre in China despite the frequent political changes in the nation throughout the early decades of the 20th century; Peking's importance was fully realized, however, only when the city was chosen as the capital of the People's Republic of China in 1949, and this political status has added much vitality to it. Indeed, few cities have ever had such rapid growth in population and geographic area, as well as in industrial and other activities. Combining both historical relics of an ancient culture and new urban construction, from fast food franchises to plush hotels for foreign tourists and corporate travelers, Peking has become a showplace of modern China and one of the world's great cities.

This article is divided into the following sections:

Physical and human geography 468
- Character of the city 468
- The landscape 468
  - The city site
  - Climate
  - Plant life
  - The city layout
  - Housing
  - Architecture
- The people 473
- The economy 473
  - Agriculture
  - Industry
  - Commerce and finance
  - Transportation
- Administration and social conditions 475
  - Government
  - Public utilities
  - Health
  - Education
- Cultural life 477
  - Museums and libraries
  - The arts
  - Recreation

History 478
- The early empires 478
- Centuries of growth 479
  - The Ming and Ch'ing dynasties
  - The modern city

Bibliography 480

Physical and human geography

CHARACTER OF THE CITY

Although much of Peking's older and more picturesque character has been destroyed in the drive since 1949 to modernize and industrialize, some parts of the city are still redolent of the past. Many fine monumental buildings, old restaurants, and centres of traditional Chinese arts and crafts remain, and the central government has taken measures to prevent Peking's further industrialization. Broad new boulevards replete with new commercial ventures have displaced the colourful stalls and markets for which the city was once famous, but the neighbourhood life of old Peking can still be glimpsed in the narrow *hu-t'ung*s (residential alleys), with their coal-burning stoves, tiny potted-plant gardens, and enclosed courtyards—some of which are still guarded by carved stone lions at their gates.

People in Peking commute by subway, bus, automobile, or bicycle and on hot summer evenings sit outside their apartment blocks to catch cooling breezes and to chat. The citizenry has a wide range of leisure pursuits, particularly those considered good for health. The ancient art of T'ai Chi ch'uan (Chinese boxing) is widely practiced, singly or in groups, along roadsides and in parks. Locals as well as tourists are attracted to the many nearby historical sites, such as the Summer Palace, the tombs of the Ming emperors, and the Great Wall. Older people, especially the men, like to huddle in tiny restaurants and tea shops or to sit alone at the Peking equivalent of corner taverns nursing a glass of beer or mao tai (a strong liquor made from sorghum). Young people are drawn to the city's many cafés and nightclubs, where the entertainment can range from DJ-run dance music to Chinese rock bands.

For all of the vicissitudes of its history, Peking continues to be a source of great pride for its inhabitants. Their obsessions are, as they have been for centuries, food and knowledge: they eat heartily when they have the means and read voraciously. Food stalls on the streets, selling a variety of cooked treats, are well-patronized, as are newspaper and magazine kiosks. The ambition of most families is to provide their offspring with a higher education or, if not that, a good job. Despite the shortage of living space in the city, families still try to accommodate parents and grandparents and to be responsible for the well-being of other elderly relatives.

THE LANDSCAPE

The city site. Peking is situated at the northern apex of the triangular North China Plain and lies at an elevation between about 100 and 130 feet (30 and 40 metres) above sea level. To the north lies the fringe of the Mongolian Plateau; to the northeast rises the mountain range of the Yen Mountains, a portion of which forms the eastern part of the concave arc that circles the Peking lowland from the northeast to the southwest to form what is known to geologists as the "Bay of Peking."

The city was built at the mouth of this embayment, which opens onto the great plain to the south and east, and between two rivers, the Yung-ting and the Ch'ao-pai, which eventually join to empty into the Po Hai (Gulf of Chihli), some 100 miles (160 kilometres) southeast of Peking. Peking is a natural gateway on the long-distance land communication route between the North China Plain and the northern ranges, plains, and plateaus. To the south of

the city the plain spreads out for about 400 miles until it merges into the lower valley and the delta of the Yangtze. On the east the plain is bounded by the sea, except for the break caused by the Shantung Hills; on the west it is flanked by the T'ai-hang Mountains, which constitute the eastern edge of the Loess Plateau (loess is loamy material deposited by wind).

Because Peking stands at the apex of the triangle, routes running across the great plain naturally converge on the city. In addition, since the dawn of Chinese history, the Yen range has constituted a formidable barrier between the North China Plain to the south, the Mongolian Plateau to the north, and the Liao River Plain in the southern region of the Northeast (formerly Manchuria). A few passes, however, cut through the ranges—the most important being the Nankow (Chü-Yung), Ku-pei, and Shan-hai—and are so situated that all roads leading from Mongolia and the Northeast to the North China Plain are bound to converge on Peking. For centuries, therefore, Peking has been an important terminus of the caravan routes leading to and from the vast Central Asian hinterland.

The metropolitan district

Since the early 15th century, the city of Peking and its surrounding territories have been organized as a metropolitan district of enormous size, having a governor—formerly appointed by the emperor himself—equal in rank to a provincial governor. This special district organization was continued by the Ch'ing dynasty and by the People's Republic of China. The present metropolitan boundary was established in 1959 and covers a territory of approximately 6,900 square miles (17,900 square kilometres). The metropolis consists of 10 urban and suburban *ch'u* (districts) and eight annexed *hsien* (counties) in peripheral areas. It may be divided into three concentric zones, based on urban functions. The central zone coincides with the central city; it is mainly occupied by old palaces, government buildings, commercial districts, and old residential areas and makes up roughly 1 percent of the total metropolitan area. Four of the *ch'u* lie within the central city. The second zone, the near suburb, comprises five *ch'u* immediately surrounding the old walled city; it is the site of the newer factories, schools, government buildings, and workers' dormitories. The outer fringe of this zone is intensively cultivated and supplies vegetables and fruits to the population of the central zone. The near suburb accounts for about 8 percent of the metropolitan area. The third zone, the far suburb, consists of one *ch'u* and eight annexed *hsien* and constitutes the remainder of the metropolitan area. This zone functions as the economic base—supplying coal, lumber, construction materials, vegetables and fruits, dairy products, water, and some grain crops to the urban population in the central zone and the near suburb. The far suburb of Peking is mainly under the administration of several *hsiang* (townships).

Climate. Though the distance from Peking to the Po Hai is only about 100 miles, the general air circulation in Peking is mainly from the northwest throughout the year. Maritime effects on Peking's weather are meagre. The climate is clearly of the continental monsoon type that occurs in the temperate zone. Local topography has a great effect on Peking's climate. Being located in the lowland area of the nation and shaded by mountains, the city is a little warmer in winter than other areas of China located on the same latitude; it has only five months each year in which the mean monthly temperature is lower than 50° F (10° C). Wind direction in Peking is also influenced by topography, with changes occurring from day to night. Generally, there are more southerly winds in the day but northerly or northwesterly winds at night.

Temperature and rainfall

The annual mean temperature of the city is 53° F (12° C). The coldest month is January, when monthly mean temperatures are 24° F (−4° C), and the warmest month is July, when the monthly mean is 79° F (26° C). In an average year, there are 132 days of freezing temperature in Peking (between October and March); the mean amount of precipitation is 25 inches (635 millimetres), with most of the annual total falling from June to August. July is ordinarily the wettest month of the year, with monthly mean precipitation averaging about nine inches.

One of the characteristics of Peking's rainfall is its variability. In 1959—an extremely wet year for Peking—the total precipitation amounted to 55 inches, whereas in 1891—an extremely dry year—only seven inches of rainfall was recorded. The average number of rainy days is about 80, and the average relative humidity for the city is 57 percent.

Winter in Peking is long and usually begins in October, when northwesterly winds gradually gain strength. Until March, this seasonal wind system dominates Peking; the Siberian air that passes southward over the Mongolian Plateau and into China proper is cold and dry, bringing little snow or other precipitation. The monthly mean temperature from December to February is below freezing. Spring, the windiest season, is short and rapidly becomes warm. Mainly because of the prevailing high winds, the evaporation rate in spring averages about nine times the total spring precipitation and frequently is sufficient to cause droughts that are harmful to agriculture. In addition to being the season of torrential rains, summer is rather hot, as warm and humid air from the southeast often penetrates into North China. Autumn begins in late September and is a pleasant, though short, season when the sky is clear and the temperature comfortable.

Plant life. The mountain areas of Peking belong to the temperate deciduous forest zone, while the plains area belongs to the wooded steppe zone. Because of continuous deforestation by humans for several centuries, however, only the mountain areas in the northeast and the western parts of the metropolitan area are still covered with mixed forests, composed mainly of pine, oak, and Manchurian birch. Different types of forest, divided into vertical zones, can be seen at higher elevations. On the lower slopes of many hills to the west of the city, human activities have resulted in the disappearance of forest cover so that only bushes and shrubs dot the landscape. On sunny slopes at elevations between about 2,300 and 5,600 feet above sea level, a variety of species occur, including Manchurian birch (*Betula mandshurica*), Dahurian birch (*B. davurica*), trembling poplar (*Populus tremula*), Mongolian oak (*Quercus mongolica*), and Liaotung oak (*Q. liaotungensis*). Between 5,600 and 6,250 feet, the mixed forest of truncated maple (*Acer truncatum*) and trembling poplar replaces all other species. Above 6,250 feet, goat willow (*Salix caprea*) becomes the dominant tree.

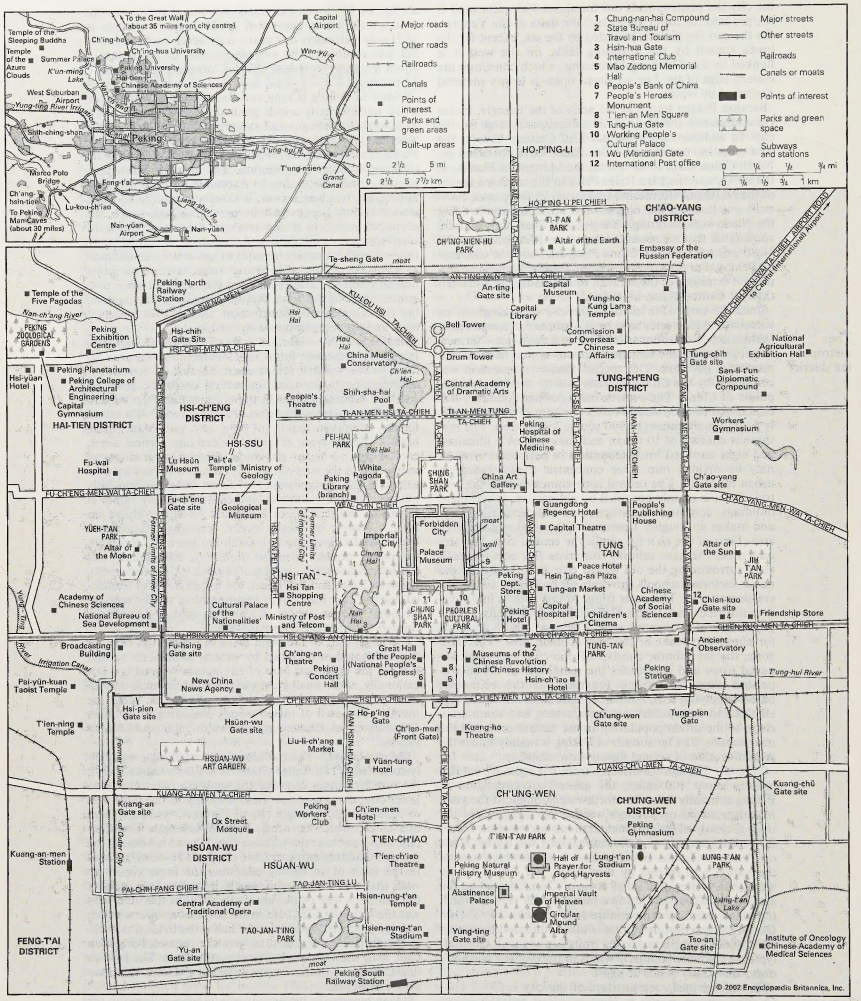
The larger part of the plain areas of Peking has been either cultivated or occupied by various settlements so that it is for the most part bare of any natural vegetation. Occasionally, some small groves of planted trees may be seen near graveyards or in the vicinity of villages; these are mainly composed of mixed woods consisting of oil pine, Chinese juniper, Chinese cypress, willow, elm, and Chinese locust.

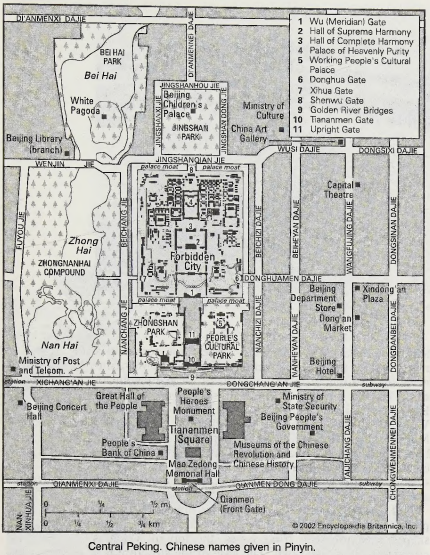
The city layout. The traditional core of Peking essentially consisted of two walled cities (the walls no longer stand), the northern inner city and the southern outer city. The inner city, also known as Tatar City, lay to the southwest of the site of the Mongol city of Ta-tu; it was in the form of a square, with walls nearly 15 miles in length. The outer city, also known as the Chinese City, was added during the reign of the Ming emperor Chia-ching (1521–66/67); it was in the form of an oblong adjoining the inner city, with walls that were 14 miles in length, including four miles of the southern wall of the inner city. Within the inner city was the Imperial City, also in the form of a square, which had red plastered walls six and a half miles in length. Within the Imperial City, in turn, was the moated Forbidden City, with walls two and a quarter miles long. The Forbidden City contains the former Imperial Palaces, which are now a museum.

Peking represents, better than any other existing city, the heritage of Chinese architectural achievement. During each dynasty in which Peking was the capital, care was consistently taken to preserve tradition when the city was rebuilt or remodeled. Few cities in the world can thus rival Peking in the regularity and harmony of its city plan.

The north-south axis

The plan was composed about a single straight line, drawn north and south through the centre of Peking, on which the internal coherence of the city hinged. All the city walls, important city gates, main avenues and streets, reli-


Major roads
Other roads
Railroads
Canals
Points of interest
Parks and green areas
Built-up areas
Major streets
Other streets
Railroads
Canals or moats
Points of interest
Parks and green space
Subways and stations
1 Chung-nan-hai Compound
2 State Bureau of Travel and Tourism
3 Hsin-hua Gate
4 International Club
5 Mao Zedong Memorial Hall
6 People's Bank of China
7 People's Heroes Monument
8 T'ien-an Men Square
9 Tung-hua Gate
10 Working People's Cultural Palace
11 Wu (Meridian) Gate
12 International Post office
Peking
HO-P'ING-LI
CH'AO-YANG DISTRICT
HSI-CH'ENG DISTRICT
TUNG-CH'ENG DISTRICT
HAI-TIEN DISTRICT
HSI-SSU
HSUAN-WU DISTRICT
CH'UNG-WEN DISTRICT
T'IEN-CH'IAO
FENG-T'AI DISTRICT
Forbidden City
Imperial City
Peking Workers' Club
Peking Gymnasium
Peking Natural History Museum
Hall of Prayer for Grain Harvests
Imperial Vault of Heaven
Circular Mound Altar
Temple of Heaven
Lung-t'an Stadium
Lung-t'an Lake
Tso-an Gate site
Yung-ting Gate site
Peking South Railway Station
Peking North Railway Station
Peking Zoological Gardens
Peking Exhibition Centre
Peking Planetarium
Peking College of Architectural Engineering
Capital Gymnasium
Hsi-yuan Hotel
Fu-wai Hospital
Liu Hsün Museum
Pai-t'a Temple
Ministry of Geology
Fu-ch'eng Gate site
Geological Museum
Yüeh-t'an Park
Altar of the Moon
Academy of Chinese Sciences
National Bureau of Sea Development
Cultural Palace of the Nationalities
Ministry of Post and Telcom
Broadcasting Building
Fu-hsing Gate site
Pai-yün-kuan Taoist Temple
T'ien-ning Temple
Hsi-pien Gate site
Hsüan-wu Gate site
Kuang-an Gate site
Ox Street Mosque
Kuang-an-men Station
Ch'ien-men Hotel
Ch'ien-men (Front Gate)
Ho-p'ing Gate
Liu-li-ch'ang Market
Yüan-tung Hotel
Kuang-ho Theatre
Kuang-chü Gate site
Ch'ung-wen Gate site
Tung-pien Gate
Ch'ien-chiao Theatre
Hsien-nung-t'an Temple
Hsien-nung-t'an Stadium
Abstinence Palace
Yung-ting Gate site
Institute of Oncology Chinese Academy of Medical Sciences
Yu-an Gate site
Nan-yüan Airport
Capital Airport
Lu-kou-ch'iao
Ch'ang-hsin-tien
Kuo-t'ai
Marco Polo Bridge
Grand Canal
Wen-yü R.
Hsiang-shan
Pei-ching University
Tsing-hua University
Chinese Academy of Sciences
West Suburban Airport
Yung-ting River Irrigation
Ch'ing-ho
Temple of the Sleeping Buddha
Temple of the Azure Clouds
Summer Palace
K'un-ming Lake
To the Great Wall (about 35 miles from city center)
To Peking Man Caves (about 30 miles)
Ch'ang-an Theatre
Peking Concert Hall
New China News Agency
Great Hall of the People (National People's Congress)
Museums of the Chinese Revolution and Chinese History
Hsin-ch'iao Hotel
Ancient Observatory
Tung-an Market
Hsin Tung-an Plaza
Peking Dept. Store
Capital Hospital
Children's Cinema
Chinese Academy of Social Science
Friendship Store
Chien-kuo Gate site
Peace Hotel
Capital Theatre
Guangdong Regency Hotel
People's Publishing House
Altar of the Sun
Jih T'an Park
Ch'ao-yang Gate site
Workers' Gymnasium
San-li-t'un Diplomatic Compound
National Agricultural Exhibition Hall
Tung-chih Gate site
Commission of Overseas Chinese Affairs
Yung-ho Kung Lama Temple
Capital Museum
Capital Library
An-ting Gate site
Bell Tower
Drum Tower
China Music Conservatory
Central Academy of Dramatic Arts
People's Theatre
Shih-sha-hai Pool
Pei-hai Park
White Pagoda
Peking Library (Branch)
China Art Gallery
Peking Hospital of Chinese Medicine
Ch'ung-shan Park
Ch'ien-men Hotel
Ch'ien-men Workers' Club
Ox Street Mosque
Hsüan-wu Art Garden
Central Academy of Traditional Opera
T'ao-jan-t'ing Park
Yu-an Gate site
Altar of Earth
Embassy of the Russian Federation
Ch'ung-nien-hu Park
Te-sheng Gate
Temple of the Five Pagodas
Nan-ch'ang River
Hsi-chih Gate Site
Hsi-chih-men Ta-chieh
Hsi-tan Shopping Centre
Ch'ien-men Ta-chieh
Ch'ung-wen Hotel
Tung-tan Park
Tung-an Market
Tung-an Men Ta-chieh
Ch'ung-wen Men Ta-chieh
Kuang-an-men Ta-chieh
Kuang-ch'i-men Ta-chieh
Hsi-chih-men Ta-chieh
Hsi-ch'ang-an Chieh
Fu-ch'eng-men Wai Ta-chieh
Hsi-chih-men Nei Ta-chieh
Ti-an-men Hsi Ta-chieh
Ti-an-men Tung Ta-chieh
Ch'ao-yang-men Wai Ta-chieh
Ch'ao-yang-men Nei Ta-chieh
Tung-ch'ang-an Chieh
Tung-ch'ih-men Nei Ta-chieh
Tung-ch'ih-men Wai Ta-chieh
Hsi-ch'ih-men Nei Ta-chieh
Hsi-ch'ih-men Wai Ta-chieh
Hsi-ch'ang-an Chieh
Hsi-chih-men Ta-chieh
Hsi-tan Ta-chieh
Hsüan-wu Ta-chieh
Ch'ien-men Tung Ta-chieh
Ch'ung-wen Men Ta-chieh
Ch'ung-wen Men Wai Ta-chieh
Ch'ien-men Hsi Ta-chieh
Ch'ien-men Tung Ta-chieh
Ch'ien-men Wai Ta-chieh
Ch'ien-men Nei



Central Peking. Chinese names given in Pinyin.

Park) on the west. Farther away from the palaces, the market area of Tung Tan to the east was balanced by the Hsi Tan market to the west; these still form two of Peking's main business districts. The T'ien-t'an (Temple of Heaven) Park to the south of the inner city is counterbalanced by the Ti-t'an (Altar of the Earth) Park to the north of the city. Of the 16 city gates constructed in Ming times, seven were located on each side of the north-south line, and two were situated on the line itself. Only a few of the old gates still stand, but their names are carried by the city streets.

The main avenues of the old city, whether running north-south or east-west, connected the gates on the opposite walls and divided the whole city into a rectangular grid. Within the walls, buildings were constructed around a courtyard or series of courtyards, with every important building facing south. Buildings often stood behind one another along a north-south line, with small courtyards in between. This prevailing southern orientation of buildings has a climatic functional basis, but it also appears to have been sanctified or conventionalized early in the Bronze Age in connection with ancestral ceremonies and with the worship of heaven and earth.

Housing. To cope with the rapid growth of population, a number of housing projects have been constructed for office and factory workers. During the first five-year plan (1953–57), housing projects were concentrated in the western outskirts of the city, where apartment buildings were erected near government offices outside the Fu-hsing Gate site. After 1958 the construction of a large group of multi-unit housing estates was begun in the northern outlying districts between the An-ting Gate site and Te-sheng Gate, centring on the residential neighbourhood of Ho-p'ing-li. The Ho-p'ing-li housing development contains primary and secondary schools, nurseries, hotels, and recreational facilities, as well as scores of four- or five-story apartment buildings. In addition, there are many groups of single-family houses in the northern suburbs, with associated parks, theatres, and recreational centres. All these buildings are supplied with gas and water, in contrast with the older parts of the city, where the existence of old stone paving hindered the development of modern structures and facilities.

In the area outside the Chien-kuo Gate site, to the east of the city, apartment buildings accommodate the families of office workers employed in nearby government office buildings. This area has also been selected as the diplomatic district and contains many foreign embassies, together with a number of Western-style houses for diplomatic representatives and their families.

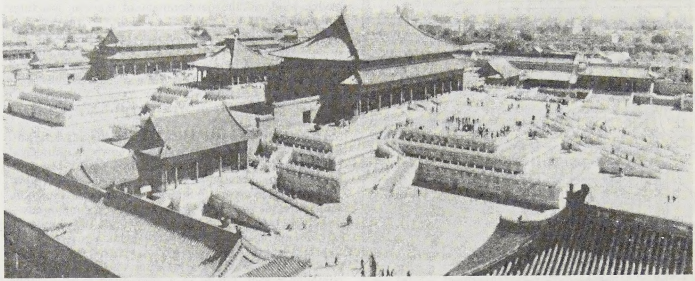
In the older districts inside the former city walls, many dilapidated houses have been pulled down in municipal housing-renewal projects and have been replaced by multistory apartment buildings. Urban-renewal projects, however, have been unable to match the growth of population. As a consequence, many traditional house compounds—originally designed centuries ago to house the families of officials during the Ch'ing, or Manchu, dynasty (1644–1911/12)—have been repaired or renovated and subdivided to provide quarters for three or four families per compound; each family in a compound faces a public courtyard and shares a common front gate with other families.

In order to minimize commuting traffic in the metropolis, many factories in the eastern and southern outskirts of the city have erected apartment buildings to house workers. These workers' residences constitute independent communities and are located so that they are easily accessible from the place of work yet are far enough away to be free from noise and smoke. Satellite towns have also been developed in the rural counties in an effort to disperse population and industries from the central city. Industries have been encouraged in rural areas in order to absorb surplus labourers and to supplement farmers' incomes.

Workers' community housing

Architecture. Peking's heritage of Chinese architectural achievement is exemplified by both private housing and public buildings. As the whole city was laid out in a rectangular street pattern symmetrically arranged around the palace compound, almost every dwelling in the city is rectangular in form, with the four sides squarely facing the cardinal directions. Most houses in the inner city were designed as residences of former officials and their families, and almost every dwelling compound is surrounded by high walls, with an open courtyard in the centre flanked by houses on the eastern, western, and northern sides, usually one story high. The former residences of high-ranking officials were composed of two or three compounds, interconnected along a north–south axis.

Stepping over the high wooden gate-sill of the front gate of a large compound, the visitor will find a brick screen wall located a few feet inside the gate—a structure that was supposed to shut off the intrusion of evil spirits, as well as to prevent curious passersby from looking inside. Beyond the screen is the outer courtyard, or service courtyard, which is flanked by houses to the east and west. There, in former days, were the kitchen and living quarters for the gatekeeper, servants, and any visiting guests and relatives. A red-painted gate leads through the north wall of the outer court into the main part of the house. This is built around three sides of the main courtyard, which, usually shaded by a large tree, was the centre of the family's life. All the windows look inward to the courtyard, and a double door opens into the courtyard from each of the three wings. The windows extend from three feet above the ground up to the deep, overhanging eaves. As they face south, the rooms in the main building get the maximum possible sunshine in winter, and the eaves provide a pleasant shade in summer, when the sun is high. The wing at the northern end of the court was intended for the head of the family and his wife. It is divided into three compartments, the central one being the living or community room, and the smaller rooms at either side being the bedroom and study. The rooms facing east and west—three on each side of the court—were for married sons and their families. This is the basic plan of all the old houses in Peking. Larger families built an extra courtyard behind the main house, because the traditional ideal was that all the existing generations should live together. Since 1949, however, as mentioned above, a great many of the old-style



The Hall of Supreme Harmony (centre), one of the former Imperial Palaces, now part of the Palace Museum complex in the Forbidden City.
Xinhua News Agency

houses have been adapted for use by several families.

While the style and architecture of private dwelling units are uniform throughout the city, the public buildings and temples are characterized by a variety of designs and structures. As the nation's political and cultural centre for more than 700 years, Peking has more buildings of historical and architectural significance than any other contemporary city in China. Since 1949 many new government and municipal buildings, combining both tradiitional and Western architecture, have added to the city's landscape.

The Imperial Palaces

The Imperial Palaces (now the Palace Museum), with their golden roofs, white marble balustrades, and red pillars, stand in the heart of Peking and are surrounded by a moat and walls with a tower on each of the four corners. The palaces consist of outer throne halls and an inner court. North of the three tunnel gates that form the front gate, called the Wu (Meridian) Gate, a great courtyard lies beyond five marble bridges. Farther north is the massive, double-tiered Hall of Supreme Harmony (T'ai-ho tien), once the throne hall. A marble terrace rises above the marble balustrades that surround it, upon which stand beautiful, ancient bronzes in the shapes of caldrons, cranes, turtles, compasses, and ancient measuring instruments. The Hall of Supreme Harmony is the largest wooden structure in China.

Behind it, beyond another courtyard, is the Hall of Complete Harmony (Chung-ho tien), where the emperor paused to rest before going into the Hall of Supreme Harmony. Beyond the Hall of Complete Harmony is the last hall, the Hall of Preserving Harmony (Pao-ho tien), after which comes the Inner Court (Nei-t'ing). The Inner Court was used as the emperor's personal apartment. It contains three large halls, the Palace of Heavenly Purity (Ch'ien-ch'ing kung), the Hall of Heavenly and Earthly Intercourse (Chiao-t'ai tien), and the Palace of Earthly Tranquillity (K'un-ning kung).

The Palace of Heavenly Purity is divided into three parts. The central part was used for family feasts and family audiences, audiences for foreign envoys, and funeral services; the eastern section was used for mourning rites and the west section for state business. The other two palaces, one behind the other, were imperial family residences. The three throne halls in the Outer Court and the three main halls in the Inner Court lie along the central axis. On either side are smaller palaces, with their own courtyards and auxiliary buildings. Behind the buildings, before the north gate of the Imperial Palaces is reached, lies the Imperial Garden. Each palace, its courtyard and side halls, forms an architectural whole.

The Temple of Heaven

Among the historical and religious structures in Peking, the Temple of Heaven (T'ien-t'an), located in the old outer city, is unique both for its unusual geometric layout and because it represents the supreme achievement of traditional Chinese architecture. Entering from the western gate of the temple, along a path that is about 1,600 feet in length and shaded by ancient cypresses, one comes to a raised passage about 1,000 feet long. This broad walk connects the two sets of main buildings in the temple enclosure. To the north lies the Hall of Prayer for Good Harvests (Ch'i-nien tien) and to the south, the Imperial Vault of Heaven (Huang-ch'iung yü) and the Circular Mound Altar (Huan-ch'iu t'an). The three buildings are built along a straight line. Seen from the air, the wall of the enclosure to the south is square, while the one on the north side is semicircular. This pattern symbolizes the traditional Chinese belief that heaven is round and earth square.

The Hall of Prayer for Good Harvests, built in 1420 as a place of heaven worship for the emperors, is a lofty, cone-shaped structure with triple eaves, the top of which is crowned with a gilded ball. The base of the structure is a large, triple-tiered circular stone terrace. Each ring has balustrades of carved white marble, which gives the effect of lace when seen from a distance. The roof of the hall is deep blue, resembling the colour of the sky. The entire structure, 125 feet high and about 100 feet in diameter, is supported by 28 massive wooden pillars. The four central columns, called the "dragon-well pillars," represent the four seasons; there are also two rings of 12 columns each, the inner ring symbolizing the 12 months and the outer ring the 12 divisions of day and night, according to a traditional system. The centre of the stone-paved floor is a round marble slab that has a design of a dragon and a phoenix—traditional symbols of king and queen. The hall has no walls, only partitions of open latticework doors.

The Imperial Vault of Heaven, first erected in 1530 and rebuilt in 1752, is a smaller structure some 64 feet high and about 50 feet in diameter. The circular building has no crossbeam, and the dome is supported by complicated span work. Its decorative paintings still retain their fresh original colours.

South of the enclosure lies the Circular Mound Altar, built in 1530 and rebuilt in 1749. The triple-tiered white stone terrace is enclosed by two sets of walls that are square outside and round inside; thus, the whole structure forms an elaborate and integrated geometric pattern. The inner terrace is 16 feet above the ground and about 100 feet in diameter; the middle terrace is about 165 feet across and the lowest terrace some 230 feet across. Each terrace is encircled by nine rings of stones. Both the Imperial Vault of Heaven and the Circular Mound Altar were erected to portray the geometric structure of heaven, as conceived by the architects of the Ming dynasty. After 1949 the whole enclosure of the Temple of Heaven was repaired; it is now a public park.

To the east of T'ien-an Men Square within the People's Cultural Park is found the Working People's Cultural Palace (formerly the Temple of the Imperial Ancestors), where the tablets of the emperors were displayed. The temple, like the Imperial Palaces in style, was built in three stonework tiers, each with double eaves. On either side are



Children sweeping the steps at the Hall of Prayer for Good Harvests, central building of the Temple of Heaven.
Emil Schulthess—Black Star

two rows of verandas surrounding a vast courtyard large enough to hold 10,000 people. Exhibitions of economic and cultural achievements, both of China and of other countries, are frequently held in the three halls. Lectures by leading scholars on science, literature, and the arts are also held there.

The Great Hall of the People

Perhaps the most imposing structure constructed in the heart of the city since 1949 is the Great Hall of the People. The Great Hall is located on the southwestern side of T'ien-an Men Square and is an immense building with tall columns of gray marble set on red marble bases of floral design. It has a flat roof with a golden-yellow tile cornice over green eaves shaped like lotus petals. The base of the building is of pink granite, and its walls are apricot yellow. Its frontage is 1,100 feet long—about the equivalent of two city blocks. The floor space of some 6,060,000 square feet (563,000 square metres) is more than that of all the palaces in the old Forbidden City. Inside the building, the ceiling and walls are rounded. The focus of the lighting system is a red star surrounded by golden sunflower petals. The grand auditorium is where the National People's Congress holds its sessions. Other components are a banquet hall for 5,000, huge lobbies, and scores of meeting rooms and offices for the standing committee of the congress.

THE PEOPLE

Population growth

Since 1949 Peking has been one of the most rapidly growing urban centres in China. During the dynastic period, when Peking was the capital, from the 13th to the beginning of the 20th century, the population of Peking seems to have increased very slowly, but in the four decades (1911–49) after the downfall of the Ch'ing dynasty it increased at an accelerating rate. This rapid increase has been attributed to at least three factors: the rural disorder on the North China Plain during the rule of the warlords caused extensive migration to the walled city of Peking; the loss of Manchuria to Japan in 1931 made Peking a shelter for thousands of refugees from the northeastern provinces; and the establishment of numerous institutions of higher education, sponsored by the government or by foreign missionaries, stimulated the growth of the student population in a city that was a cultural as well as a political capital.

The phenomenal increase in Peking's population, however, occurred after the city had resumed its role as the national capital. In 1949 Peking was the fourth largest city in the country; some 10 years later it had become the second largest. This increase resulted primarily from the necessary migration of thousands of government workers to the new capital, the rapid rate of industrial development in the municipality during the 1950s and '60s, and the extension of the municipal boundary in 1959 to include large rural areas. The effect of these circumstances has been a dramatic growth in the population of Peking since 1949.

The overwhelming majority of Peking's population is Han (Chinese). Hui (Chinese Muslims), Manchus, and Mongols constitute the largest minority groups. Peking residents speak a dialect that forms the basis for Mandarin, or *p'u-t'ung hua* ("common language"), which all educated citizens can speak.

Peking has a small but growing community of foreign residents—mainly diplomats, journalists, business people, teachers, and students—who have become an important aspect of the city's life. Government authorities have made great efforts to meet the special demands of foreigners. Foreigners generally send their children to Western-style schools (although the children are free to attend Chinese schools); in addition, the government has provided an international club with dining and recreational facilities, although few foreigners use it.

Since the late 1970s efforts have been made to restore the ancient temples, churches, and seminaries of various religions (most prominently Buddhism) that were damaged during the Cultural Revolution (1966–76). Some of the restored structures are again being used by worshipers, monks, and pupils.

THE ECONOMY

Agriculture. Although Peking has been the capital of China for hundreds of years, it is far away from the nation's traditional key economic area, the productive Yangtze Valley. In the past the task of feeding the large urban population in Peking was facilitated by transporting grain from the south through the Grand Canal or by sea. Since the first decade of the 20th century, railways have played an important role in bringing food supplies to the capital. The city, however, has striven for a high degree of self-sufficiency in secondary food supplies, such as vegetables, fruits, fish, and poultry, and in a number of construction materials. The expansion of municipal boundaries in 1958 and 1959 was partially aimed at this goal, and the municipality has succeeded in maintaining an adequate supply of vegetables and fruits. Agricultural reforms since the early 1980s have given individual households greater freedom of production and resulted in greater and more varied food supplies in the city.

Thus, unlike in most of the world's major cities, agriculture forms a significant part of Peking's economy. Vegetables are grown in a belt that encircles the city and is covered by a network of irrigation channels. The channels are supplemented by a large number of electric pumps. These irrigation measures have made intensive farming possible in an area prone to frequent droughts. Vegetables are grown in the winter months in hothouses. This large farm belt is also one of the most highly mechanized agricultural areas in China. The municipal government has encouraged the expanded cultivation of fruit and nut trees, and large areas of wasteland to the north and east of the city have been reclaimed for fruit-tree planting. The large-scale planting of fruit and other trees has also reduced con-



The People's Heroes Monument, in front of the Great Hall in T'ien-an Men Square.
CLICK/Chicago

siderably the wind erosion of the soil. While the hill areas produce large quantities of pears, persimmons, apples, chestnuts, and walnuts, the reclaimed lowlands are covered with vineyards and peach orchards. Part of the Peking fruit crop is shipped by train to the cities of the Northeast and Inner Mongolia.

Industry. One of the main differences between the imperial capital of former times and present-day Peking is that the imperial capital was a consumption rather than a production centre, receiving supplies of all kinds from other parts of the country. Since 1949, however, Peking has emerged as China's most industrial city after Shanghai in terms of total production value, but concern for the adverse effects of industrialization on the city's environment has more recently led to a curtailment of industrial expansion.

Iron and steel manufacturing

Among the large industrial establishments is the Sh'ou-tu Iron and Steel Works, located about nine miles west of the old city. The Sh'ou-tu plant was originally started in 1920, based on local ore and anthracite in the Western Hills; after the Japanese occupation of 1937, it produced a meagre amount of pig iron. Since 1958 the plant has been enlarged and is now one of the largest steel plants in China. Its high-quality steel production supplies such Peking industries as machine building, electrical engineering, and precision instrument manufacturing. A number of smaller finishing mills have also been established to produce such items as cold-drawn bearing steel and flat spring steel for tractor and automobile accessories, seamless tubes for high-pressure boilers, and magnetic steel for machine tools and radios. Peking is also an important centre of machinery manufacture. Most of these factories were built in the suburbs east and south of the central city, where extensive tracts of level land were available and where the prevailing northwest winds would carry industrial pollutants away from the densely populated areas. With the rapid expansion of cotton cultivation in Hopeh Province in the 1950s and '60s, Peking also became one of China's major textile centres. Peking's petrochemical industry expanded rapidly when an oil pipeline was constructed in the mid-1970s to link the city with the Ta-ch'ing oil field in the Northeast. The petrochemical industry is dominated by the Yen-shan Petrochemical Corporation, located southwest of the city.

As a former imperial capital, Peking has a rich accumulation of art treasures, produced by a variety of arts and handicraft industries that were intimately connected with court life and imperial needs. Much of the traditional handicraft industry has been reorganized and reequipped. Among the city's well-known handicraft products are rugs and carpets, porcelain and chinaware, jade and ivory sculpture, brass ware, enamel and lacquer ware, lace, and embroidery.

Commerce and finance. With the rapid growth of industry and population in the city, many services, operated by both private companies and government agencies, have also greatly expanded.

Banking institutions

China has a nationwide and centralized system of banking, in which the state-owned People's Bank of China plays a key role. This bank, with its head office in the inner city, functions as the agent of the national treasury. All funds of state-owned industrial enterprises, as well as of national, provincial, and local governments, the People's Liberation Army, and the network of cooperatives, are deposited with the People's Bank. The bank uses the working capital at its disposal for the operations of the economy and the government. The People's Bank thus acts as cashier of the national budget and of a large part of the nation's financial operations. Through its branch offices and savings account centres, the bank also serves the daily needs of people in the city and in the suburbs. Working in cooperation with the People's Bank and under its supervision are four other banking institutions: the People's Construction Bank, which has the task of financing and supervising all basic construction projects, such as new railways; the Agricultural Bank of China, which specializes in agricultural investment; the Bank of Commerce and Industry, which provides loans to industry; and the Bank of China, which handles most international trade and foreign exchange. In

Zhang Shuyuan—Xinhua News Agency



Agricultural produce market in northern Peking.

addition, several foreign-owned banks have representative offices in the city to assist foreign nationals doing business in China.

Because of the symmetrical layout of the city, Peking lacks the compact central business district that characterizes most cities in the Western world. Historically, the market areas of the city were situated at two street intersections to the southwest and southeast of the Imperial Palaces, and the Tung Tan and Hsi Tan neighbourhoods are still major shopping centres. Since 1990, however, Western-style shopping malls and department stores have been established in various parts of the city. One of the most vibrant retail areas is along Wang-fu-ching Ta-chieh (Wangfujing Dajie). As part of a 20-year development plan for this shopping street that began in 1991, it was transformed in 1999 when storefronts were beautified and all bicycles and motor vehicles, except city buses, were excluded. The Peking Department Store, a state-owned enterprise, still operates there, but it has been overtaken by gigantic, privately-owned shopping malls such as Hsin Tung-an (Xindong'an) Plaza. In 1996 construction began on an enormous shopping complex at the southern end of Wang-fu-ching Ta-chieh. Similar shopping districts can be found in other parts of the city, such as Chien-kuo-men-wai and San-li-t'un, both of which are near diplomatic compounds. There is still a Friendship Store in Chien-kuo-men-wai. In the past, when it was the only place to buy Western goods, it mainly served foreign visitors, although some Chinese—usually cadres or those who received foreign-currency remittances from relatives living abroad—were allowed to shop there. Although anyone can shop there now, it has been passed up by more modern establishments.

Shopping facilities

In addition to these modern establishments, traditional markets that still serve a local function are spread around the city. The Peking markets have a long history, and each has developed its own reputation for special commodities and services. The restored Liu-li-ch'ang Market is located just south of the Ho-p'ing Gate in the old outer city. The area acquired its name (which means "the glazier's shop") from the colourful glazed tiles that were made there during the Ming dynasty, but in the latter part of the 18th century it gradually became a market for curios, antiques, old books, paintings, works of ancient Chinese calligraphers, and paper. It is still a centre for traditional art shops. Ta-chan-lan, just west of Ch'ien-men Ta-chieh (Qianmen Dajie), was rebuilt in 1998 and many of the Ch'ing period shops there were restored. This market contains shops specializing in silk, tea, herbal medicines, food, and clothing. P'an-chia-yüan-chiu-hua Shih-ch'ang—which used to be popular with China's national minorities, though they are now outnumbered by Han Chinese—sells a wide variety of ornaments, many of them made of metal, and numerous other objects. Ya-pao Lu, just south of Ch'ao-yang-men-wai Ta-chieh, is popular with Russians and eastern Europeans. Most of the Chinese shop owners speak at least some Russian, signs are written in Cyrillic, Russian food is served, and most of the products are Russian-made clothing and daily-use items. Ya-t'ing Hua Niao Shih-ch'ang, just outside the southeast corner of T'ien-t'an Park, is a flower and bird market that also sells kittens and Pekinese dogs, kites, and many other items.

Tourism

Tourism has become increasingly important to the Peking economy. Tourists are drawn not only to the historical and cultural attractions of the central city but also to such sites in the rural areas as the Great Wall and the tombs of the Ming emperors to the north of the city and the Peking-man caves to the southwest. There has been a steady increase in tourist services provided in Peking. Many hotels and hostels have been built, and old ones have been renovated and enlarged to meet the ever-greater demand.

Transportation. Peking is the railroad centre of China, forming the terminus of a number of lines in the national railroad network. The major lines radiating from Peking provide connections with Shen-yang (Liaoning Province), Shanghai Municipality, Canton (Kwangtung Province), Hong Kong, Pao-t'ou (Inner Mongolia Autonomous Region), and T'ai-yüan (Shansi Province). The local lines serve outlying districts within the metropolitan area.

Rail connections

Peking is linked by direct express with several other large urban centres, and it is also connected by express trains with Moscow; P'yŏngyang, North Korea; and Ulaanbaatar, Mongolia.

In addition to the railroads that connect Peking with distant cities, an ambitious national highway building program is linking Peking with many other cities throughout the country. Radiating from the city are more than a dozen major highways, which are used by the growing number of privately owned automobiles as well as trucks and long-distance bus services. Peking's road-transport system, though being improved rapidly, is still inadequate and cannot keep up with the rapid increase in motor vehicles; as a result, traffic congestion is often severe. The once-ubiquitous bicycles and three-wheeled cycle carts are still heavily used for short-distance transport. In the rural areas of the municipality, ox- or horse-drawn carts are still sometimes used to transport goods.

Peking is also a centre of China's civil air transport. The nation's major domestic air routes link Peking to regional centres at Shanghai, Canton, and Urumchi (Wu-lu-mu-ch'i, in the Uighur Autonomous Region of Sinkiang) as well as to other major cities, and international air carriers provide service to major overseas destinations. In 1999 a new terminal building was completed at Peking's Capital Airport, located northeast of the central city.

Peking's intracity commuting services are provided primarily by a network of local and express buses. Taxis of various types are also an important means of getting around. In 1969 a subway line—the first in China—was opened, running east-west through the centre of the city; this line was supplemented by a separate loop line opened in 1987 that follows the old limits of the inner city beneath the Second Ring Road. A third line which extends eastward from the original east-west line was completed in 1999 and connected to the existing line in 2000. Also in 2000 construction began on the first north-south line and on an elevated light rail line that will extend the east-west line to the eastern suburbs. The plan is to create a complete transit network consisting of at least eight lines.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Peking is one of the four centrally administered (*i.e.*, province-level) *shih* (municipalities) in China (the others being Shanghai, Tientsin, and Chungking), and there is no governmental tier between it and the central government. The municipality is divided administratively into four urban and six suburban *ch'u* (districts) and the eight *hsien* (counties) in the peripheral areas. Peking's municipal government is part of the hierarchical structure of the Chinese government that extends from the national organization, through the provincial apparatus, to the municipal and, ultimately, neighbourhood levels. Paralleling this governmental structure is that of the Chinese Communist Party (CCP). As in all of China, real power in Peking is held by the local CCP, but local government institutions perform various formal functions. The Peking Municipal People's Congress, the formal governmental decision-making body, follows the guidance of the local CCP in issuing administrative orders, collecting taxes, determining the budget, and implementing economic plans. Under the direction of the local CCP, a standing committee of the Municipal People's Congress recommends policy decisions and oversees the operation of municipal government. Executive authority is formally assigned to the Peking People's Government, the officers of which are elected by the congress. The local government consists of a mayor, vice mayors, and numerous bureaus in charge of public security, the judicial system, and other civil, economic, social, and cultural affairs.

As a subdivision of the municipality, each district or township has its own mayor. Below the urban district, police substations supervise the population and subdistrict or neighbourhood offices handle civil affairs in their areas. Residents' committees help mediate disputes, conduct literacy campaigns, and promote sanitation, welfare, and family planning.

As the national capital, Peking houses all of the most important governmental and political institutions in the na-

The State Council

tion. These include the National People's Congress, nominally the supreme organ of state power; the State Council, the highest executive organ of the state; and the various administrative departments under the jurisdiction of the State Council, including the ministries and commissions in charge of foreign affairs, internal affairs, public security, national defense, justice, finance, culture, health, education, nationality affairs (concerning minority groups in the country), agriculture, and various branches in industry. Peking is also the headquarters of the parallel CCP organizations—the National Party Congress, the Central Committee and Politburo, and the Secretariat. In addition to the above, the highest organ of state concerned with the maintenance of law and order—the Supreme People's Court—is located in Peking.

Water supplies

Public utilities. Peking is not on a river, and most water for municipal consumption has to be brought from elsewhere. Some water, however, comes from shallow wells, which are common throughout much of the region. These provide some villages with drinking water and supply water for irrigation.

From the earliest history of the city, use has been made of local rivers and springs. To the east is the Pai River, which, after being joined by the Ch'ao River at the Mi-yün Reservoir, forms the Ch'ao-pai River, which then flows southward. The tumultuous Yung-ting is a large river, draining the Shansi Plateau and the northwest of Hopeh. After following a twisting course through the mountains, it reaches the Peking plain, passes under the Marco Polo Bridge, nine miles southwest of Peking, and then turns southward to meet the Grand Canal north of Tientsin. The flow of the Yung-ting is irregular; in the rainy season it rises rapidly, carrying with it large quantities of silt, which raise the level of the riverbed considerably. At Marco Polo Bridge it is 50 feet above the level of the city, thus constituting a hazard when the river is in flood but facilitating canalization and irrigation. Several springs rise at the foot of the Hsiang Hills and on Yü-ch'üan Hill, both to the northeast of the city. During the Ch'ing dynasty these springs were tapped by means of an aqueduct that conveyed water for the city moat and for the three lakes near the Imperial Palaces. The moat around the city walls became an important means of water distribution for the municipality.

In the early 1950s, large-scale water-conservation projects were begun to provide more water for the expanding urban area. Two large artificial lakes have since been made: the Mi-yün Reservoir, northeast of the city, and the Kuan-t'ing Reservoir, in the northwestern mountains beyond the Great Wall. These regulate the flow of the Pai and Yung-ting rivers upstream, storing water at times of heavy discharge and releasing it when the rivers are low. Two lesser projects have also been carried out: the construction of the Ming Tombs Reservoir, whose waters feed a hydroelectric power station and irrigate the neighbouring countryside, and a hydroelectric power station near Mo-shih-k'ou, which uses the waters of the Yung-ting and feeds them back into the river through an ancient canal. The hydroelectric station at Kuan-t'ing is the largest source of electricity for the metropolitan area of Peking.

The Peking Water Company, which was originally formed in 1908, is responsible for supplying water for the municipality. Water is obtained from the waterworks within the old city, as well as from several waterworks in the suburbs; the network of mains and pipes extends to every part of the municipality. The Kuan-t'ing Reservoir, the largest man-made lake in the North China Plain, supplies the metropolitan area of Peking with most of its water.

Peking is rapidly developing a modern sewage system. An underground sewage system was installed as early as the 15th century, which extended nearly 195 miles. Later this network became clogged with effluent and silt. A new system was installed in the early 1950s, which, within the walled area, was partly based on the rehabilitation of the old system. All open sewers, characteristic of many Chinese cities in the past, have been eliminated, and new pipes have been laid throughout the densely populated areas of the municipality.

The headquarters of the Peking General Post Office is located on the east side of T'ien-an Men Square. It provides more comprehensive services than do post offices in Western cities, handling mail, telegrams, long-distance telephone calls, and the distribution of newspapers and magazines. There are also more than 350 branch offices and stations in the city and the suburbs. The Peking General Post Office also maintains several service centres at certain busy traffic points, such as the Ch'ien Gate and Peking Railway Station, and Wang-fu-ching Ta-chieh. The service centre on Wang-fu-ching Ta-chieh takes subscriptions to foreign newspapers and magazines and sells single issues.

Health. Chinese medical services employ both the traditional Chinese and the Western systems of medicine. Peking has some of the best hospitals of both systems in the nation.

The Peking Union Medical College Hospital, the earliest modern hospital to be established in China, was founded in 1921 as an affiliate of the Peking Union Medical College, which from 1915 onward was supported by the China Medical Board of the Rockefeller Foundation. The hospital is a large polyclinic combined with an institute of gynecology and pediatrics. It is the largest hospital in Peking. Since 1949 many new hospitals, clinics, and sanitariums have been built. People's Friendship Hospital, a gift of the Soviet Union at the peak of Sino-Soviet friendship in the 1950s, is located in the T'ien-ch'iao neighbourhood in the outer city. Until 1960 the Soviet staff not only treated patients but also gave advanced training to Chinese medical personnel on the latest Soviet methods.

Many hospitals are clinical teaching hospitals attached to medical schools; these include the Fu Wai Hospital of the Chinese Academy of Medical Science, located outside the Fu-ch'eng Gate site; the First Affiliated Hospital of the Peking Medical College, near the Hsi-an Gate site; and the Affiliated Hospital of the Research Institute of Chinese Medicine, located in the western suburban town of Hsiao-yüan. The largest pediatrics facility in China, the Peking Children's Hospital, is situated just outside the Fu-hsing Gate site and provides a homelike atmosphere for its young patients. In the hills to the west, a sanitarium in the traditional Chinese architectural style has been built for students from Asian countries. The most specialized hospitals are located in the old residential districts of the city and include the Stomatological Hospital (dealing with diseases of the mouth), the Chest Surgery Hospital, the Hospital of Plastic Surgery, an older Peking Children's Hospital, the Hospital of Chinese Medicine, and the Ji-suitan Hospital, which specializes in traumatic (accident-related) and orthopedic surgery.

Traditional medicine

Chinese traditional medicine comprises prescribing Chinese herb drugs and employing acupuncture (a medical practice encompassing therapeutics and anesthesia that involves inserting needles into various parts of the body). Doctors in both Chinese and Western systems are trained in many universities and institutes in Peking. Students in the Western-style medical schools are taught the essentials of Chinese traditional medicine, and Chinese traditional medical students are expected to be familiar with the essentials of Western medicine. Diagnoses of difficult cases are sometimes made only after consultations with specialists of both systems.

For the rural areas of the metropolis, each county has a well-equipped hospital located in the county seat. But the basic unit of the medical service is the health centre in the work unit, which is equipped with complete clinical facilities. Medical teams of doctors, nurses, and public health personnel are sent frequently from the teaching centres in the city to the work units to assist in the general medical work and to keep the local medical staff in touch with the latest medical progress and techniques.

The schools

Education. The educational system of Peking was massively disrupted by the Cultural Revolution, and throughout the 1980s it was in the process of being restored. Nevertheless, there has been no radical change in the structure of the school system in Peking since 1949; it is still largely composed of six years of primary education, which is universal, and six years of secondary education. For preschool-age children, kindergartens and nurseries are operated by factories, business enterprises, government offices,

and city block cooperatives. Their main function is to permit mothers to work, and although some of these facilities are located in the residential neighbourhoods, many are located on the premises of places of employment. In the suburban or rural areas, kindergartens are often temporarily set up during such events as the harvest.

In the early 1950s, efforts were made to abolish private schools and to put all education in the hands of the state. As a result, all private schools run by missionaries were taken over by the government. It was soon realized, however, that the government could not possibly provide enough schools for the millions who required education. The government then reversed its policy to some degree and decided to encourage the establishment of schools by local organizations, such as factories, business concerns, collectives, and communes. A major part of the finance for elementary education is now met by sources other than the central government, though the government does provide considerable subsidies to schools in difficult financial situations.

The curriculum in the primary school is basic. The early part of primary education is largely devoted to reading, writing, and mathematics, and, in the last two years of primary school, nature study, history, geography, and physical education are added to the curriculum. Closely related to primary education are movements for eliminating illiteracy and simplifying Chinese characters.

The secondary schools in Peking are mainly of three types: the general middle school, the normal school, and the vocational and technical school. The general middle school is of the academic type, with a curriculum designed to prepare students for college. The normal school trains teachers for primary schools. Vocational and technical schools, created to provide skilled workers in various fields, have developed rapidly in the city.

Peking's position as a centre of higher education in China has been further strengthened since 1949, and within the metropolitan area there are numerous institutions of higher learning. A scientific and educational district has emerged in the northwestern outskirts of the city, against the background of K'un-ming Lake, the Summer Palace, and the Western Hills. This area contains such prestigious institutions as Peking University and Ch'ing-hua University. Peking University was founded in 1898 and is one of the largest comprehensive universities in China. In 1953 the university moved from its old site at Sha-t'an, in the inner city, to the present campus, which previously belonged to the missionary-established Yen-ch'ing University. The new campus has been expanded considerably to the east and south. Ch'ing-hua University is the most renowned institution for science and engineering in China.

Institutions of higher education

Within the academic district are also found the People's University of China, an institute of ideological training, and the Central Institute of Nationalities, which accepts students from the various autonomous regions and districts. Situated in the same area are the Peking Normal College, the Peking Medical College, and the Central Conservatory of Music. Numerous institutions of higher education are also located in this area; these include the Peking Institute of Steel and Iron, the Peking Institute of Aeronautics and Astronautics, the Peking Petroleum Institute, the Peking Institute of Geological Prospecting, the Peking Institute of Forestry, and the Peking Mechanical Institute of Agriculture. The Chinese Academy of Sciences, the most prestigious research institute in the nation, is also located in this district.

Foreign-language institutions

Besides technical colleges, there are a number of foreign-language institutions in Peking where foreigners are employed to teach alongside their Chinese colleagues. The Peking Foreign Language Institute, located on the northwestern outskirts of the city, is the largest institute of its kind in China. The Second Foreign Language Institute was set up in 1964 for training personnel for the New China News Agency. There are also a number of small institutes of language training run by the government and by the Peking Radio Station. The School of Foreign Languages is an institute at the secondary level, which accepts children as young as eight years of age; half of them learn English, while others study French, German, Spanish, Arabic, Russian, and Japanese.

CULTURAL LIFE

Peking has been the magnificent centre of traditional Chinese culture and learning since the Ming dynasty. Emperors and courtiers patronized the arts, especially painting and calligraphy. Precious objects from other parts of the empire and from foreign countries poured into the capital. This role of cultural centre was continued during the Ch'ing dynasty, although the century of political and social upheaval that began in the mid-19th century led to an overall cultural decline in Peking. This decline continued into the modern period and affected not only Peking but the whole of China. With their defeat in 1949, the Nationalists shipped a huge quantity of art treasures to Taiwan. On the mainland, subsequently, many family heirlooms were purchased by the state for very low prices and were then sold for export or used to enhance museum holdings.

Until the early 1960s traditional arts, crafts, and scholarship were encouraged by the Chinese Communist Party, but the Cultural Revolution swept all before it. Art objects that were not deliberately smashed by the Red Guards were confiscated (some were returned to their former owners during the 1980s). The Cultural Revolution also put a virtual end to traditional Chinese scholarship, as many academics were sent to the countryside or imprisoned. Since the late 1970s, however, the government has made a concerted effort to restore damaged treasures and to revive the work of traditional artists and scholars. Since much of this activity has taken place in Peking, the capital has undergone something of a cultural renaissance.

Museums and libraries. The Palace Museum, containing the city's greatest collection of art treasures, is housed in the main buildings of the former Imperial Palaces. Many of the halls are kept as they were in dynastic times, each constituting a museum in itself; others are kept for the display of some of the priceless treasures from China's past. Of special interest are the porcelains and enamels, works in embroidery and precious metals, and the stone carvings and scrolls.

The Palace Museum

The Museum of Chinese History is located on the eastern side of T'ien-an Men Square. There are thousands of historical relics and documents on display, arranged to give a comprehensive survey of the progress of human achievement in China, beginning with the appearance of the prehistoric Peking man some 500,000 years ago, then continuing through the last 6,000 years of Chinese history. The Museum of the Chinese Revolution occupies a wing of the museum building. Its purpose is to trace the history of China since the mid-19th century, laying particular stress on the revolutionary movement and the part played by the Communist Party.

The Peking Planetarium, the first of its kind in China, is located outside the Hsi-chih Gate site. It is an educational centre for the dissemination of general knowledge in astronomy and natural science by means of exhibitions, forums, films, projector demonstrations, and telescopic observations of the heavens. The planetarium proper is situated on the south side of the building. Behind the main building stand an observatory and meteorological station.

The Peking Library, which holds the historical collections of the National Library of China, is located in the northwest of the city, near the university district. The library inherited books and archives from the Imperial Wen-yüan Ko library, a famous collection that has existed for more than 500 years and that, in turn, included books and manuscripts from the Imperial Library of the Southern Sung dynasty, established some 700 years ago. Also in its holdings are other collections from imperial libraries of the Ch'ing dynasty, imperial colleges, and private owners. Among them are rare copies of ancient manuscripts and books of five dynastic periods from the Sung to the Ch'ing, including a vast number of manuscript volumes on different subjects, copies of Buddhist sutras dating to the 6th century, old maps, diagrams, and rubbings from ancient inscriptions on metal and stone; in addition, it possesses the *Yung-lo ta-tien* ("The Great Canon of the Yung-lo Era") of the Ming dynasty and a copy of the *Ssu-k'u ch'üan-shu* ("Complete Library of the Four Branches of Literature"), dating from the Ch'ing dynasty. In the late 1980s the collection of modern holdings of the National

The Peking Library

Library was moved from the Peking Library to a new facility in the western outskirts of the city.

The arts. Traditional Peking opera (*ching-hsi*)—with its elaborate and stylized costumes and makeup, cacophonous music, and spectacular dance and acrobatic routines—has been revived, after an attempt during the Cultural Revolution to adapt the form to modern revolutionary themes. The opera has great appeal to older people but less for the young, who instead prefer motion pictures, television, and popular music. Musical performances of all kinds are given in Peking; the city boasts a symphony orchestra and Western-style opera and ballet companies and hosts visits by foreign orchestras and performers.

A variety of Chinese plays are staged each year, and works by Shakespeare and other Western dramatists have again been playing in Peking. The city's bookshops, almost denuded of cultural works and foreign literature during the Cultural Revolution, are now well stocked. In addition, calligraphy and the painting of non-revolutionary subjects have again become popular pastimes.

Recreation. As the residence of the imperial families through several dynastic periods, Peking is well known for its numerous parks and playgrounds; few cities in China have as large a proportion of land within the central city allocated for recreational uses. Among the most popular of Peking's parks are Chung-shan Park, Pei-hai (Beihai) Park, Ching-shan Park, the Summer Palace, and the Peking Zoological Gardens.

Chung-shan Park adjoins the west wall of T'ien-an Men Square; it is the most centrally located park in Peking and encloses the former Altar of Earth and Harvests, where the emperors made offerings to the gods of earth and agriculture. The altar consists of a square terrace in the centre of the park, to the north of which is the 550-year-old Hall of Worship, now the Sun Yat-sen Memorial Hall, which is used for public meetings. The Hall of Worship is the oldest wooden structure in Peking; its simple form, masterly design, and sturdy woodwork bear the characteristic marks of early Ming architecture.

The Water Pavilion, built out over a lotus pond on three sides to provide a gathering place for scholars and poets, is also located in the park. Scattered among the park's pools, goldfish enclosures, rocky hills, weeping willows, pines, cypresses, bamboos, and flowers are pavilions, kiosks, and towers, which are typical of Chinese garden landscape.

Pei-hai Park

Pei-hai Park lies to the northwest of the Forbidden City. It covers some 176 acres (71 hectares), half of which is water. It contains Pei Hai, the most northerly of the three lakes—called "seas" (*hai*)—that lie roughly north and south within the Imperial City. Pleasure grounds, lakes, and buildings have existed on this site for 800 years. As the lakes were deepened and dredged, the excavated earth was used to build hillocks and islands of great beauty. In 1651 a Manchu emperor built the White Pagoda, the most striking landmark in the park, on the top of a hill. Many of the buildings in the park are used for serious pursuits, as well as recreation. They include an Institute of Research on the History of Chinese Classical Literature and a Natural History Museum, as well as the Peking Library. Pei Hai is crowded with rowboats in summer, and it freezes over to become a natural skating rink in winter.

Ching-shan (Prospect Hill) Park, also known as Mei-shan (Coal Hill) Park, is a man-made hill, more than a mile in circumference, located north of the Forbidden City. The hill, from the top of which a spectacular panorama of Peking can be seen, has five ridges, with a pavilion on each. The hill was the scene of a historical tragedy when, in 1644, at the end of the Ming dynasty, the defeated Ming emperor hanged himself on a locust tree on its east slope. In the northern part of the park is Peking Children's Palace, with recreational, athletic, and educational facilities.

The Summer Palace

The Summer Palace, lying close to the Western Hills, is about six miles northwest of the Hsi-chih Gate site. It is the largest park on the outskirts of Peking and is noted for its artificial landscaping, which provides an inimitable blend of woods, water, hills, and architecture. The Summer Palace covers more than 800 acres, four-fifths of which consists of K'un-ming Lake, the remainder being man-made hillocks. More than 100 buildings—halls, towers, pavilions, bridges, and pagodas—lie scattered throughout the park; a marble boat, two stories high and some 80 feet long, located at the northwest corner of the lake, is one of the major attractions. Connecting the buildings and courts along the shore of the lake are a series of richly painted covered promenades.

Peking Zoological Gardens is located on the western outskirts of the city. The zoo was established toward the end of the 19th century and was named the "Garden of Ten Thousand Animals" (Wan-sheng Yuan). Its collection is actually less than half that size, but it is the largest zoo in the country, with animals from all parts of China and the world; one of the zoo's most popular attractions is its collection of giant pandas.

Peking Workers' Stadium, located in the eastern part of the city, seats 80,000 and is the largest sports facility in China. Football (soccer) games and field and track events are often held there. The Capital Gymnasium in the western outskirts has facilities for a variety of indoor games; it has enough space for 24 table tennis matches to be held simultaneously. A movable floor facilitates formation of an ice rink for figure skating and ice hockey.

Cuisine

The Chinese love of good food is world-renowned, and Peking is one of China's culinary showcases. All of the regional cuisines are represented among the city's hundreds of restaurants, although the Peking style predominates. The best-known dish to foreigners is Peking duck, which is the specialty of several establishments.

History

THE EARLY EMPIRES

With but few interruptions, Peking has been the capital of China for some 700 years, and in the number of years as the imperial capital it is exceeded only by Sian (Ch'ang-an) in Shensi Province and Lo-yang in Honan Province. In prehistoric times the area around Peking was inhabited by some of the earliest known human beings. Between 1918 and 1939 the fossil remains of Peking man (formerly *Sinanthropus pekinensis;* now known as *Homo erectus pekinensis*), who lived about 500,000 years ago, and of Upper Cave man, who lived about 50,000 years ago, were unearthed at Chou-k'ou-tien, a village about 30 miles southwest of Peking.

Chi—the capital of Yen

While long periods in Peking's early history remain blank, it is certain that, some 3,000 years ago, Neolithic communities settled on or near the site where Peking now stands. During the Chan-kuo (Warring States) period (475–221 BC) of the Chou dynasty (*c.* 1111–255 BC), one of the powerful feudal states, the kingdom of Yen, established its capital named Chi near the present city of Peking; this was the first capital city to be associated with the site. The city was destroyed by the troops of Shih huang-ti, founder of the Ch'in dynasty (221–206 BC).

During the Ch'in dynasty, the Yen capital was incorporated into one of the 36 prefectures then established throughout the country. A new town was built during the Han dynasty (206 BC–AD 220) and was known as Yen. Throughout the Han dynasty and the turbulent centuries that followed, however, the place remained a provincial town, witnessing most of the time the fateful struggle between the Han Chinese to the south and the nomadic Hsiung-nu, or Huns, to the north.

During the period of San-kuo (Three Kingdoms; AD 220–280), the city was again called Yen. The northern border of ancient China ran close to the present city of Peking, and northern nomadic tribes frequently broke in from across the border. Thus the area that was to become Peking became an important strategic as well as a local political centre.

For nearly 300 years (from the end of the Hsi [Western] Chin dynasty in 316/317 to the beginning of the Sui dynasty in 581), the northern territory, including the site where Peking now stands, was largely under the control of invading nomads. It was not until the T'ang dynasty (618–907) that it was recovered by the Han people and became known as Yu-chou. By the middle of the T'ang dynasty, measures were taken to prevent the nomadic Tangut



Ta-kuan Park modeled after the garden described in the novel *Dream of the Red Chamber.*
Song Lianfeng—Xinhua News Agency

tribes of Tibet, such as the Hsi Hsia, and the Khitans (a Turco-Mongolian people from Manchuria) from raiding the borderlands and the local capital. The position of Yu-chou consequently became increasingly important. On the fall of T'ang China, a number of states emerged in North China. One of these states was established by the Khitans; after destroying the city of Yu-chou, the Khitans established the Liao kingdom (947–1125) and built one of their capitals on approximately the same site, calling it Nanking ("Southern Capital") to distinguish it from other capitals in their Manchurian homeland. The Liao capital was bounded by a square wall, almost 14 miles in circumference and some 32 feet high. It had eight gates and enclosed a fine imperial palace in the centre, which indicated the strong influence of Chinese city planning.

In the mid-12th century, when the Juchen, a Tungus people from eastern Manchuria, defeated the Liao and established the state of Chin, the Liao capital was rebuilt as the new Chin capital and renamed Chung-tu ("Central Capital"). Chung-tu under the rule of the Juchen was constructed on a larger scale, with splendidly decorated palaces and halls.

Between 1211 and 1215 the Mongols, under the leadership of Genghis Khan, one of the great conquerors of history, repeatedly attacked and finally took the city from the Chin. In the battle the palaces of Chung-tu were set on fire and blazed for more than a month. When all China fell to the Mongol hordes, Kublai Khan (1215–94), a successor to Genghis Khan and founder of the Yüan, or Mongol, dynasty (1206–1368), determined to build a new capital at Peking, abandoning the old city of Karakorum in Mongolia. In 1272 he named the new capital Ta-tu ("Great Capital"); under Mongol control it became for the first time the political centre of all China.

The Mongol capital of Ta-tu

Ta-tu was larger than any of its forerunners and was rebuilt slightly northeast of the old site. The square of the outer wall measured about 18 miles in length and enclosed an area of more than 20 square miles. The city walls were built with pounded earth, and once a year labourers were called in to repair them with mud. The Imperial Palace, which was approximately to the west of the existing one, was situated in the southern half of the capital city. The chief palace architect at the time was an Arab, appointed by Kublai Khan. The city of Ta-tu exemplified the imposing and variegated architecture of the Mongol period. The square walls and the 12 gates were all modeled on the Chinese plan, but the inner chambers and living quarters were often in the style of Mongolia or of Central Asia. It was at this time that the T'ung-hui Canal was dug and connected to the Grand Canal, so that boats transporting the tribute rice from the provinces south of the Yangtze could sail into one of the new lakes inside the city. Ta-tu, which had magnificent imperial palaces and treasures drawn from every corner of the country, was the scene of stupendous feasts given by the khan (ruler) on state occasions. These characteristics and the well-organized post-stages on the roads leading to the city astounded Marco Polo, the Venetian traveler who visited Ta-tu in the 1280s.

CENTURIES OF GROWTH

The Ming and Ch'ing dynasties. In the middle of the 14th century, Chu Yüan-chang headed a peasant revolt that overthrew the Mongol dynasty and, as the Hung-wu emperor, established the Ming dynasty (1368–1644). He moved the capital to Chin-ling, near his own hometown in Anhwei Province, and called it Nanking; Ta-tu was renamed Pei-p'ing ("Northern Peace") and was placed under his son's rule. On Chu's death (1398) the throne passed to his grandson in Nanking, but his son, Chu Ti (also called the Yung-lo emperor), who ruled Pei-p'ing, usurped the throne. In consequence, in 1403 the city was renamed Peking ("Northern Capital"), and in 1421 it was officially made the capital city of the Ming dynasty.

Peking in the Ming period grew on a yet grander scale than under the Mongols. The former city walls and the extant moats, palaces, and temples were mostly built in the 15th century. The old city of Ta-tu, including the palaces, was largely demolished. The new city was situated farther southwest, thus leaving the northern part of the Mongol city derelict while at the same time slicing off one gate from the east and west walls, respectively. In 1553 an outer city wall was begun, to include the increasing number of inhabitants living outside the city, but, when the entire construction was subsequently found to be too costly, the plan was abandoned on the completion of the south wall, thus producing the present shape of the old city. Unlike the city wall of pounded earth of Mongol times, the walls of the Ming city were faced with a layer of bricks, in order to prevent weathering.

The 15th-century city

In 1644 Peking was taken over by Li Tzu-ch'eng, who led a peasant uprising against the Ming regime. Li's army held it for only 40 days, for the Manchus were simultaneously preparing an incursion south of the Great Wall, and—thanks to the treachery of a Ming general who opened the gate—they swept down on the city. Peking fell intact and in the same year was declared the Manchu capital by Shun-chih, the first emperor of the Ch'ing dynasty (1644–1911/12).

Peking remained superficially the same throughout Ch'ing times. The city plan was unaltered, though many palaces, temples, and pavilions were added outside the walls to the west, notably the Yüan-ming Yüan (old Summer Palace), built in the 17th century, and the I-ho Yüan (new Summer Palace), built in the late 19th century. The old Summer Palace was completely destroyed by fire in 1860 by the allied British and French troops during the "Arrow" War (1856–60). In the same year, as a result of the treaties of Tientsin in 1858, a permanent British embassy was established in the city, and a legation quarter, situated to the southeast of the palace ground, was reserved for British and other embassies. The legation quarter was besieged for nearly two months by the Boxer rebels in 1900.

The modern city. After the revolution of 1911, Peking remained the political centre of the Republic of China until 1928, when the Nationalists moved the capital to Nanking; Peking was again called Pei-p'ing. The city came under increasing pressure from the Japanese, who established the puppet state of Manchukuo in Manchuria in 1931. In July 1937 fighting broke out between Chinese and Japanese troops near the Marco Polo Bridge, southwest of the city; Pei-p'ing was subsequently occupied by the Japanese until 1945. After World War II the city reverted to the Nationalists, who were defeated by the Communists in the ensuing civil war. In 1949, with the establishment of the People's Republic of China, Peking (with its old name restored) was chosen as the capital of the new regime. The city soon regained its position as the political centre of China. (S.-d.C./D.M.Bo.)

In the 1950s and '60s urban development projects widened the streets and established the functional districts that characterize the modern city, but political campaigns culminating in the Cultural Revolution (1966–1976) delayed many of these projects. Beginning with the economic reforms of the early 1980s, the pace of change accelerated, and Peking has changed dramatically. New shopping centres and residential buildings have appeared throughout the city, and high-tech industrial parks have been established, especially in the suburbs. In an area located between Peking and Ch'ing-hwa universities there is a high-tech district, developed with government backing, called "Silicon Valley" because of its concentration of technology companies. Alongside small Chinese start-up companies are the research and development offices of multinational corporations. Local shops sell computer components, particularly to young people building or upgrading their computers, and the latest prices are posted daily on a Web site. One striking change, noticeable particularly in the modern shopping centres, is the emergence of a consumption-oriented middle class similar to that found in Hong Kong, Singapore, Seoul (South Korea), and other Asian cities undergoing rapid economic growth. Peking is continuing to expand its subway system Express highways now traverse the city and connect it to distant destinations. At the same time, Peking, like other major cities, is facing the growing problems of air pollution, traffic congestion, and overcrowding. In 2001 Peking was selected by the International Olympic Committee to host the 2008 Summer Games; this accelerated plans to reduce pollution, improve transportation, and beautify the city.

(Ed.)

BIBLIOGRAPHY

General works: ROSS TERRILL, *Flowers on an Iron Tree: Five Cities of China* (1975); FELIX GREENE, *Peking* (1978); and DAVID BONAVIA, *Peking* (1978), describe city life in Peking. A number of important works on scenic and structural aspects of Peking were completed during the 1930s; among them are L.C. ARLINGTON and WILLIAM LEWISOHN, *In Search of Old Peking* (1935, reprinted 1987); and JULIET BREDON, *Peking: A Historical and Intimate Description of Its Chief Places of Interest,* 3rd rev. ed. (1931). OSVALD SIRÉN, *The Walls and Gates of Peking* (1924), and *The Imperial Palaces of Peking* (1926, reprinted 1976), give details on architectural features of the old walled city. HEDDA MORRISON, *A Photographer in Old Peking* (1985), is a record of the city in the 1940s. NORMAN A. CHANCE, *China's Urban Villagers: Life in a Beijing Commune* (1984), provides insight into economic and political factors. For demographic information, see the annual *China Urban Statistics.*

Guidebooks: WILLIAM LINDESAY and WU QI, *Beijing* (1999); TOM LE BAS (ed.), *Insight Guide: Beijing,* 4th ed. (2000); PETER NEVILLE-HADLEY, *Cadocan Guides: Beijing* (2000); LIU JUNWEN, *Beijing: China's Ancient and Modern Capital* (1991); and FREDERIC M. KAPLAN, JULIAN M. SOBIN, and ARNE J. DE KEIJZER, *The China Guidebook,* 13th ed. (1993), include detailed descriptions of the city and cover its history, society, and culture.

History: Historical development is covered by RODERICK MACFARQUHAR, *The Forbidden City* (1972); and ZHOU SHACHEN, *Beijing Old and New: A Historical Guide to Places of Interest* (1984; originally published in Chinese, 1982). FRANK DORN, *The Forbidden City* (1970), associates the details of the Imperial Palaces with historical events. Other studies include PETER FLEMING, *The Siege at Peking* (1959, reissued 1984); and GEORGE N. KATES, *The Years That Were Fat: The Last of Old China* (1952, reissued 1976), covering the period of the 1930s. Accounts of life in the city in the early decades of the 20th century are found in H.Y. LOWE, *The Adventures of Wu: The Life Cycle of a Peking Man* (1940, reprinted 1983); HAROLD ACTON, *Peonies and Ponies* (1983), a novel set in Peking's expatriate community in the 1930s; and PU YI, *From Emperor to Citizen: Autobiography of Aisin-Gioro Pu Yi,* new ed., trans. from Chinese by W.J.F. JENNER (1987), the memoir of the last Chinese emperor.

(D.M.Bo./Ed.)

Human Perception

Perception, or perceiving, refers to the process whereby sensory stimulation is translated into organized experience. That experience, or percept, is the joint product of the stimulation and of the process itself. Relations found between various types of stimulation (*e.g.*, light waves and sound waves) and their associated percepts suggest inferences that can be made about the properties of the perceptual process; theories of perceiving then can be developed on the basis of these inferences. Because the perceptual process is not itself public or directly observable (except to the perceiver himself, whose percepts are given directly in experience), the validity of perceptual theories can be checked only indirectly. That is, predictions derived from theory are compared with appropriate empirical data, quite often through experimental research.

This article is divided into the following sections:

General considerations 481
- Classical problems 482
  - Sensing and perceiving
  - Temporal (time) relations
  - Perceiving as synthesizing
- Primary tendencies in perceptual organization 484
  - Gestalt principles
  - Context effects
  - Perceptual constancies
- Individual differences in perceiving 486
  - Perceptual learning
  - Sex differences in perceiving
  - Cultural influences

Perception of movement 489
- Visual cues to movement 489
- Nonvisual cues 490

Space perception 490
- General considerations 490
  - Primary gravitational effects
  - Visual factors in space perception
- Perception of depth and distance 491
- Interrelations among the senses 492
  - Factors of constancy
  - Path recognition: navigation in space
  - Social and interpersonal aspects of space perception

Time perception 494
- Sequential activities 494
  - Adaptation to successive events
  - Adaptation to periodic change
- Perception of sequence and duration 494
  - The psychological present
  - Perception of sequence
  - Perceived duration
  - Estimating duration

Illusions and hallucinations 496
- The nature of illusions 496
- Types of illusory experience 496
  - Stimulus-distortion illusions
  - Perceiver-distortion illusions
  - Other illusory experiences
  - Tactual illusions
  - Anchor-effect phenomena
  - Intersensory effects
  - Illusions of psychiatric significance
- The development of modern understanding of hallucinations 499
- The nature of hallucinations 499
  - Attention
  - Reduced sensory activity
- Induction of hallucinations 500
  - Direct brain stimulation
  - Sleep
  - Hypnagogic hallucinations
  - Dreams
  - Excessive excitation
  - Loss of sleep
  - Hypnosis and trance states
  - Sensory defects
  - Psychological factors
  - Chemical factors

Bibliography 502

General considerations

Historical background

Historically, systematic thought about perceiving was the province of philosophy. Indeed, perceiving remains of interest to philosophers, and many issues about the process that were originally raised by philosophers are still of current concern. As a scientific enterprise, however, the investigation of perception has especially developed as part of the larger discipline of psychology.

Philosophical interest in perception stems largely from questions about the sources and validity of what is called human knowledge (see EPISTEMOLOGY). Epistemologists ask whether a real, physical world exists independently of human experience and, if so, how its properties can be learned and how the truth or accuracy of that experience can be determined. They also ask whether there are innate ideas or whether all experience originates through contact with the physical world, mediated by the sense organs. For the most part, psychology bypasses such questions in favour of problems that can be handled by its special methods. The remnants of such philosophical questions, however, do remain; researchers are still concerned, for example, with the relative contributions of innate and learned factors to the perceptual process.

Such fundamental philosophical assertions as the existence of a physical world, however, are taken for granted among most scientific students of perceiving. Typically, researchers in perception simply accept the apparent physical world particularly as it is described in those branches of physics concerned with electromagnetic energy, optics, and mechanics. The problems they consider relate to the process whereby percepts are formed from the interaction of physical energy (for example, light) with the perceiving organism. Of further interest is the degree of correspondence between percepts and the physical objects to which they ordinarily relate. How accurately, for example, does the visually perceived size of an object match its physical size as measured (*e.g.*, with a yardstick)?

Questions of the latter sort imply that perceptual experiences typically have external referents and that they are meaningfully organized, most often as objects. Meaningful objects, such as trees, faces, books, tables, and dogs, are normally seen rather than separately perceived as the dots, lines, colours, and other elements of which they are composed. In the language of Gestalt psychologists, immediate human experience is of organized wholes (*Gestalten*), not of collections of elements.

A major goal of Gestalt theory in the 20th century was to specify the brain processes that might account for the organization of perception. Gestalt theorists, chief among them the German-U.S. psychologist and philosopher, the founder of Gestalt theory, Max Wertheimer and the German-U.S. psychologists Kurt Koffka and Wolfgang Köhler, rejected the earlier assumption that perceptual organization was the product of learned relationships (associations), the constituent elements of which were called simple sensations. Although Gestaltists agreed that simple sensations logically could be understood to comprise organized percepts, they argued that percepts themselves were basic to experience. One does not perceive so many dis-

crete dots (as simple sensations), for example; the percept is that of a dotted line.

The role of brain function

Without denying that learning can play some role in perception, many theorists took the position that perceptual organization reflects innate properties of the brain itself. Indeed, perception and brain functions were held by Gestaltists to be formally identical (or isomorphic), so much so that to study perception is to study the brain. Much contemporary research in perception is directed toward inferring specific features of brain function from such behaviour as the reports (introspections) people give of their sensory experiences. More and more such inferences are gratifyingly being matched with physiological observations of the brain itself.

Many investigators relied heavily on introspective reports, treating them as though they were objective descriptions of public events. Serious doubts were raised in the 1920s about this use of introspection by the U.S. psychologist John B. Watson and others, who argued that it yielded only subjective accounts and that percepts are inevitably private experiences and lack the objectivity commonly required of scientific disciplines. In response to objections about subjectivism, there arose an approach known as behaviourism that restricts its data to objective descriptions or measurements of the overt behaviour of organisms other than the experimenter himself. Verbal reports are not excluded from consideration as long as they are treated strictly as public (objective) behaviour and are not interpreted as literal, reliable descriptions of the speaker's private (subjective, introspective) experience. The behaviouristic approach does not rule out the scientific investigation of perception; instead, it modestly relegates perceptual events to the status of inferences. Percepts of others manifestly cannot be observed, though their properties can be inferred from observable behaviour (verbal and nonverbal).

One legacy of behaviourism in contemporary research on perception is a heavy reliance on very simple responses (often nonverbal), such as the pressing of a button or a lever. One advantage of this Spartan approach is that it can be applied to organisms other than man and to human infants (who also cannot give verbal reports). This restriction does not, however, cut off the researcher from the rich supply of hypotheses about perception that derive from his own introspections. Behaviourism does not proscribe sources of hypotheses; it simply specifies that only objective data are to be used in testing those hypotheses.

Behaviouristic methods for studying perception are apt to call minimally on the complex, subjective, so-called higher mental processes that seem characteristic of adult human beings; they thus tend to dehumanize perceptual theory and research. Thus, when attention is limited to objective stimuli and responses, parallels can readily be drawn between perceiving (by living organisms) and information processing (by such devices as electronic computers). Indeed, it is from this information-processing approach that some of the more intriguing theoretical contributions (*e.g.*, abstract models of perception) are currently being made. It is expected that such practical applications as the development of artificial "eyes" for the blind may emerge from these man–machine analogies. Computer-based machines that can discriminate among visual patterns already have been constructed, such as those that "read" the code numbers on bank checks.

CLASSICAL PROBLEMS

Sensing and perceiving. Many philosophers and psychologists have commonly accepted as fundamental a distinction made on rational grounds between sensing and perceiving (or between sensations and percepts). To demonstrate empirically that sensing and perceiving are indeed different, however, is quite another matter. It is often said, for example, that sensations are simple and that percepts are complex. Yet, only if there is offered some agreed upon (a priori) basis for separating experiences into two categories—sensations and percepts—can experimental procedures demonstrate that the items in one category are "simpler" than those in the other. Clearly, the arbitrary basis for the initial categorization itself cannot be subjected to empirical test. (See also SENSORY RECEPTION.)

Problems of verification aside, the simplicity–complexity distinction derives from the assumption that percepts are constructed of simple elements that have been joined through association. Presumably, the trained introspectionist can dissociate the constituent elements of a percept from one another, and in so doing, experience them as simple, raw sensations. Efforts to approach the experience of simple sensations might also be made by presenting very simple, brief, isolated stimuli; *e.g.*, flashes of light.

Perceptual learning

Another commonly offered basis for distinction is the notion that perceiving is subject to the influence of learning while sensing is not. It might be said that the sensations generated by a particular stimulus will be essentially the same from one time to the next (barring fatigue or other temporary changes in sensitivity), while the resulting percepts may vary considerably, depending on what has been learned between one occasion and the next (see below *Perceptual learning*).

Some psychologists have characterized percepts as typically related to external objects and sensations as more nearly subjective, personal, internally localized experiences. Thus, a spontaneous pain in the finger would be called a sensation; however, if the salient feature of experience is that of a painfully sharp, pointed object, such as a pin located "out there," it would be called a percept.

The above definitional criteria all relate to properties of experience; that is, they are psychological. An alternative way of distinguishing between sensing and perceiving that has become widely accepted is physiological-anatomical rather than psychological. In this case, sensations are identified with neural events occurring immediately beyond the sense organ, whereas percepts are identified with activity farther "upstream" in the nervous system, at the level of the brain. This assignment of anatomical locations to sensory and to perceptual processes seems consistent with psychological criteria. That is, the complexity and variability of percepts (both a product of learning) are attributed to the potential for physiological modification inherent in the vastly complex neural circuitry of the brain.

Temporal (time) relations. Clearly, many subjective processes (such as problem solving) take time to run their courses. This is true even for such relatively simple activities as perceptual discriminations in the size of different objects. It is not readily apparent, however, whether

By courtesy of Bell Laboratories

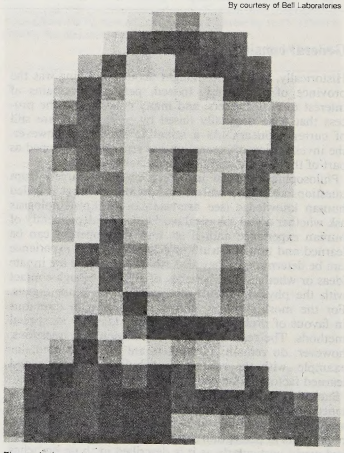


Figure 1: An ambiguous picture. Increasing viewing distance permits more precise perception (see text).

percepts themselves—which, for example, might enter as elements in problem solving—take time to form. To the naïve observer, percepts probably do seem essentially instantaneous: the moment a square is shown, a square seems to be seen. Yet, experimental evidence suggests that percepts, even of simple geometric forms, follow a measurable, developmental time course. In some instances the temporal development of percepts is relatively long (on the order of seconds), and in some it is quite brief (on the order of thousandths of a second).

Pictures that are incomplete or ambiguous provide good examples of relatively long-term temporal development of percepts. Look at Figure 1 and continue looking until you see something more than a pattern of black, gray, and white patches. Abruptly, you probably will perceive a familiar face that, on subsequent viewing, will reappear to you without difficulty. How long it takes for such a percept to develop will vary considerably from one person to another, perhaps revealing fundamental differences among individuals in their speed of perceptual processing. It might be instructive to show Figure 1 to several people, and with the aid of a stopwatch, measure the time it takes each of them to achieve the desired percept, both initially and then on some later occasion. (Figure 1 commonly is seen by most people as the face of Abraham Lincoln.)

A somewhat different way in which percepts may change with time is illustrated in the section *Illusions and hallucinations,* below. On initial viewing of this type of drawing, one will probably immediately see a meaningful picture. After continued gazing at the drawing, the initial percept may abruptly be replaced by another. Thereafter, the two percepts should alternate with the passage of time. Stimuli of this sort (which can yield more than one percept) raise such questions as, for example, what determines the initial percept; why do some people first see a vase whereas others see two profiles; why does the initial percept give way to the alternate; what determines the rate of fluctuation from one percept to the other; do differences from one person to another in the rate of fluctuation of ambiguous figures indicate fundamental differences in perceptual activity? Tentative answers to such questions continue to be proposed.

Instances of slowly developing percepts require relatively simple procedures to uncover. Those percepts with a very rapid time course may be studied with the aid of instruments known as tachistoscopes that permit the durations of visual stimuli to be precisely controlled. Sophisticated electronic tachistoscopes flash reliably for periods as brief as one millisecond (one-thousandth of a second). Such precision permits study of the short-term development (microgenesis) of such percepts as those of simple geometric figures. Thus, it has been found that perception of a small black disk is disrupted (masked) by the rapidly successive tachistoscopic presentation of a second stimulus: a black ring that fits snugly around the disk. Indeed, as far as experimental subjects can tell, the disk target simply does not appear, though when flashed without being followed by the ring, it is readily detectable. Other target and mask stimuli also have been successfully employed.

Backward masking, or metacontrast

It has been theorized that it takes time for any percept to develop; that one's experience of a figure such as a disk develops from the figure's centre outward; that one's percept of the disk becomes stable only when the outer contour is appreciated; and that the ring functions in backward masking (metacontrast) because in the course of its own emergence as a percept the inner contour of the ring "absorbs" the perceptually developing contour of the disk. (Unless the viewer becomes aware of its contour, the disk theoretically cannot be perceived.) This interpretation is consistent with evidence of an optimal time interval between disk and ring onsets (about 30 to 50 milliseconds) for the best masking effect. Thus, masking is most evident at the moment developing awareness of disk contour is held to coincide in space and time with the initial perceptual growth of the ring's inner contour. If the ring is presented too soon or too late, theoretically, the contour absorption on which masking presumably depends is ineffective.

While such theories are controversial, masking experiments in general do clearly indicate that in human beings there is a brief period (100 to 200 milliseconds at most) during which a percept is highly vulnerable to disruption. Whatever its exact mechanism, the phenomenon of masking manifestly demonstrates that percepts do not emerge instantaneously and full-blown at the moment of sensory stimulation. Thus, assuming that percepts are synthesized from simpler elements, relatively complex percepts would be expected to take longest to develop and, hence, to be most vulnerable to masking. Yet empirical studies show just the opposite, indicating that the more complex the visual target, the more difficult it is to mask.

Perceiving as synthesizing. The organization apparent in percepts has been attributed by some to learning, as being built up through arbitrary associations of elements that have repeatedly occurred together in the person's experience. Other theorists (particularly Gestaltists) stress the view that perceptual organization is physiologically inborn, being inherent in innate aspects of brain functioning rather than depending on a synthesizing process of learning to combine simpler elements into more complex, integrated wholes. One way of resolving such theoretical disputes would be to deprive people from birth of all visual sensory experience and, hence, of all opportunity for visual perceptual learning. Then at the time normal sensory function was restored, they would need to be tested to determine what perceptual functions, if any, were intact. Such a strategy was proposed in a letter to the British philosopher John Locke by a fellow philosopher William Molyneux in 1690. Molyneux's suggestion waited until the 20th century to be taken seriously, after surgical methods had been found to restore the sight of people born blind because of cataract (clouded lens within the eye).

The use of cataract surgery to study perception

After removal of their cataracts, such newly sighted people are found to be normally sensitive to changes in intensity of illumination and to colour. Though they are able initially to tell when a figure is present, they cannot at first discriminate one simple shape from another, nor can they readily remember the shape of a just-exposed object. This deficiency extends to such socially important visual stimuli as people's faces. Only after a long and painstaking period of experience—perhaps of several months duration—do such seemingly primitive visual performances as discriminating a square from a triangle come easily. Until then, the person must count corners, for example, to achieve accurate discrimination.

Findings derived from cataract surgery have provided a rich source of hypotheses for further research, including posited neurophysiological mechanisms (*e.g.,* assemblies of brain cells) that might serve as the medium for the structural changes presumed to accompany perceptual learning. This situation led to experimental attempts to show how, through repeated stimulation, the perceptual system could progress from performance of only very primitive functions to the highly complex operations (such as form identification and discrimination) that are characteristic of the mature organism.

A host of experiments using laboratory animals as subjects (*e.g.,* pigeons, rats, cats, dogs, monkeys, and chimpanzees) have been conducted to determine, under rigorous experimental control, the extent to which learning early in life contributes to later perceptual functioning. By analogy with humans born with cataract, such animals were deprived of visual experience from as close to birth (or hatching) as possible; *e.g.,* chimpanzees were reared in darkness. In another type of experiment, animals were reared in environments that provided more than the normal amount and variety of stimulation or were exposed to specific stimuli they might not ordinarily encounter. These research strategies are said to provide impoverished or enriched environments. Experiments of both sorts have consistently provided verification of the general hypothesis that early perceptual experience plays an important role in later perceptual (as well as intellectual and emotional) development, even producing changes in brain weight and biochemistry. This research also offers a strong scientific rationale for efforts to enrich the environments of so-called disadvantaged or culturally deprived children.

Studies of human infants indicate that their early per-

ceptual experiences are not the "blooming, buzzing confusion" postulated by the U.S. psychologist William James late in the 19th and early in the 20th century. Rather, even infants one or two days old are capable of refined visual discriminations. Recording of the infants' visual fixations as their eyes move indicates them to have reliable preference for one paired stimulus over another, giving evidence of visual discrimination. Research employing this technique shows that preferences among various visual patterns or shapes generally are related to the complexity and novelty of the stimuli, for infants and for older human subjects as well. Evidence of this sort seems out of keeping with the findings of cataract surgery, which suggest that figural discrimination is not innate for the visually naïve. It may be, however, that adults who were blind with cataracts at birth suffer more than mere lack of normal visual experience; they are not quite comparable to visually naïve, but otherwise intact, infants. It may be that visual experience is necessary, not to generate pattern perception but to maintain it; that is, the infant's built-in (innate) perceptual abilities may somehow deteriorate through disuse.

The visual cliff technique

Similar research has dealt with visual depth perception in laboratory animals and human babies. One technique (the visual cliff) depends on the evident reluctance of young animals to step off the edge of what seems to be a steep cliff. The so-called visual cliff apparatus in one of its versions consists of a narrow platform on which the subject is placed and two wide platforms on either side of it. Although both flanking platforms are equally and only slightly lower than the central platform, the subject sees visual patterns designed so that one looks much deeper than the other. Typically, the subject explores the central platform and then investigates the flanks, finally stepping down onto the shallow-appearing side. By this response, the subject indicates sensitivity to visual depth cues. To discover if prior visual experience is necessary for the typical avoidance of the flanking platform that looks deepest requires subjects able to locomote well from the start (*e.g.*, chicks) or those deprived of visual experience (*e.g.*, rats) until their locomotor ability has developed. Research with the visual cliff clearly demonstrates the presence of depth perception in visually naïve subjects.

In summary, there is evidence among cataract sufferers of the necessity of early visual experience in human visual pattern discrimination; laboratory animals reared in impoverished or enriched environments demonstrate the importance (if not necessity) of early visual stimulation for perceptual development; human infants and other visually naïve animals behave as if they are innately capable of pattern and depth perception (*i.e.*, without the need for learning). These data together suggest that some basic visual functions, including pattern perception, are built in but that visual experience serves to maintain and elaborate them.

Structuralism

In a theory called structuralism, that everyday perceptual experience is structured or synthesized from "sensations," psychologists such as the English-U.S. introspectionistic psychologist Edward Bradford Titchener even devised a formal method of introspection for experimentally analyzing (or taking apart) percepts in an effort to reveal their constituent elements. The procedure required that the introspecting experimental subjects learn to avoid reporting on their experiences as they naïvely seemed. To establish this way of treating experience required careful training. One consequence of this training is that the observer's introspective reports may be contaminated by his expectations and hence may, in all honesty, reflect little more than his theoretical biases.

But the problem remains interesting: If percepts are indeed syntheses of simpler elements, can those elements be made to appear in experience? If so, what will they turn out to be? Can this problem be investigated without recourse to the structuralist's method of introspection and reliance on the reports of strongly biassed observers?

Evidence that percepts have constituent elements emerged serendipitously from research on stabilized retinal images. The image cast on the retina of the eye by a fixed object normally is continually moving because the perceiver himself is always in motion. Even when dampened by physical restraint, some residual movement will be left, attributable largely to high-frequency tremors (nystagmus) of the eyeballs. If the perceiver functioned as if he were a camera, the normal instability of the retinal image would produce a blurred percept and a concomitant impairment of visual acuity.

It is not feasible to eliminate eye movements, but it is possible to stabilize or fix the location of the retinal image by coupling the source of the image to the eyeball itself. An optical lever system can be so adjusted that when the eye moves the image source moves with it, and potential motion in the retinal image is eliminated. As expected, visual acuity is slightly enhanced when the retinal image is kept motionless. A remarkable, unexpected finding, however, was that such stabilized images rapidly seem to disappear, the perceiver losing awareness of them. It would seem that some movement in retinal image is needed to maintain perception over extended periods of time.

The optical lever system

One limitation of the optical lever system is that it permits the use of only very simple targets, such as straight, vertical lines. With a different device (in effect, a miniature projector attached to the eyeball), stabilized images of complex patterns may be presented. Complex patterns are found to produce percepts that are relatively slow to deteriorate; furthermore, they do not disappear in toto. The manner of the fragmentation is perhaps revealing of the way in which complex percepts are synthesized. Speaking metaphorically, observing how percepts "come apart" under retinal stabilization may be very much like discovering the structure of a rock by striking it with a powerful hammer blow.

Indeed, under retinal stabilization, single lines seem to disappear and reappear in a unitary (altogether) fashion. In a figure comprised of several lines (say, a square), percepts of parallel lines are likely to disappear and reappear together; proximity also affects the joint perceptual fate of pairs of lines. Retinally stabilized segments of such geometric figures as circles and triangles can seem to disappear and reappear without implicating the entire figure. In the disappearance of percepts of triangles, lines rather than angles are the functional units. (This finding is embarrassing to earlier theorizing about the crucial role of angles in the development of the neural network underlying the percept of a triangle.)

Clearly, with stabilized images, the constituent perceptual elements of complex geometric forms are lines, straight or curved; and lines with the same orientation are likely to have similar perceptual fates, as though forming a higher-order component of complex patterns than do individual lines. These conclusions are remarkably similar to those drawn from studies of the effect of visual stimuli on the electrical activity of single neurons in the cerebral cortex. A finding of major theoretical significance is the failure of percepts of circles, squares, and triangles to act as units. Such percepts are treated in classical Gestalt theory, however, as though they are basic and unitary and not readily decomposable.

PRIMARY TENDENCIES IN PERCEPTUAL ORGANIZATION

Gestalt principles. Gestalt theory was meant to have general applicability; its main tenets, however, were induced almost exclusively from observations on visual perception. Whatever their ultimate theoretical significance, these observations have been raised to the level of general principles. It is conventional to refer to them as Gestalt principles of perceptual organization.

The overriding theme of the theory is that stimulation is perceived in organized or configurational terms (*Gestalt* in German means "configuration"). Patterns take precedence over elements and have properties that are not inherent in the elements themselves. One does not merely perceive dots; he perceives a dotted line. This notion is captured in a phrase often used to characterize Gestalt theory: "The whole is more than the sum of its parts."

Of the many principles of organization that have been enunciated by various Gestalt theorists, the most general is referred to as *Prägnanz*. In effect, according to the prin-

The notion of "good" configurations

ciple of *Prägnanz*, the particular perceptual configuration achieved, out of a myriad of potential configurations, will be as good as prevailing conditions permit. What constitutes a "good" configuration, or a poor one, is unfortunately not clearly specified, though several properties of good configurations can be listed, chief among them being simplicity, stability, regularity, symmetry, continuity, and unity. What happens when these properties of figures come into conflict is not specified, but should be possible to determine empirically.

The principle of closure often operates in the service of *Prägnanz;* for example, a circular figure with small gaps in it will be seen as a complete or closed circle. Similarly, if a portion of the image of a figure falls on the blind spot of the retina, a complete figure often will still be perceived. Some distortions from good configuration may be so large as to preclude closure; in those cases, the figures may be a source of tension for the observer.

Prägnanz may also be achieved through good continuation; this principle describes a tendency for smooth continuity of contour to be dominant over discrete, irregular, abruptly changing contours. Thus, a figure composed of the overlapping outlines of an ellipse and a rectangle will probably be seen as such rather than as three figures, each with irregular, noncontinuous borders.

Closure and good continuation represent two of the factors that are held to determine what percepts will emerge from a complex stimulus. Implicit in them (and in the general principle of *Prägnanz*) is the assumption that whenever possible some figure will be perceived; more specifically, that the visual field will be articulated into figures and patterns of figures. It is understood that such emerging patterns are not *in* the stimulus. Although they are *permitted by* the stimulus, they are created by the perceptual system; that is, by the perceiver himself.

In the illustrations in Figure 2, in the panel on the left, the vertical distance between elements is less than the horizontal distance. By virtue of this differential proximity, the elements become perceptually organized into columns. In the right-hand panel, similarity, another principle of organization, is operative. Here, by virtue of similarity in brightness, the visual field tends to be perceptually articulated into alternating sets of black and gray rows.

It is not at all obvious why organization by similarity should occur; physical stimulation allows but does not demand it. Clearly in that case the articulation of the visual

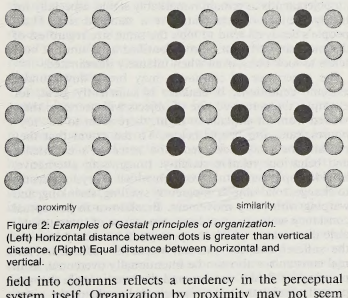


Figure 2: *Examples of Gestalt principles of organization.* (Left) Horizontal distance between dots is greater than vertical distance. (Right) Equal distance between horizontal and vertical.

field into columns reflects a tendency in the perceptual system itself. Organization by proximity may not seem to reveal anything more than a close correspondence between perception and stimulation. (Though as argued by the Gestalt theorist Kurt Koffka, it is not an adequate explanation to say that "things look as they do because they are what they are.") Yet, when a proximity pattern like the one shown in Figure 2 was briefly presented and subjects were asked (under guise of another task) to reproduce what they saw, many people failed to indicate a differentiated percept of columns. Instead, they reproduced a homogeneous matrix of elements. After repeated exposures, some of those subjects began to draw proximity-based columns of elements. Organization according to the principle of proximity seems to be neither universal nor, for those who achieve it, immediate.

Proximity patterns

In the latter experiment, people who failed to obtain the differentiated percept of columns scored significantly lower on a test of verbal intelligence than did those who succeeded at some point in the experiment. Perhaps the Gestalt principles of organization apply to perceivers (such as Gestalt theorists) whose intellectual development has reached a high degree of maturity. When organized percepts are easy to come by, gradations in intelligence do not seem to matter; when some barrier to organization is imposed (as by brief stimulus exposure), however, then the effect on perception of such differences among individuals may show up.

If people see a pattern of columns in the left panel of Figure 2 because that is how the stimulus is constructed, then why do some people not see it that way? Both achieving and failing to achieve an organized percept must be explained. Surely, part of the explanation must lie in the nature of the perceptual process itself. Thus, the experimental results indicate that perceptual organization is not universal and immediate; rather, they support the major tenet of Gestalt theory that things look as they do because of the organization imposed by the perceptual process (*i.e.*, by the perceiver).

Gestalt views of movement

One Gestalt principle, that of common fate, depends on movement and is quite striking when observed. According to the principle of common fate, stimulus elements are likely to be perceived as a unit if they move together. An illustration of this principle is provided by a well-camouflaged object, such as a military vehicle; when stationary, the elements of the vehicle are integrated, through proximity, similarity, and so on, into patterns of background elements, and the object is difficult to detect. But it is easy to see it once it starts moving; with all of its elements moving in unison, the vehicle is readily perceived as a unitary figure, clearly segregated from its background.

Movement is also at the heart of a set of observations of considerable significance in the historical development of Gestalt theory. These observations concern circumstances in which people perceive movement in the absence of actual physical motion of the stimulus. One familiar instance of this class of events is referred to as the phi phenomenon. In simplest form, the phi phenomenon can be demonstrated by successively turning two adjacent lights on and off. Given appropriate temporal and spatial relations between the two lights, an observer will perceive the first light as if it were moving from its location to that of the second light. The phi phenomenon is basic to the eye-catching displays used on theatre marquees and to cinematic and television presentations. The motion-picture screen, for example, presents a series of briefly flashed, still images; the movement people see is a creation of their own perceptual systems (see below *Perception of movement*).

It is the lack of one-to-one correspondence between stimulation and perception, as dramatically illustrated in the phi phenomenon, that underscores the Gestaltists' dissatisfaction with stimulus-bound models of perception and their insistence on the priority of patterns and relations. What people perceive is determined not only by what is present at the point under direct observation but also by what is occurring in the total stimulus context or display.

Context effects. One of the simplest instance of relational (or context) effects in perception is that of brightness contrast. Thus, the apparent brightness of a stimulus depends not only on its own luminance but also on that of the surrounding stimulation. The same gray square looks whiter against a dark background and blacker when placed in a bright surround. Similarly, a white or gray patch will take on an apparent hue that is complementary to the colour of the surround (*e.g.*, the patch will seem tinged with yellow when it is placed against a blue background).

Analogous context effects

Analogous context effects are evident in many commonplace experiences. A man of average height seems to be a runt when he is on a basketball court with much taller players; yet the same man looms like a giant when refereeing a game played by little boys. It is known that a typical winter's day seems delightfully balmy when temperatures rise after a week of subfreezing weather.

To the Gestaltist, contrast effects dramatize the relational nature of perception. They also play a significant role in a more recently developed adaptation-level theory, which also provides a general perceptual model. At the core of the model is the notion that the manner in which a stimulus is perceived depends not only on its own physical characteristics but also on those of surrounding stimuli and of stimuli previously experienced by the observer. In other words, the perceiver is said to be perceptually adapted to past sensory stimuli; his adaptation level forms a kind of zero point against which any new stimulus is perceived. An example is provided by the almost overwhelming silence one experiences when the sound of an air conditioner (to which he has adapted) suddenly ceases.

Gestalt theorists also attached significance to the observer's history of stimulation; indeed, some of them interpreted so-called figural aftereffects within a Gestaltist model of brain functioning. Figural aftereffects refer to changes in the perceived shape or location of a figure following its inspection; for example, a curved line will appear to get straighter after prolonged inspection. Or the distance between two parallel lines seems to change as an aftereffect of previous inspection.

In a typical experiment one looks at a point adjacent to a dark vertical bar (the inspection figure) on a screen. Following this inspection period, the dark bar is replaced by two identical pairs of vertical lines, one pair on either side of the region where the bar had been, the second pair alongside in a region not previously exposed to the inspection figure. The subject again fixates the same point. A figural aftereffect shows up as a greater apparent distance between the pair of lines surrounding the region of the inspection figure even though the other pair is actually identical. This distortion is not simply a generalized contrast effect because it occurs only in the small area along the borders of the inspection figure; that is, the effect is localized and restricted.

It thus has been speculated that visual exposure to a figure induces in the brain a condition of localized satiation.

Localized satiation in the brain

The passage of electrical activity is assumed to be impeded in satiated areas of the brain. Moreover, it is postulated that the perceived distance between two borders of a figure is directly related to the time it takes for electrical currents to pass between them. Thus, it is held that one effect of satiation is to increase the apparent distance between the borders of a figure that straddles a satiated region. Whatever the merits of such physiological speculations, they have stimulated a vast amount of research on figural aftereffects. Good evidence for similar effects in other senses, such as touch, also has been obtained. Clearly, perception can be influenced not only by the context of current background but also by the residues (after-effects) of previous stimulation.

Concurrent visual stimulation may modify one's acuity in detecting auditory stimuli. Similar interactions are claimed to occur for other combinations of senses. Some dentists report success in using audioanalgesia, in which stimulation with sound waves is said to reduce the experience of pain in the mouth. The high specificity of some of the reported sensory interactions seems to preclude an explanation that concurrent stimulation works by changing the subject's general level of alertness. However these intersensory effects might be mediated, they do suggest that the brain does not function as a collection of entirely independent sensory channels. As a physical system, the brain follows physical principles; thus overlapping and spreading or waning fields of neural excitation in the brain have been theorized to underlie such phenomena as closure and audioanalgesia. Köhler referred to these models of neural analogues of perceptual phenomena as physical Gestalten; unfortunately, there is little direct physiological evidence for them.

An alternative to field effects in brain functioning is the assumption that local stimulation gives rise, in one-to-one fashion, to a mosaic of local responses. Implicit in the mosaic hypothesis is a kind of telephone switchboard model of the brain as a machine in which the electrical activity is strictly confined to separate pathways of neurons that are well insulated (isolated) from one another. The Gestaltists rejected this model because in its early formulations it did not explain intersensory and intrasensory perceptual phenomena. A more sophisticated machine model, however, provides for fieldlike effects through the operation of complex networks of neural elements. It is held that electrical activity remains confined to discrete pathways, but that these pathways do not simply travel straight through the system; that they also interconnect, with both excitatory and inhibitory consequences. Supporting evidence comes from records of the electrical activity in single neurons in the cat brain; when the cat's eye is probed by a small spot of light, a specific area on the retina can be found that serves to excite a given brain neuron.

Further mapping of the cat retina often uncovers inhibitory areas adjacent to the one that is excitatory; that is, when light strikes those retinal areas the activity in the brain neuron being monitored is depressed. The excitatory and inhibitory areas thus comprise the brain neuron's retinal receptive field. Analogous inhibitory effects have also been found in research on the eye of the crab, *Limulus*. Such context effects as brightness contrast could be based on these simple inhibitory mechanisms. It remains to be seen, however, just how many perceptual phenomena that fit Gestalt field theory also can be handled by sophisticated variants of the machine or mosaic model.

Perceptual constancies. Even though the retinal image of a receding automobile shrinks in size, the normal, experienced person perceives the size of the object to remain constant. Indeed, one of the most impressive features of perceiving is the tendency of objects to appear stable in the face of their continually changing stimulus features. Though a dinner plate itself does not change, its image on the retina undergoes considerable changes in shape and size as the perceiver and plate move. What is noteworthy is stability in perception despite gross instability in stimulation. Such matches between the object as it is perceived and the object as it is understood to actually exist (regardless of transformations in the energy of stimulation) are called perceptual constancies.

Dimensions of visual experience that exhibit constancy include size, shape, brightness, and colour. Perceptual constancy tends to prevail for these dimensions as long as the observer has appropriate contextual cues; for example, perception of size constancy depends on cues that allow one a valid assessment of his distance from the object. With distance accurately perceived, the apparent size of an object tends to remain remarkably stable, especially for highly familiar objects that have a standard size. Thus, people's heads all tend to look the same size regardless of distance; similarly, an object identified as a lump of coal tends to look black even when intensely illuminated.

Loss of perceptual constancy

The experience of constancy may break down under extreme conditions. If distance is sufficiently great, for example, the perceived size of objects will decrease; thus, viewed from an airplane in flight, there seem to be "toy" houses, cars, and people below. To the extent that they prevail, the constancies lend the perceiver's experience and behaviour relative stability. Imagine an alternative, kaleidoscopic perceptual world in which everything seems to change, solid objects apparently swelling, shrinking, and warping with every movement. Breakdown in perceptual constancy seems to complicate the course of some psychiatric disorders in which the perceptual boundary between the sufferer and the external world is weakened. Normal constancies also can be intentionally overcome, as in paintings of flabby watches and distorted people that apparently depict the unique perceptual world of the artist.

INDIVIDUAL DIFFERENCES IN PERCEIVING

Theoretical assertions about perceiving are often made as though they apply indiscriminately to all organisms, or at least to all people. Perhaps perceptual principles of such great generality eventually will be uncovered. In the meantime it is evident that there are clear differences in perceptual functioning among individuals, among classes of individuals, and within the same individual from one occasion to another.

Perceptual learning. That perceptual functioning should change with the perceiver's age is expected on the grounds

that psychological development stems from maturation and learning. Indeed, empirical evidence for age-related changes in perceiving is substantial. There are, for example, reliable data that perceptual constancies are enhanced with the person's increasing age, improvement leveling off at about age ten. Similarly, there is a great deal of evidence for both decreased and increased susceptibility to various optical illusions with increasing age (see *Illusions and hallucinations*). Those illusions that become less pronounced with increasing age probably depend on the subject's changes in scanning and on his increased ability to segregate parts of a pattern from one another; illusions that become more pronounced probably reflect the operation of expectancies that develop through experience. Anatomical and physiological changes in the eye itself also may account for some age-related perceptual changes. (W.N.D.)

Optical illusions among older persons

Historically, the perceptual role of learning was a source of controversy. Vigorous denials that perceiving is influenced by learning are found in arguments of early Gestalt psychologists (*e.g.,* Max Wertheimer, 1880–1943, a German). By contrast, heavy reliance is placed on learning processes in the writings of the German philosopher and scientist H.L.F. von Helmholtz (1821–94). Today, there is virtually full agreement that perceiving is modified by learning. Disputes now focus on the process of perceptual learning itself. Most theoretical alternatives reflect two underlying themes: discovery and enrichment. The discovery thesis is reflected in Eleanor J. Gibson's view that perceptual learning is a process of discovering how to transform previously overlooked potentials of sensory stimulation into effective information. Enrichment theories depict perceptual learning as enriching sensory experience with specific associations and with rules for its interpretation that derive from past experience. Discovery theories propose that perceptual modification results from learning to respond to new aspects of sensory stimuli, while enrichment theories hold that such modification results from learning to respond differently to the same sensory stimuli.

Theories: discovery versus enrichment

Direct confrontations of these positions are rare, their advocates tending to differ in their selection of experimental procedures and learning situations. It may be that discovery and enrichment theories are compatible, simply accounting for different forms of perceptual learning.

General acceptance of the perceptual role of learning should not be taken to endorse the claim that perceiving originally depends on learning. Indeed, studies of human newborn and very young infants indicate highly organized and stable perceptual functions. Learning is to be regarded as supplementary to unlearned factors that mediate perceiving.

Effects of practice. The most direct examination of perceptual learning is provided by investigating the effects of practice. In so-called detection tasks the observer is required to detect the presence or absence of a selected stimulus. For example, effects of practice on visual acuity were studied by requiring observers to detect simple orientation (left or right) in a row of leaning letters; *e.g.,* ε ε ε ε ε . Practice tended to lower acuity thresholds, defined as the lowest intensity of illumination at which each observer could detect the orientation. Or, observers were asked to say when they just could see that an approaching pair of parallel bars was double. With practice they continued to report seeing the narrow space between the bars at increasing distances. Such improvements suggest that sensitivity to simple (unidimensional) stimuli is not immutable, being modifiable through practice.

Detection of stimuli

Improvement is not limited to simple variables. In one visual-search procedure, subjects scanned a long list of letters to find a single letter that appeared only once. Search time was reduced by a factor of 10 following extensive practice, after which 10 different letters could be detected as quickly as a single letter. Practice effects with complex targets also have been studied. In one experiment, two rows of figures were displayed on each trial, one with four simple outlines of geometrical figures, the other containing three complicated figures. Subjects were to guess or detect which one of the simple figures was concealed (embedded) in all three of the complex figures. Again, ability to identify the correct simple figure improved with practice.

Tasks involving absolute judgment require much more of the observer than does simple or complex detection. For example, he may be asked to estimate the diameters of circular targets numerically (*e.g.,* in inches or centimetres). In a similar study, two groups of subjects made absolute judgments of widely varying distances outdoors, both before and after interpolated activity. One group spent the interpolated period estimating a large number of other distances, none the same as in the original series. The other group spent the interval on unrelated paper-and-pencil work. In the first (extra-practice) group, judgments became more accurate and less variable than among the pencil-and-paper workers. Increased precision following practice also has been reported for absolute judgments of odour intensities, and of multidimensional visual (colour) and auditory stimuli. Improvement with practice is observed even when the subject remains uninformed of his accuracy; correcting him seems to confer slight benefit.

Many studies have failed to establish a clear basis for observed improvements in altered perceptual sensitivity or discriminability. For example, better performance on an acuity test may result from adopting a new criterion of visual doubleness or from learning how to use characteristics of blur to infer slant among leaning Es. Such uncertainties cloud the theoretical and practical significance of much available data.

U.S. psychologist William James (1842–1910) probably introduced the notion that practice in labelling stimuli can alter their discriminability. Indeed, sometimes vague visual forms that are distinctively named are easier to discriminate (acquired distinctiveness). If several such stimuli have the same verbal label, discriminability may be reduced (acquired equivalence).

Labelling effects in the laboratory have been discouragingly fragile, however, and factors that favour them are poorly understood. Perhaps labelling affects one's efforts to discover distinguishing characteristics of stimuli. Having him learn distinctive labels may encourage him to analyze sensory features more fully. Or it may be that he begins to perceive a compound stimulus that includes the visual form and its associated label. If labels differ, the presumed compound stimuli are different, and discrimination should be enhanced. These hypotheses express both the discovery and enrichment theses.

Effects of perceptual assumptions. According to one version of the enrichment thesis, exposure to recurrent regularities among stimuli prompts one to assume specific relationships between the environment and his sensory experience. For example, one learns that a continuous sequence of projective transformation (*e.g.,* the circular profile of a dinner plate seems to become elliptical) is associated with changing positions of the object in view, or that continuous symmetrical expansion of the retinal image is associated with approach. In addition, one presumably learns to make assumptions about what is called reality; *e.g.,* despite alterations in retinal image, one perceives the plate to stay the same size. Psychologists Adelbert Ames, Jr., and Egon Brunswik proposed that one perceives under the strong influence of his learned assumptions and inferences, these providing a context for evaluating sensory data (inputs). In keeping with enrichment theory, Brunswik and Ames contended that sensory stimuli alone inherently lack some of the information needed for mature, adaptive perceiving; enrichment was held necessary to reduce ambiguity.

Much of the evidence for the contention that all perceiving is modified by one's assumptions comes from investigations in which most of the visual, everyday stimuli are eliminated. Often, the subject may view an isolated target in total darkness or look at a motionless display while keeping his head steady. To show that learned assumptions about physical size affect perceived distance, the observer may be asked to judge how far he is from a rectangle of light displayed against total darkness. He is told at one time that the rectangle is a calling card; at another it is called a business envelope. His assumptions about these objects in relation to the size of his retinal

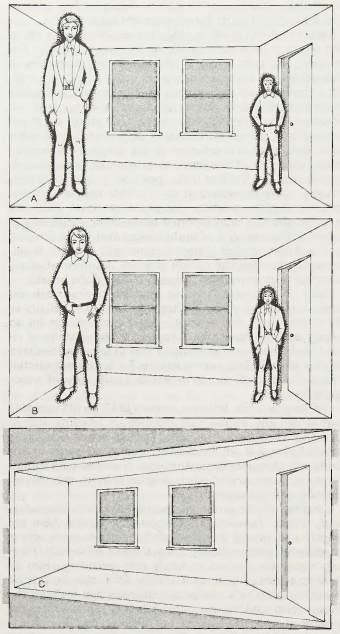


Figure 3: Perception modified by learned assumptions in a distorted room. The same men change places in A and B in a room modelled after C.

image are invoked as prompting him to say that the "envelope" looks more distant than does the "calling card." Dramatic examples of this effect were invented by Ames, including his famous distorted room (see Figure 3).

Ames held that perceiving under unusual conditions (*e.g.*, in a dark room) follows the same principles that govern more ordinary experience. The special conditions are said to permit experimental scrutiny of the same processes that are so difficult to examine under ordinary, uncontrolled conditions.

An opposing view is that such perceptual assumptions and inferences operate *only* under specific experimental conditions. It is asserted that only when commonly available sources of information are eliminated is the subject forced to rely on assumptions.

In the tradition of Helmholtz, Ames and Brunswik seemed to liken perceiving to reasoning, although not as a conscious process. They held that perceptual assumptions, once established, are influenced only slightly by logic. Although the floor and ceiling of the distorted room are sloped and all windows are of different size, it projects the same retinal pattern as a normal room; and a naïve subject will report that he sees an ordinary room. But even after he explores the room he remains likely to say it looks rectangular as before, despite his new informaton. Comparable observations have been reported for a variety of situations. Familiarization or instruction seems to have little effect on long-established perceptual assumptions.

Reasoning and perceiving

(W.E.)

Psychoanalytic theory explicitly calls for motivational influences on such functions as memory, thinking, and perceiving. In particular, the theory is concerned with unconscious motives and conflicts and with unconscious defenses (such as repression) used to control them. According to the psychoanalytic hypothesis, there should be wide perceptual variation among individuals in response to stimuli that have motivational significance. At any rate, a host of experiments have been designed to show that perceiving is indeed subject to unconscious influences.

Psychoanalytic theory

In some studies, for example, it seemed that so-called obscene words flashed on a screen had to be exposed longer than apparently neutral (control) words before their meaning could be perceived. In the other studies, children of poor families have been found to overestimate the size of coins as compared with the judgments of children of richer families. One major problem with such research lies in finding or creating appropriate experimental and control stimuli. Considering differences in the use of language, for example, it is most unlikely that what once were widely called obscene words would currently evoke the conflicts and defenses of more than a few subjects.

Assuming suitable stimuli can be found, an even more serious problem arises around the interpretation of the subjects' behaviour; for example, do people really find it more difficult to recognize obscene words or are they simply reluctant to admit recognition? Problems of this sort have plagued researchers, and unambiguously interpretable experiments in this field are most difficult to produce. The hypothesis of such individual influences as motivation on perception remains appealing and viable, but unproved.

(W.N.D.)

Information discrepancy. Striking examples of perceptual learning are observed when one receives sensory data that contradict earlier experiences. For example, spectacles containing a wedge prism will bend light rays to displace images on the retina. An object thus will be seen as if it were somewhere other than its ordinarily perceived position. The subject's initial attempts to touch the target will be misdirected, and there is a discrepancy between its location as seen and as felt. A right-angle prism will tilt the visual scene to any desired degree, altering the customary direction in which retinal images move. Usually, images of stationary objects move parallel to the direction of head movement; now their motion is at an angle to the head's path.

However, if an observer wears such eyeglasses for an extended period, objects no longer seem displaced, nor does the scene continue to appear tilted. The observer has adapted to the prismatic distortions and comes to perceive the environment as he did pre-experimentally. Similarly adaptation to the perceptual aftereffects rapidly occurs after the prism is removed in such experiments.

Adaptation

Adaptation may be interpreted as perceptual learning that results from exposure to discrepancy. People who wear prism spectacles during active, self-initiated movement tend to show a greater degree of adaptation than do those who sit still or who are moved passively. Apparently conditions that heighten exposure to discrepancies facilitate adaptation. It seems likely that adaptation reflects a learning process during which the perceiver re-evaluates one or more sources of sensory information to reduce his experience of discrepancy. For example, information generated by receptors that respond to tension in skeletal muscles may be re-evaluated to resolve a discrepancy between felt and seen position.

It often is suggested that adaptation to prism eyeglasses may involve the same processes that serve perceptual development in infants. Indeed, some conditions that experimentally facilitate adaptation to prism distortion also seem necessary for everyday perceptual development (*e.g.*, active, self-initiated movement). In work reported by Richard Held (*Scientific American*, November 1965), actively moving kittens developed visually guided movements normally. When each of these was yoked to a littermate that was pulled passively over the same path, the passive partner failed to develop normal perceptual function. Yet both kittens apparently received identical visual stimuli.

The effects of learning on perceiving are varied. Most of these involve learning to respond to new stimuli or to

make new responses to old stimuli. The one case consists of differentiating previously neglected stimulus characteristics; the other is a matter of re-evaluating stimuli and learning to respond to them differently.

(W.E.)

Sex differences in perceiving. It is difficult to assess the degree to which differences related to the sex of the perceiver are biologically based or are the cultural product of traditional differences in sex role. Biological sex and sex role thus far have been hopelessly confounded in experiments with human subjects.

Sex differences in perceiving, whatever their basis, can be illustrated in research on differences in the style with which people perceive. This stylistic difference emerges in extremes of response to context. If a person perceives the world as highly differentiated, he tends to resist contextual influences and is said to be field independent; the person who perceives in an extremely diffuse style, the field-dependent individual, tends to be highly susceptible to contextual effects. Thus, field-independent people are superior in locating a simple visual figure (*e.g.*, a triangle) embedded in a complex pattern; similarly, field-independent subjects can better adjust a rod in a tilted frame to the true vertical when no other visual cues to verticality are present.

Both age and sex are found to be implicated in these differences in perceptual style. Specifically, field dependence declines with increasing age, as does the closely related susceptibility to optical illusions. In North American studies, female subjects tend to be more field dependent than are males, especially after puberty. Perhaps these results are distinctive of cultures in which females are at least implicitly trained to be passive and perceptually diffuse, and in which males are encouraged to assume an active, perceptually articulated stance. This hypothesis has received some support in studies of the parent–child interactions characteristic of the early years of the two types of subject.

Cultural influences. Beyond sex differences in perceiving that seem to be culturally imposed, there is evidence for more general cultural influences on perception. The burden of much research is to show that the type of physical environment people construct for themselves or choose to inhabit can influence their style of perceiving. There are African groups (*e.g.*, Zulu and San), for example, whose environments are virtually lacking in rectangular forms, by contrast with the carpentered, right-angled world of people in Western cultures. People in these African groups also make no use in their art work of two-dimensional representations of three-dimensional objects. Such differences in visual environments show up in tests of susceptibility to illusions. Zulu and San subjects are relatively resistant to those visual illusions that depend for their effectiveness on the subjects' treating the lines comprising the pictures as borders of three-dimensional, rectangular objects. Analogous effects with different classes of illusion have been shown for other peoples who live in a perceptually unique environment.

(W.N.D.)

Perception of movement

Most animals (including people) move in search of food that itself often moves; they move to avoid predators and to mate. Animals must perceive their own movements to balance themselves and to move effectively; without such perceptual functions the chances for survival would be sharply reduced.

VISUAL CUES TO MOVEMENT

The eye is by far the most effective organ for sensing movement. Some animals are especially sensitive to visual stimuli that move in specific ways. For instance, electrical patterns from the eye of a frog show that some elements in the organ respond only when the stimulus is about the size of a fly moving in the insect's range of speed. Generally the eyes of lower animals seem to respond selectively to what is of importance to survival. In these animals the eye's retina does much of the visual processing. This is an economical arrangement since the animal tends to respond only to essential stimuli, the brain having little to do but relay signals to the motor system. It is an inflexible mechanism, however; higher animals process visual information in more elaborate ways, the brain being more heavily involved. Thus, some cells in the visual area of the cat's brain respond only to moving stimuli, sets of movement-detector cells functioning specifically for each direction across the field of view. Features of human visual experience also suggest that movement detectors exist in the human brain.

Eye and brain

Each retina in most higher animals has a central (foveal) zone for detailed colour and pattern vision and a surrounding peripheral zone that effectively is sensitive only to the grosser features of the outer visual field. The peripheral retina is especially sensitive to movement (often a signal of danger), which induces a reflex turning of the eyes to project the image on the fovea and permit the moving object to be recognized.

Visual stability. Mechanisms have evolved that yield stable, clear visual input despite swaying and other blurring factors. In a reflex mechanism called optokinetic nystagmus, the eyes pursue a moving scene to keep the image stationary on the retina. When they can move no farther, they snap back and pursue the scene again in a to-and-fro alternation of slow pursuit and quick return. These eye movements are readily observed in people who are looking at a moving pattern of stripes or turning their heads, this response being inhibited only when something stationary is visually fixated.

Similar nystagmic movements are triggered by impulses arising in the inner ear when the head moves. These persist even when the eyes are closed and may be felt by pressing the eyelids lightly as one rotates the whole body.

Nystagmus; saccades

In a related stabilizing activity the eyes scan in quick jerks (saccades) with short fixations; *e.g.*, in reading. Normally the eyes cannot move steadily over a stationary scene but make a series of stationary images (like still photographs); visual function tends to be suppressed when there is saccadic blurring. Yet the eyes can follow a steadily moving object smoothly.

When one looks from one point to another, movements of the retinal image are the same as those produced by a moving scene on a stationary eye. It might be thought that the sensory structures found in the eye muscles would provide the cues for judging whether it is the eye or the scene that has moved. Yet we see the scene as stationary only when we move our eyes voluntarily and not when they are moved passively by the finger. This suggests that motor-nerve signals inform us whether our eyes are moving, rather than the sensory structures in the eye muscles. When the eye is moved by pushing it with the finger there is no normal motor discharge to inform the brain, and changes in retinal image are perceived as movement of the scene. Indeed, people with paralyzed eye muscles experience the scene as moving when they try to move their eyes. When the motor discharge thus generated is not accompanied by the expected image motion, the person falsely perceives the scene and the eye to be moving together.

Relative visual movement. A visual field containing familiar objects provides a stable framework against which relative motion may be judged. People often report that an isolated point of light in a dark room is moving when it is not; the experience is known as autokinetic movement. It was observed in 1799 by Alexander von Humboldt while he was watching a star through a telescope, and he attributed it to movement of the star itself. Not until about 60 years later was the effect shown to be subjective, apparently arising from instability in the sense of eye position without a visual frame of reference.

Autokinesis; induced movement

Similarly, if a small object is presented in a frame with nothing else in view, movement usually is attributed to the object even when only the frame moves. This induced movement effect reflects our tendency to use the larger surround as a stable frame of reference. Recall the illusion that your train is moving when it is really the moving train alongside that, seen through the window, is falsely accepted as the frame of reference.

People cannot perceive very slow movement; below a minimum speed (about that of the minute hand on a

watch) movements become imperceptible and can only be inferred (as in remembering the previous position of the hour hand). At high speeds, one perceives a blurred streak rather than a definite object in motion.

Movement aftereffect. When a parade is interrupted after some minutes, the pavement may seem to move in the opposite direction to the marchers who have passed. Phenomena similar to this movement aftereffect occur in other senses. For instance, after disembarking, a sailor feels the land to be rolling like a ship as the result of kinesthetic and vestibular aftereffects. The visual movement aftereffect probably arises when movement detectors in the brain that respond to the original direction of motion become fatigued, leaving predominant those detectors that respond to contrary movement.

Apparent movement. Motion-picture film is a strip of discrete, still pictures but produces the visual impression of continuous movement. Stationary light bulbs coming on one after the other over the theatre entrance also produce an impression of steady movement. In part, such effects of apparent movement (called the visual phi phenomenon) depend on persistence of vision: visual response outlasts a stimulus by a fraction of a second. When the interval between successive flashes of a stationary light is less than this visual-persistence time, the flicker will appear to fuse into a continuous light. The flicker frequency at which this occurs is called the perceiver's flicker-fusion frequency (or critical flicker frequency) and represents the temporal resolving power of his visual system at the time. Another process on which apparent movement depends is a tendency (called visual closure or phi) to fill in the spaces between adjacent visual objects. This means that the movement detectors of the visual system are triggered as effectively by a closely spaced pair of lights alternately going on and off as by a single light moving back and forth. It would seem that two aspects of visual function (flicker fusion and phi) make the motion-picture industry possible.

Flicker fusion; phi

Stroboscopic effect. When a rotating electric fan is illuminated by a flashing light source (called a stroboscope) so that a flash arrives whenever a fan blade passes a fixed position, the blades will seem to stand still. This is a useful way of observing fast-moving objects such as machinery or insect wings. If the flashes occur less frequently, the object will seem to move slowly in its actual direction; when the flashes arrive more frequently, the object will seem to move backward, as stagecoach wheels may do in the cinema.

Visual movement in depth. An object moving directly away from an observer provides fewer visual cues of movement than it would be moving across the field of view. However, changes in retinal-image size are produced that give a clue to its movement. Thus a stationary, but shrinking, luminous object in the dark is seen as if it were receding. Other clues to movement in depth are changes in the convergence angle of the two eyes, in the focussing mechanism, and in the haziness and brightness of the object.

NONVISUAL CUES

Auditory. Sound waves travel well in water, and fish are accordingly able to rely heavily on acoustic cues to detect moving objects. Land animals, although typically more visually oriented, also make some use of such cues, including changes in intensity (loudness) and small differences in the time at which the wave reaches each ear. Some animals (*e.g.*, rabbit, horse) have mobile external ears that track moving sound sources. Bats vocalize high-frequency waves and are able to detect objects by sonic reflection (a technique similar to sonar).

Kinesthetic. Kinesthesis here refers to experiences that arise during movement from sense organs in the membranes lining the joints and from the sense of effort in voluntary movement; receptors in muscles seem to have little role in the perception of bodily movements. Depending on speed of motion and the joint involved, blindfolded people can detect a passive joint movement as small as a quarter of a degree. People vary widely in the accuracy with which they can actively produce movement of a given extent; this ability also varies with direction of movement and the opposing friction, mass, and springiness.

Kinesthetic perception may persist for a limb that has been amputated, giving rise to a hallucinatory experience known as the phantom limb. The patient may experience vividly the "movement" of the absent part; a recent amputee may attempt to stand on his missing leg or to grasp with his missing hand.

Phantom limb

Vestibular system. Vestibular structures, enclosed in a fluid-filled cavity in the region of each inner ear, include the utricle, a small sac containing minute sensitive hairs associated with tiny sandlike granules called otoliths. The utricle functions as a linear accelerometer. When the head tilts relative to gravity or is accelerated, the relatively dense otoliths deflect the hair cells and nerve impulses are transmitted to the brain. At constant velocity the otoliths become stable, stimulation ceases, and a person must rely on other cues (*e.g.*, by observing the passing scene) to detect his motion.

Vestibular structures for each ear also include three fluid-filled semicircular canals, each in a different plane. Each canal has a swelling (ampulla) that contains the cupula, a cluster of sensitive hairs embedded in a jellylike mound. As the head moves in the plane of a given canal, motions of the fluid deflect the cupula to produce nerve impulses. These travel through the brainstem to other brain and spinal centres that mediate equilibrium or balance and that generate nystagmic eye movements.

Equilibrium

Taken together, the semicircular canals serve as a rotary accelerometer. If a person is rotated at constant velocity and then is suddenly stopped, the cupula is redeflected to give a feeling of rotation in the opposite direction; this also gives rise to dizziness and postrotational nystagmus. Dancers and skaters learn to overcome such effects by concentrating on some fixed, visible object; with their eyes closed they are as likely to fall as anyone.

Overstimulation of the vestibular system (*e.g.*, on a ship or airplane) may induce motion sickness. A person with vestibular function totally destroyed is not subject to motion sickness; but if the vestibular mechanism is impaired only on one side, each movement of the head can be nauseating. Such a patient takes a long time to compensate for this imbalance.

In outer space there is no gravitation to mediate feelings of up or down, although these still may arise from visual cues. The utricles and vestibular canals still respond to movements of the head, however, and serve orientation within the spacecraft. (I.P.H.)

Space perception

Human beings have been interested in the perception of objects in space at least since the question of how a picture or image arises in the human eye was first asked. It was popularly thought in ancient Greece that objects are seen because they emit what was imagined to be a continuous series of extremely thin "membranes" in their own image, which fall upon the eye and there merge into the picture that is perceived. Only after long experimental research was a more tenable conception reached, in which space was described in terms of three dimensions or planes: height (vertical plane), width (horizontal plane), and depth (sagittal plane). These planes all intersect at right angles, and their single axis of intersection is defined as being located within perceived three-dimensional space—that is, in the "eye" of the perceiving individual. Man does not ordinarily perceive a binocular space (a separate visual world from each eye), but sees a so-called Cyclopean space, as if the images from each eye fuse to produce a visual field akin to that of Cyclops, a one-eyed giant in Greek mythology. The horizontal, vertical, and sagittal planes divide space into various sectors: something is perceived as "above" or "below" (the horizontal plane), as "in front of" or "behind" (the vertical plane), or as "to the right" or "to the left" (of the sagittal plane).

GENERAL CONSIDERATIONS

The study of space perception developed rapidly in the second half of the 19th century and still more rapidly

during the 20th. Many modern psychologists who deal with perceptual function hold that the study of space perception is rapidly becoming a distinct branch of psychology in its own right. This special field of investigation within psychology concentrates on the factors contributing to the perceptual articulation (organization) of objects in space (*e.g.,* on cues to depth perception, movement, form, colour, and their interactions), or dwells on particularly interesting special problems such as that of amodal perception (*e.g.,* the question of how one perceives that there are six sides on a cube, even though only three of them can be directly visible at once). Another special problem is the effect of the so-called mirror-ego on perception. It is questioned, for example, whether, when one's skin is touched while he is looking at himself in front of a mirror, he experiences the touch as if it were in the mirror image or perceives it to be in contact with his own body as it actually appears in front of the mirror.

Perception and orientation

Space perception research also has tended to become a distinct specialty because of indisputable evidence that behaviour in this area is involved in orienting the individual to his environment. Specifically, man's orientation in space typically seems to reflect his strivings (*e.g.,* to seek food or to avoid injury). He would not be capable of effective orientation behaviour, however, unless the environmental information reaching him through the various sense organs were articulated (perceptually organized) in one way or another, permitting objects in space to be perceived at least approximately in a way that corresponds to their physical "reality" (*i.e.,* to their "actual" condition as stimuli). Such perception is called veridical perception, despite philosophical disputes about perceptual truth or verity (see EPISTEMOLOGY). Without some degree of veridicality of what is perceived to be physical space, one should be unable to efficiently seek food, flee from his enemies, or take part in social interaction. This means that, in service to one's veridical perception (direct perception of stimuli as they exist), one must also experience changing stimuli as if they were stable (invariant); that is, one experiences so-called perceptual invariances. Even though the sensory image of an approaching tiger grows larger, for example, one tends to perceive that the animal's size remains unchanged. One must be capable of perceiving objects in his environment as having relatively constant characteristics (as to size, colour, and so on) despite considerable variations in stimulus conditions.

Primary gravitational effects. Not all perception of space, however, is veridical; *i.e.,* it may fail to correspond to reality. A large number of perceptual effects have been demonstrated in which space as perceived deviates from reality in some systematic way. These are cases of nonveridical perception; for example, it has been demonstrated experimentally that the three basic spatial planes (horizontal, vertical, and sagittal) dominate the ability of the individual to localize visual objects in nearby space. Carefully controlled experiments show that objects tend to be perceived as lying closer to these basic dimensions or planes than they really are. Part of the explanation of these perceptual discrepancies in visual experience may lie in the force of gravity, which undeniably affects the individual from the moment of conception onward, pulling a man's viscera and muscles "down" toward the centre of the earth. The various perceptual phenomena indicative of nonveridical perception nevertheless do not generate chaos in one's perception of space, as might theoretically be expected; indeed, most often they serve to simplify and to articulate the perceived characteristics of surrounding space. If all of the mass of sensory information available at a given moment were perceived veridically, the flood of data might so confuse the perceiver that he would be unable to orient himself appropriately. In other words, some degree of selectivity in perception in general appears to be demanded for the survival of the individual. Ideally, information about the environment is perceived only as it is relevant to the goals, needs, or physiological state of the individual at a given moment.

Selective perception

Visual factors in space perception. On casual consideration, it might be concluded that the perception of space is based exclusively on vision. After closer study, however, this so-called visual space is found to be supplemented perceptually by what are clearly identifiable as cues from auditory (sense of hearing), kinesthetic (sense of bodily movement), olfactory (sense of smell), and gustatory (sense of taste) experience. In addition to these, a number of other spatial cues, such as vestibular stimuli (sense of balance) and other modes for sensing body orientation, must be taken into account. These various (*e.g.,* visual, olfactory) "spaces" are not found to be perceptually independent of one another; powerful experimental evidence shows them to interact in producing unified perceptual experiences.

It is nevertheless undeniable that the individual ordinarily receives the largest part of his information about the environment through the sense of sight; balance or equilibrium (vestibular sense) seems to rank next in importance. It has been shown that in a state of total darkness the orientation of an individual in space is mainly dependent on sensory data deriving from vestibular stimuli. The dominance of visual stimuli in the perception of space becomes particularly understandable when it is noted that vision is a so-called distance sense; it is capable of gathering information from extremely distant points in the environment, reaching out to the stars themselves. Hearing is also considered a distance sense, as is smell, though the space they encompass is considerably more restricted than that of vision. All the other senses, such as touch and taste, are usually considered to be proximal (proximate or direct) senses—*i.e.,* conveying information about the space in direct contact with the perceiving individual's body. Strictly speaking, the latter senses may, on occasion, function over distances. The smell of rotten eggs (produced by hydrogen sulfide gas) several yards away may be associated with a taste experience if the gas dissolves in the perceiver's saliva.

The eye has often been compared to a camera; this is, of course, a very rough comparison, but it is possible to think of the retina (the back surface of the inside of the eye) as if it were the film in a camera and of the lens within the eye as analogous to the single lens of the camera (see SENSORY RECEPTION: *Human vision*). Just as in a portrait photographer's camera, the picture (image) that is projected from the environment onto the retina is upside down. While it might be expected that the perceiver therefore would experience space as turned upside down, normally this does not happen. Some sort of perceptual mechanisms within the individual enable him to see his environment right side up; the exact nature of these mechanisms is under study, yet remains poorly understood. But the process of perception seems to involve at least two inversions: one (optical) inversion of the image on the retina; another (perceptual) associated with nerve impulses in the visual tissues of the brain.

Experiments have been carried out with people who were wearing special optical devices (*e.g.,* as spectacles) that reverse the right–left or up–down dimensions of images as normally received at the surface of the retina. At first the person becomes disoriented, but when he wears such a device all the time, he gradually learns to cope with space correctly while the special glasses are on. Indeed, he reorients himself to his environment to the point that he visually begins to perceive objects as if they were right side up again. This is evidence that the individual can become visually adapted in his perception of space to a new set of stimulus cues that deviate considerably from those previously learned. The process changes direction when the optical distorting device is removed; at first the basic visual dimensions appear reversed to the subject who has adapted to the spectacles. Within a short time, however, a new adaptation occurs; the subject quickly reorients himself to the earlier, well-learned, normal visual cues and perceives his environment as normal once more.

PERCEPTION OF DEPTH AND DISTANCE

The perception of depth and distance depends on information transmitted through various sense organs. Sensory cues indicate the distance at which objects in the environment are located from the perceiving individual and from each other. Such sense modalities as seeing and hearing are known to transmit depth and distance cues and are

in substantial degree independent of one another. Each modality by itself can produce consistent perception of the distances of objects. Ordinarily, however, the individual relies on the collaboration of all his senses, (so-called intermodal perception).

Gross tactual–kinesthetic cues. In his perception of the distances of objects located in nearby space, one depends on his tactile (tactual or touch) sense. In tactile experience, however, kinesthesic experience (sensations of muscle movements and of movements of the sense-organ surfaces) normally is so closely associated that investigators lump the two as tactual–kinesthetic cues.

Man's own body as a perceptual standard

Tactual–kinesthetic sensations enable one to differentiate his own body from the surrounding environment. The body of the individual seems to function as a perceptual frame of reference—that is, as a standard against which the distances of objects are gauged. One's perception of his own body may vary from time to time, and so its role as a perceptual standard is not always consistent. It has been found that the way in which the environment is perceived also affects one's perception of his own body. Thus, for example, it has been shown experimentally that one experiences his own arm as growing longer when he is using it to point at some object off in the environment.

Cues from the eye muscles. When one looks at an object at a distance, the effort arouses activity in two eye-muscle systems called the ciliary muscles and the rectus muscles. The ciliary effect is called accommodation (focussing the lens for near or far vision), and the rectus effect is called convergence (moving the entire eyeball). Each of these muscle systems contracts as a perceived object approaches. The effect of accommodation in this case is to make the jellylike lens more convex, while the rectus muscles rotate the eyes to converge on the object as it comes nearer. One's experience of these muscle contractions provides him with cues to the distance of objects.

Beyond the cues of accommodation and convergence, the size of the retinal image also serves to indicate how far one is from an object. The larger the image on the retina, the closer one judges the object to be. Some workers theorize that these sensory cues activate inherited tendencies immediately to perceive such sensory attributes as size without the need for learning (the so-called nativistic theory). Modern efforts to study these cues have been especially directed toward physiological changes in the body that may be related to depth perception; whether one's perception of depth is totally inborn (nativistic), and thus independent of learning, however, remains most controversial.

Accommodations and convergence provide reliable cues when the perceived object is at a distance of less than about 30 feet (nine metres) and when it is perceived binocularly (with both eyes at once).

Visual cues. Perhaps the most important perceptual cues of distance and depth depend on so-called binocular disparity; that is, since the eyes are imbedded at different points in the skull, they receive slightly different (disparate) images of any given object. The two retinal images of the same object seem to be combined perceptually in the brain into one three-dimensional experience. The degree of disparity between the two retinal images, so-called binocular parallax, depends on the difference between the angles at which an object is fixed by the right eye and by the left eye. Thus, in looking at the indicator needle on a pressure gauge, for example, one will make slightly different readings because of parallax if he uses first the left eye alone and then the right eye. The greater the parallax difference between the two retinal images, the closer the object is perceived to be.

Binocular disparity

The phenomenon of binocular disparity functions primarily in near space because with objects at considerable distances from the viewer the angular difference between the two retinal images diminishes. Visual disparity nevertheless can be used in estimating greater distances by simply using optical devices that magnify the parallax distance separately for each eye. Such devices include artillery range-finding devices and old-fashioned, three-dimensional picture viewers called stereoscopes.

In what is called visual movement parallax, distance cues are obtained from retinal changes that depend on the interposition of objects in space. Thus, when the individual moves his head either from side to side, or forward and backward, the retinal image of a nearby tree moves more, while that of a distant tree moves less. Unlike binocular disparity, which functions only in binocular vision, movement parallax is especially important for judging distance when only one eye is used (monocular vision).

Another group of visual images called perspective projections provide perceptual cues that are independent of whether there is monocularity or binocularity of vision. Although estimates of distance based on such perspective effects as the apparent distant fusing of railroad tracks in a single point are incompletely understood, they are often considered to depend heavily on learning. Such phenomena illustrate the tendency of the individual to integrate perceptions into consistent and invariant wholes. Experiences of perspective may be generated by putting appropriate lines in an oil painting (linear perspective), by gradations in the tint of the paint (colour perspective), and by viewing the surface of the earth from an airplane (aerial perspective).

Still another group of visual cues of depth and distance consists of apparent differences in object brightness. In experimental studies it is found that the brighter an object appears, the closer it seems to be. Thus, a white card against a dark background seems to recede or to move forward as the level of illumination on the card is experimentally varied. Similar effects can be induced by changing the colour (hue) of an object—*e.g.*, from bright red to dark red.

Auditory cues. Auditory cues for depth perception include sound intensity (loudness), auditory pitch, and the time lapse between visual perception and auditory perception (for example, one hears a distant cannon after seeing the flash and smoke of the explosion).

Hearing in depth

Changes in pitch function as depth cues because when a moving object (*e.g.*, an automobile) is emitting sound waves (*e.g.*, from its horn), the pitch of the sound seems to rise when the object is approaching the perceiver; to fall when it is moving away.

INTERRELATIONS AMONG THE SENSES

The development of human ability to perceive space normally depends on the interaction of the senses (modalities) of sight and touch. Toward the end of its first year, a child eagerly begins to touch and to explore objects with his hands. Compared with his visual apparatus, which begins to function more efficiently at a later time, the child's sense of touch at this stage of development is delicately and effectively sensitive. By touching objects that are placed at various heights and distances from him, the child learns excellent tactual ways of perceptually evaluating what must be a highly ambiguous world as transmitted by his relatively immature visual apparatus.

The part played by other modalities (*e.g.*, hearing) does not appear to be as fundamental in perceptual learning among young children. By contrast, people are most seriously hampered in learning a detailed, well-articulated perceptual appreciation of space without vision or touch; most people (including those who are blind) may find it difficult to imagine how impoverished an understanding of space one would have if there were only auditory cues. It is well known that people with full sensory endowment learn to locate the sources of sounds by consulting both their visual and tactual experiences. There are subtler forms of sensory interaction as well; for example, one is more accurate in turning a pointer toward a distant source of sound (*e.g.*, a buzzer) if the pointer is simultaneously illuminated whenever there is buzzing. Such anchoring of one sense to another as a basis for spatial perception is vividly illustrated by experiments with laboratory animals.

Experiments with animals

Young monkeys have been taught to discriminate, on the basis of tactual perception alone, differently shaped objects (balls, cubes, cones, and cylinders, all about the size of a matchbox). That is, each animal was given two of these objects at a time and was allowed to feel them with its hands without seeing them. The animal's choice of one specific shape (say, a cube) and its rejection of

another (say, a cone) were rewarded with bits of food. When this selective tactual response had been learned, the animal was given the same tasks of selection, this time to be performed visually (by looking at pictures of the same objects). Under these conditions the animal often was able to discriminate the visually presented figures correctly. This phenomenon (called cross-modal learning) may be explained as a transfer effect based on earlier visual-tactual learning.

Success in orientation, in moving about effectively and without accident in everyday pursuits, is most probable when environmental information at any given time is available through as many senses as possible. When the range of sensory stimuli that forms the usual basis for the experience of perceptual space is experimentally reduced, one's success in orientation is impaired. When visual cues are sharply reduced for sighted people, they complain of disorientation, akin to groping around in a diffuse, undifferentiated fog. Indeed, injuries are all too common even when one seeks in the dark to negotiate such familiar territory as his own bedroom.

Apparently, by learning about systematic relationships that exist among a number of simultaneously available stimuli, people can perceive distances more or less correctly. The retinal image of an object 100 yards (91 metres) away is so much smaller than the image of the same object at a distance of 10 yards (9 metres) that (if they relied solely on the sizes of retinal images) people might perceive moving objects to shrink or grow rather than to recede or approach.

Experiments have shown that the distance (in depth) between selected objects in photographs is most accurately estimated when the objects have been filmed in a richly organized environment—*e.g.,* many people standing at different distances from the camera. On the other hand, it is most difficult indeed to reliably perceive even the relative depth of two vertical rods when they are presented against a background in which other cues have been reduced or eliminated.

Factors of constancy. In general, the considerable range of sensory information offered by isolated local stimuli is typically controlled, sifted, and corrected by the perceiver. The specific nature of these "corrections to conform to reality" generally depends on the unique combination of simultaneously presented stimuli at any given moment. In this way, spatial perception tends to ensure that a person experiences the continually changing circumstances of the environment in which he moves and lives with some degree of stability or constancy.

This "realistic" perception, based on an awareness of the apparently real, physical world, serves as the basis for practical adjustment and survival and appears clearly in a group of perceptual experiences of so-called object constancy. In the experience of size constancy, for example, within a radius of a few hundred yards from the observer, the size of objects is perceived to remain roughly constant no matter where they are. Constancy of form means that an object is perceived to retain its fixed characteristic shape regardless of variation in the angle at which it is observed; for example, a pencil seen end-on only shows a small circular profile but is still perceived as a pencil. Colour constancy clearly illustrates the way in which the brightness and hue experienced over the surface of an object is determined by direct comparison with other objects; a lump of coal still is perceived as black whether the Sun is shining brightly or whether there is a dull, leaden sky.

Path recognition: navigation in space. Different animal species are equipped in various ways for the recognition of their path of movement. Earthbound species of animals often use olfactory (smell) signals in recognizing paths of varying distance; this is encountered both among social insects (*e.g.,* ants) and among many mammals (the dog sniffing from tree to tree). Certain insect larvae can retrace their path of movement by using as guides extremely fine webs or filaments spun during their advance. Many species seem to navigate by the Sun, the angle between the rays of the Sun and the organism's own path being held constant. The ability of migratory birds to orient themselves by stars in twilight or at night is most impressive (see BEHAVIOUR, ANIMAL: *Migratory behaviour*). In addition to cues offered by the stars and other heavenly bodies, such stimuli as the effects of gravity, temperature changes, and direct visual observation of landmarks such as rivers all seem to contribute to path recognition among animals.

Use of smell to judge distance

With the burgeoning of aviation and space technology, there has been a growing research effort to increase understanding of the sensory basis for human navigation in space. Reliable perception of the vertical and horizontal dimensions and preservation of perceptual constancy for these dimensions during flight are based on the parallel activity of vision and the vestibular sense (balance, or equilibrium). Even in flying small aircraft it has been shown that the pilot becomes disoriented unless he preserves visual control of the horizontal dimension; he cannot safely rely on the information conveyed by his sense of balance to inform him that a wing tip, for example, is dipping dangerously low. While the sense of balance ordinarily depends on the effect of gravity, the movement of an airplane produces additional (so-called centrifugal and centripetal) forces, particularly as the plane tilts or changes direction, which easily mislead a person's vestibular (balance) receptors. For this reason, in high-altitude flying the horizontal line of the surface of the Earth is simulated for the pilot by an optical display that works on the same principles as does a television screen.

Even greater demands on the human senses of vision and balance are made in space flights, since a person is effectively weightless in outer space. At one laboratory maintained by the U.S. Navy, an enormous, very slowly rotating cylindrical chamber has been constructed. Variations in the perceptual sensitivity of people remaining for variable periods (even days at a time) in this simulated "outer space" environment are studied in an effort to anticipate the short-term and long-term effects of interplanetary flight.

Social and interpersonal aspects of space perception. Many animal species that use nests, lairs, or dens and care for their young act to defend a specifically delimited territory against interlopers. This process is clearly observable in birds and among seals during breeding season. Apparently territorial behaviour depends on a rather precise perception of space, since the animal ceases its defensive manoeuvres when an interloper passes out of the territory by moving across the "border." The social distances maintained by primates (*e.g.,* human beings and apes) are theorized to be the result of territorial groupings. Modern architecture is held to be dominated by an inborn human tendency to divide groups into small, separate family territories (just as birds do), the result being such structures as apartment houses. Geographical distance may also be maintained to separate individuals who belong to different social classes or races, as is observed in the ethnic neighbourhoods or ghettos of large cities.

Territorial distance

Theories of space perception. An early theory put forth by the Anglican bishop George Berkeley at the beginning of the 18th century was that the third dimension (depth) cannot be directly perceived in a visual way since the retinal image of any object is two-dimensional, as in a painting. He held that the ability to have visual experiences of depth is not inborn (nativistic) but can only result from logical deduction based on empirical learning through the use of other senses (*e.g.,* touch).

Although modern research fails to verify the importance Berkeley placed on reason in perception, contemporary theories still include both nativistic and empiricistic considerations. According to empiricistic theorists, spatial perception develops primarily as the result of learning; indeed, during the early stages of development, the individual gradually does seem to learn something about the significance of observable (empirical) spatial cues. It is just as difficult, however, to deny the nativistic concept that space perception is based on the innate (hereditary) structure of sense receptors and the nervous system; for example, the ability to detect differences in the size of two-dimensional figures seems to be inborn (that is, the differences seem to be immediately recognizable without training or learning).

(K.V.J.v.F./E.J.J./P.Y.K.)

Time perception

The human experience of change is complex. One primary element clearly is that of a succession of events; but distinguishable events are separated by more or less lengthy intervals that are called durations. Sequence and duration are fundamental aspects of what is perceived in change.

Manifestly, duration is relative to the events people isolate in the sequences through which they live: the duration of a kiss, of a meal, of a trip. A given interval always can be subdivided into a sequential chain delimiting briefer durations, as with the regular units that provide empirical measures of time: the second, the day, the year.

Indeed, human experience is not simply that of one single series of events, but of a plurality of overlapping changes. The duration of a radio program, for example, can combine with that of a breakfast, both being inserted within the longer period of an ocean voyage.

Early philosophical views of time

Man seems to be unable to live without some concept of time. Ancient philosophies sought to relate the concept of time to some objective reality to which it would correspond. René Descartes (1596–1650) inaugurated a critical era of philosophy by stressing the ancient problem of the origin of ideas, including the idea of time. Immanuel Kant (1724–1804), providing a radical answer to the epistemological problem of time, wrote that we do not appreciate time objectively as a physical thing; that it is simply a pure form of sensible intuition. Other philosophers of the 18th and 19th centuries sought to explain the notion of time as arising from association and memory of successive perceptions.

A move to empirical psychology emerged with the growth of research on the introspective data of experience. From about mid-19th century, under the influence of the psychophysical notions of Gustav Theodor Fechner, psychologists conducted experiments to study the relationship between time as perceived and time as measured in physics. Their work with adults gradually spread to the study of children and of animals. The psychologists then broadened their investigations of time to cover all forms of adaptation to sequence and duration.

SEQUENTIAL ACTIVITIES

Adaptation to successive events. One may respond to stimulation in an immediate way (as in unconditioned reflex action) without taking the element of time into account. Stimulation, however, can also signal an event to follow; then it has meaning only as part of the sequence of which it is the first term: bell announcing dinner, a road sign, or an approaching danger. People react to such stimuli with anticipatory behaviour that is adapted to a stimulus or action that has not yet occurred. The principles that govern such time-binding adaptation are none other than those of conditioning. One event becomes conditioned as the signal for another stimulus that is to be sought or avoided.

Time-binding adaptation as conditioning

The bottle-fed infant who initially reacts to the nipple on his lips with a simple sucking reflex is gradually conditioned to stop crying when he sees the bottle (the signal for feeding). Later he may learn to react to even more secondary signals that announce the arrival of the bottle; *e.g.,* being lifted from the crib or hearing the sounds of his mother warming the milk in the kitchen. His behaviour has come to incorporate the temporal dimension of the events.

According to the principles of instrumental conditioning, one stimulus becomes the signal for an ensuing event only if the second stimulus elicits an adaptive reaction (consummatory or aversive) and only if the order of the sequence is repeated. Conditioning tends to be established most rapidly when the interval between the signal (conditioned stimulus) and the unconditioned stimulus is quite brief. Ivan P. Pavlov estimated that the optimum interval for such a sequence was 0.5 second, which corresponds approximately to the intervals characteristic of sequences that are most accurately discriminable perceptually (see below).

Aside from adapting the individual to the order of a sequence, conditioning also adapts to the duration between signal and immediately effective stimulus. Response to signal tends to occur after about the same interval that separated the two stimuli during conditioning. Thus, an animal may be trained to delay a response for some time after the signal (delayed conditioning).

This form of adaptation is most pervasive in human behaviour, permitting people to anticipate sequences of events in their environment so that they can prepare to cope appropriately with what is yet to happen.

Adaptation to periodic change. In 1912 one of Pavlov's students (I.P. Feokritova) demonstrated that a dog accustomed to being fed every 30 minutes would begin to drool toward the end of each half-hour period. It was clear evidence of conditioning to time; the between-feedings interval itself served as a conditioned stimulus.

That discovery underscores the ever-present periodicity of daily living, especially on the biological level: rhythms of activity and sleep, rhythms of eating and lovemaking. As conditioning intervenes, anticipatory experiences of hunger, fatigue, or arousal serve our adaptation to ecological demands.

Circadian rhythms in metabolic activity

Allowance should also be made for the daily, or circadian, rhythms in metabolic activity (*e.g.,* daily cycles of temperature change). There is evidence that these fundamental biological functions can synchronize with the rhythmic phases of environmental (exogenous) change. Thus within a few days after a factory worker has been assigned to the night shift, highs and lows of his daily fluctuations of temperature will be inversed. The rhythmic changes in body temperature persists, nevertheless, suggesting an innate (endogenous) basis for circadian phenomena. Such a hypothesis would mean that the gradual establishment of human circadian rhythms of sleep or temperature results from maturation of the nervous system rather than from conditioning in the strict sense. Experiments begun in 1962, in which men lived in caves or other enclosures for months deprived of temporal cues from the environment, also demonstrated the enduring nature of rhythms in body temperature and in sleep–wakefulness. The rhythmic periods, however, sometimes expanded, the subject beginning to live on an approximately two-day cycle without being aware of it.

Through conditioning to time and by way of circadian rhythms, human physiology provides a kind of biological clock that offers points of reference for temporal orientation.

PERCEPTION OF SEQUENCE AND DURATION

The psychological present. To perceive is to become aware of stimulation. Awareness of sequence or duration may, at first glance, seem inconsistent with the definition of perceiving. In a mathematical sense, certainly, the present is only a point along the continuum of becoming, an instant when future is transformed into past. Nevertheless, there is indeed a more prolonged psychological present, a brief period during which successive events seem to form a perceptual unity and can be apprehended without calling on memory. There is a perceptual field for time just as there is a visual field. The rate or speed of a sequence determines the limits of the time field.

When a metronome tics two or three times a second, one perceives an integral sequence, becoming aware of a rhythmic auditory series characterized by a perceptually distinct frequency. When the ticks come less often, however—at intervals of three seconds, say—the frequency or sequence no longer is perceived. Each physically discrete sound impulse remains an isolated perceptual event; each tick is no longer perceived as belonging to the same temporal field as the one that follows. Similar effects can be achieved by playing a recording of music or speech at a very slow rate. Music or spoken sentences are recognizable only when their elements (melody, rhythmic patterns, phrase) are presented at an optimal speed that permits significant perceptual unity; that is, only when they belong to the relative simultaneity of the psychological present.

Relation of perceived time to number of elements presented

The perceived field of time also depends on the number of stimulus elements presented. When a clock strikes three or four times, one knows without counting that it is three or four o'clock. At noon one must count; the first chimes no longer belong to the psychological present that includes the last. Most people also can repeat a series of letters or

numbers they hear, so long as there are no more than seven or eight elements. This ability varies with the degree of perceptual (*e.g.,* semantic) organization among the elements. While most adults can apprehend only about eight letters, they can grasp and repeat without fault sentences of 20 to 25 syllables (see also ATTENTION).

Perception of sequence. A series of physically discrete stimuli that impinge too rapidly on a sensory structure (*e.g.,* flashes of light on the retina) may produce perceptual fusion; the flashes will be indiscriminable and will appear to be uninterrupted light. The experience of fusion yields to one of discontinuity over distinctive critical ranges of frequency for some of the senses: visual flicker appears under prescribed experimental conditions at about 60 flashes per second, auditory flutter at about 1,000 interruptions per second, and tactual vibration at about 4,000 pulses per second. These values depend on differences in the persistence of the receptor systems (*e.g.,* how long an image is seen after removal of the stimulus).

The question of perceiving sequence hardly has meaning for the senses of taste and smell. Hearing appears to be particularly adapted to temporal perception, since the pattern of auditory excitement shows little inertial lag, closely following the physical duration of successive stimuli. Tactual function can give comparable results, but hearing has the practical superiority in everyday experience of reception at a distance.

When two heterogeneous stimuli (*e.g.,* a flash and a click) are successively presented, the critical threshold for passing from perceived simultaneity to an awareness of succession is found for intervals that vary between 0.02 to 0.1 second, depending on the training of the subjects. The maximum interval for perceiving sequence is more difficult to measure. The minimum time intervals are largely determined by the immediate physiological conditions of direct perceiving, while the maximum intervals are obscured by the effects of other cognitive activities. Determining when direct perception ends and when memory takes over is difficult.

At any rate, awareness of unitary sequence ceases for pairs of auditory or visual stimuli when the interval between them increases to approximately two seconds. For perceptually organized stimuli (as in a rhythm, a melody, or a phrase) the interval may reach five seconds, as indicated by one's ability to reproduce the pattern.

Between the upper and lower limits there are optimal values that seem most likely to produce perception of sequence. In the simple case of two homogeneous stimuli the optimum interval seems to be about 0.6 to 0.8 second. This is inferred from a series of clues: the same interval defines the tempo most frequently adopted in spontaneous motor activity (*e.g.,* tapping, walking) and corresponds to the heart rate. It is the interval that is most precisely reproduced by subjects in experiments; shorter intervals tend to be overestimated and longer ones underestimated. Stimuli repeated at that rate are subjectively judged to proceed most comfortably, without appearing to rush each other as in faster tempos and with no tendency to be separately perceived as at slower frequencies.

Full and empty duration

Perceived duration. Duration, the interval between two successive events, may be distinguished as full or empty (filled or unfilled) in terms of the sensory stimulation that intervenes. An empty interval is bounded by two perceptually discrete stimuli (*e.g.,* two clicks in succession); a duration is full when there is continous stimulation, being delimited by an onset and cessation (*e.g.,* a light stays on throughout the interval). To experience an empty duration is to perceive sequence, while full duration corresponds to the temporal length of a stimulus.

Human subjects need a minimum of about 0.1 second of visual experience or about .01 to .02 second of auditory experience to perceive duration; any shorter experiences are called instantaneous. Direct, unitary perception of duration occurs up to a maximum period of approximately 1.5 to 2 seconds from the beginning to the end of a continuous sensory stimulus.

This roughly two-second maximum for directly perceived duration seems to have a biological basis and can be considered the upper temporal limit of some sort of integrated neural mechanism. The immediate physiological process triggered by a stimulus endures beyond the period of stimulation, and may be measured as the duration of electrical impulses (*i.e.,* in the optic nerve) evoked by simple stimulation. This initial activity appears to be integrated subjectively into a cognitive unit that embraces the rapidly ensuing perceptual processes as well. The optimum range of 0.6 to 0.8 second noted earlier seems to represent the typical duration of this integrating mechanism, as inferred from studies of sensory physiology and from reaction-time experiments.

At any rate, only within these limits can the quality and precision of direct human perception (as opposed to estimation and recall) of duration be studied. Such perception can be absolute or relative. Absolute perception corresponds to estimates expressed in subjectively qualitative terms as long or short. In making such estimates, people can discriminate four to five different durations between 0.1 and 1.0 second and six to seven between 0.5 and 5.0 seconds. In studies of relative perception, subjects attempt to reproduce intervals that are presented, or are told either to produce durations of specified length or to compare two successively presented durations. These tasks, especially comparison, give rise to constant time-order errors; that is, errors in estimation that depend on which interval is presented first.

Experimentally, the perception of empty duration is found to vary with the sense that marks the limits. With duration constant, interval estimates tend to be greater (1) when the limits are visual rather than auditory or tactual, (2) when they are of low intensity, or (3) when auditory limits are higher pitched. If the unfilled limits are defined by successive stimuli from different places, duration appears longer when the distance between the two sources is greater; this is called the S effect or kappa effect. The reverse is the tau effect, in which the distance is perceived as being wider when the interval between successive stimuli is longer.

Kappa and tau effects

The perception of filled duration also varies with the stimulus. Holding the interval constant, interrupted stimulation (*e.g.,* several successive clicks) appears to last longer than does a continuous stimulus; and auditory stimuli appear to last longer than visual. Filled durations seem longer as stimulus intensity (*e.g.,* loudness) or auditory pitch rises.

One interval can be perceived as longer or shorter than the next when the difference is about 7 to 10 percent (both full and empty durations). This relative difference threshold is lowered by practice. Such studies also reveal that apparent duration remains proportional to the objectively measured length of the interval.

Estimating duration. When an interval lasts more than a few seconds, it no longer is directly perceivable as a whole, but its length can be estimated on the basis of memory function. Since common experience shows how imprecise these estimates are, people generally calculate time from such indicators as the position of the sun or with clocks and watches. Duration then is inferred rather than perceived.

Estimates, however, often are made, including those of absolute duration in which an activity is appreciated as brief or prolonged. Lacking a watch, one may make crude estimates based on such quantitative aspects of activity as distance travelled, number of dishes washed, or number of pages read. Or one may estimate directly as in subjectively counting seconds.

Several important factors influence the subjects estimation of time:

Type of activity. The more often a task is broken up or interrupted, the longer it seems to take. As a corollary, a period of doing nothing appears longer than an equally long period when one is doing something. Similarly, relatively passive activities appear longer than do those requiring active participation; *e.g.,* time passes faster for the student who is taking notes than for one who passively listens.

Level of motivation. The more one is motivated by a given task, the shorter it appears to last. Clearly, motivation and the type of activity pursued are interdependent

factors. Lack of motivation tends to interrupt attention to a task; a task in which perceptual focus frequently shifts rarely corresponds to one for which there is strong motivation. The more one notices change during an interval, the longer it is judged to be. More generally, it may be said that time has subjective duration only when one notices it; *e.g.,* in awaiting the arrival of a friend (as opposed to the actual meeting) or in hoping to finish a task (in contrast to working at it).

Change as a factor in judging time

Personality traits. Although inadequacies in quantifying personality traits and difficulties in studying estimates of time spans exceeding a few seconds have hampered scientific study, simple observation reveals marked individual differences in the ability to estimate time. Sex differences have not been reliably established, but the influence of age is well known. Experimental data indicate that children use the same criteria as do adults, but give more variable estimates of duration. One reason for this seems to be that they are less able to compensate for differences in the nature of a task or in personal motivation; also they are inexperienced in making inferences based on the volume of work they have accomplished. Elderly people tend to find time shorter, probably since they are likely to notice long-accustomed changes less frequently.

Children are as accurate as adults in reproducing various series of metronome clicks that last about two seconds or less. But estimates of longer intervals require processes for organizing experience that develop only with age, and very young children seem to depend only on limited criteria: "It lasts because it's longer; because there's more of it; because it goes faster." According to Jean Piaget, estimates based on more or less explicit comparison with standard units of duration imply concrete cognitive operations that are developed only after about the age of seven or eight. Adolescents typically construct more sophisticated notions of time abstracted from such concrete experimental data.

The influence of body temperature

Physiological effects: drugs. The precision with which time is perceived has not been found to be related to heart rate or to electroencephalographic data. It has been shown, however, that perception of time as in clapping or counting accelerates or decelerates with the rise and fall of body temperature. The precise metabolic basis for such temperature effects awaits further study.

Ethical considerations sharply limit the dosage level of drugs employed for experiments on human beings. Understanding of the interactions between drug effects and personality traits in studies of time estimation is, therefore, quite incomplete. Within the dosage ranges investigated, however, stimulating drugs (*e.g.,* thyroxine, caffeine, amphetamines) produce overestimates of duration, while depressants and anesthetics (*e.g.,* barbiturates, nitrous oxide) promote underestimates. Under the influence of hallucinogens (*e.g.,* marijuana, mescaline, LSD), subjects tend to estimate absolute duration as very long. In addition, a marijuana user may underestimate the speed of a motor vehicle, increasing the chances of accident.

Sensory deprivation and hypnosis. Relatively complete sensory deprivation (such as may be experienced, for example, by persons undergoing prolonged stays in experimental isolation chambers) compresses the experience of time to the point that short or long intervals (from about a minute to a day) seem to pass about twice as fast as usual. Time spent under these unpleasant conditions paradoxically seems shorter than normal time. Thus, the 58 objective days of a subject's first stay in a cave were underestimated as 33 days.

Under hypnosis, durations ordinarily are estimated at least as precisely as ever. Time distortion, however, can be readily induced among hypnotized subjects by simple suggestion. Such a subject, for example, may be exposed to two clicks that delimit an objective, 10-second interval but be told that it lasts 10 minutes. On being asked to count objects for 10 minutes, he may report having counted several hundreds without difficulty over what the experimenter's stopwatch shows to have been 10 seconds.

(P.F.)

Illusions and hallucinations

Illusion and hallucination are terms formulated to describe subjective (perceptual) experiences that contradict objective "reality" as it is defined by general agreement among people. Such experiences are not necessarily signs of psychiatric disturbances; they are or have been regularly and consistently reported by virtually everyone. Generally, illusions comprise misinterpretations of "real" sensory stimuli (*e.g.,* the child who perceives tree branches at night as if they are hobgoblins). Hallucinations are experiences that seem to originate when no external source of such stimulation appears (*e.g.,* one's name is called by a voice that no one else seems to hear).

THE NATURE OF ILLUSIONS

Illusions, then, are special perceptual experiences in which information arising from "real" external stimuli leads to an incorrect perception, or false impression, of the object or event from which the stimulation comes.

Causes of illusions

Some of these false impressions may arise from factors beyond an individual's control (such as the characteristic behaviour of light waves that makes a pencil in a glass of water seem bent), from inadequate information (as under conditions of poor illumination), or from the functional and structural characteristics of the sensory apparatus (*e.g.,* distortions in the shape of the lens in the eye). Such visual illusions are experienced by every sighted person.

Another group of illusions results from misinterpretations one makes of seemingly adequate sensory cues. In such illusions, sensory impressions seem to contradict the "facts of reality" or fail to report their "true" character. (For more profound philosophical considerations, see EPISTEMOLOGY.) In these instances the perceiver seems to be making an error in processing sensory information. The error appears to arise within the central nervous system (brain and spinal cord); this may result from competing sensory information, psychologically meaningful distorting influences, or previous expectations (mental set). The driver who sees his own headlights reflected in the window of a store, for example, may experience the illusion that another vehicle is coming toward him even though he knows there is no road there.

TYPES OF ILLUSORY EXPERIENCE

Stimulus-distortion illusions. This type of illusory sense perception, common to mankind, arises when the environment changes or warps the stimulus energy on the way to the person, who efficiently receives it in its distorted pattern (as in the case of the "bent" pencil referred to above).

Auditory illusions. A common phenomenon is the auditory impression that a blowing automobile horn changes its pitch as it passes an observer on a highway. This is known as the Doppler effect, for C. Doppler, an Austrian physicist, who in 1842 noted that the pitch of a bell or whistle on a passing railroad train is heard to drop when the train and the perceiver are moving away from each other and to grow higher when they are approaching each other. The sound heard is also affected by such factors as a wind blowing toward or away from the person.

Another auditory illusion was described in 1928 by P.T. Young, an American psychologist, who tested the process of sound localization (the direction from which sound seems to come). He constructed a pseudophone, an instrument made of two ear trumpets, one leading from the right side of the head to the left ear and the other vice versa. This created the illusory impression of reversed localization of sound. While walking along the street wearing the pseudophone he would hear footsteps to his right when they actually came from the left.

Uses of the pseudophone

When two sources of sound in the same vicinity emit sound waves of slightly different frequencies (*i.e.,* vibrations per second), there will come intervals when waves from both sources arrive at the ear in phase (simultaneously) and produce the experience of a combined, louder sound. These intervals of combined sound will be perceived as "beats," or periodic alternations of sound intensity. When such auditory beats occur too rapidly to

be discriminated, the resulting experience may be that of a harsh, continuous noise. Another case of interference results when two tones sounded together produce a subjectively heard third tone. When this third tone is lower in pitch than the original two, it is called a difference tone; *i.e.,* its frequency is the difference between the frequencies of the two original tones. When the third tone is higher, it is called a summation tone; *i.e.,* its frequency is the sum of the frequencies of the two original tones. Piano tuners depend in part upon their ability to employ these experiences in a reliable way in tightening and loosening the strings of the instruments.

Visual illusions. Numerous optical illusions are produced by the refraction (bending) of light as it passes through one substance to another in which the speed of light is significantly different. A ray of light passing from one transparent medium (air) to another (water) is bent as it emerges. Thus, the pencil standing in water seems broken at the surface where the air and water meet; in the same way, a partially submerged log in the water of a swamp gives the illusion of being bent.

Rainbows also result from refraction. As the sun's rays pass through rain, the droplets separate (refract) the white light into its component colours (see LIGHT). As rays of white light from any source pass through a prism, they are refracted to give the appearance of a spectrum of colour as in the rainbow of a summer morning. Another illusion that depends on atmospheric conditions is the mirage, in which, for example, the vision of a pool of water is created by light passing the layers of hot air above the heated surface of a highway. In effect, cooler layers of air refract the sun's rays at different angles than do less dense strata of heated air, giving the appearance of water where there is none; nearby objects may even appear to be reflected in it. Under unusual conditions, more elaborate mirages may appear as cities, forests, "unidentified flying objects," oases, and even as the images of ships in a nearby body of water plying the sky of a desert.

Perceiver-distortion illusions. Some illusions are related to perceiver characteristics such as brain function. When an observer is confronted with a visual assortment of dots, for example, the brain may appear to group the dots that "belong together." These groupings are made on the basis of such things as observed similarity (*e.g.,* red versus black dots), proximity, common direction of movement, perceptual set (the way one is expecting to see things grouped), and extrapolation (one's estimate of what will happen based on an extension of what is now happening).

Time-induced error is an example of a Gestalt illusion that occurs over brief time intervals. Two images of the same line, for example, will appear to differ in length if they are flashed quickly one after the other.

Closure

Closure (another Gestalt term) is the illusion of seeing an incomplete stimulus as though it were whole. Thus, one unconsciously tends to complete (close) a triangle or a square with a gap in one of its sides. In watching movies, closure occurs to fill the intervals between what are really rapidly projected still pictures, giving the illusion of uninterrupted motion.

Visual illusion. The "figure and ground" illusion is commonly experienced when one gazes at the illustration of a white vase the outline of which is created by two black profiles (see Figure 4). At any moment one will be able to see either the white vase (in the centre area) as "figure" or the black profiles on each side (in which case the white is seen as "ground"). The fluctuations of figure and ground may occur even when one fails deliberately to shift attention, appearing without conscious effort. Seeing one aspect apparently excludes seeing the other.

Another example of ambiguity and object reversibility is the Necker cube (see Figure 5), which may seem to "flip-flop"; younger people tend to perceive these reversals more readily than do their elders.

The so-called Müller-Lyer illusion (also Figure 5) is based on the Gestalt principles of convergence and divergence: the lines at the sides seem to lead the eye either inward or outward to create a false impression of length. The Poggendorff illusion depends on the steepness of the intersecting lines. As obliqueness is decreased, the illusion becomes less compelling. In the Zöllner illusion, the cross-hatching disturbs the perception of parallel lines. A figure seen touching converging lines, as in the Ponzo illusion, creates the impression of size larger than does another figure placed between the lines where they are farther apart. In a related experience, linear perspective creates the illusion that parallel lines or contours (such as railroad tracks) converge as they recede from the viewer. If it were not for these converging lines, a figure in the distant background might appear smaller than would an identical figure in the foreground.

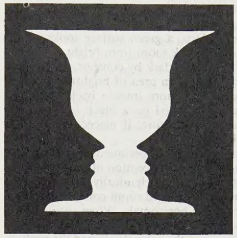


Figure 4: Ambiguous figure seen as either white vase or two black profiles.

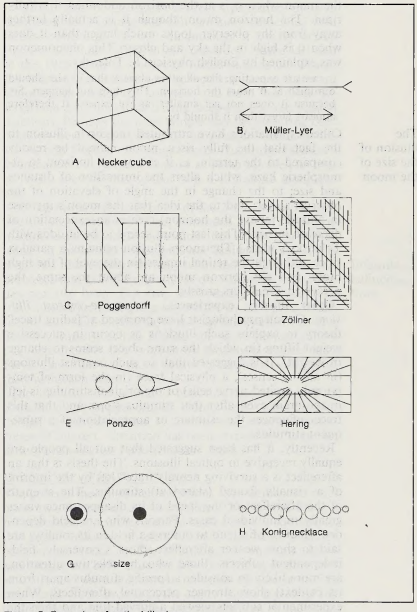


Figure 5: *Examples of optical illusions.*
(A) Cube changes orientation. (B) Lines are equal in length. (C) Lines covered by rectangles are straight. (D) All long lines are parallel. (E) Circles are equal in size. (F) Horizontal lines are parallel. (G) Black dots are equal in size. (H) Tops of circles are on a straight line.

Colour in visual illusions

Visual illusions include a variety of contrast colour phenomena. A successive contrast occurs when, after one has stared at a red surface, a green surface looks much brighter. As one enters a dark room from bright sunshine, the room at first seems quite dark by contrast. A simultaneous contrast occurs when an area of brightness is seen against a less intense or a more intense background. If a gray patch of paper is placed on a black background, it looks whiter than it did before; if placed on a white background, it looks darker.

In studies of visual verticality, experimenters investigated the conditions that determine perception of the "upright." A tilted chair that could be mechanically controlled by the subject was placed in a slanted room containing visual indicators of verticals and horizontals. When various persons were asked to sit in the chair and align themselves in a vertical position, some of the subjects aligned themselves with the "true vertical" determined by gravity, while others experienced the illusion of verticality by aligning themselves with the visual directions they saw in the slanted room. Closing the eyes made "true" alignment easier.

Staring at a single bright spot in an otherwise darkened room creates the illusion that the stationary light is moving (autokinetic effect). One theory to account for this is that the impression is caused by minute eye movements of the observer. The so-called phi phenomenon is an illusion of movement that arises when stationary objects, light bulbs for example, are placed side by side and illuminated rapidly one after another. The effect is frequently used on theatre marquees to give the impression of moving lights.

There is a well-known apparent difference in the size of the moon when it is at the horizon and when it is fully risen. The horizon moon, though it is actually farther away from the observer, looks much larger than it does when it is high in the sky and closer. This phenomenon was explained by English physicist S. Tolansky:

> . . . we are expecting, like all other objects, that its size should diminish as it nears the horizon. This does not happen. So, because it does not get smaller, as we expect, it therefore appears larger than it should be.

The illusion of the size of the moon

Other explanations have attributed the moon illusion to the fact that the fully risen moon cannot be readily compared to the terrain, as it can at the horizon; to atmospheric haze, which alters the impression of distance and size; to the change in the angle of elevation of the observer's eyes; and to the idea that the moon's increase in altitude (above the horizon) creates an expectation of decrease in size. (This last point seems to be at odds with Tolansky's ideas.) The moon illusion remains a paradox since, although the retinal images (in the eye) of the high moon and the horizon moon are about the same, the perceived size differs grossly.

Other illusory experiences. *Successive-contrast illusions.* Gestalt psychologists have proposed a "fading trace" theory to explain such illusions as occur in successive weight lifting (in which the same object seems to change in weight). It is suggested that, in such contrast illusions (in all the senses), a physical trace (in the form of temporarily excited nerve cells) of the original stimulus is left in the brain even after that stimulus stops; and that this trace influences the estimate or appreciation of a subsequent stimulus.

Recently, it has been suggested that not all people are equally receptive to optical illusions. The thesis is that an aftereffect is a surviving neural "trace" left by the imprint of a visually fixated (stared at) stimulus. The strength of the aftereffect or the speed of its disappearance varies greatly in individual cases. Persons who are field dependent (that is, who tend to observe a field in its totality) are said to show weaker aftereffect traces. Conversely, field-independent subjects (those who, by selective attention, are more likely to consider a specific stimulus apart from its context) show stronger perceptual aftereffects. When experimental subjects viewed a curved line and then the Hering figure (displayed elsewhere in this article), it was found that those who exhibit potent aftereffects (and are field independent) are able to counteract optical illusions, while those who have weaker aftereffects are more prone to experience visual illusions.

Tactual illusions. The skin contains numerous "spots" that respond selectively to either cold or warmth, but generally not to both. It can happen, however, that a very warm stimulus will produce a sensation of cold when placed on a spot that responds to cold. Thus, when a warm stimulus is perceived as cold, the illusion is called that of "paradoxical cold."

Illusions of hot and cold

"Paradoxical heat," a less frequent experience, results from stimulating warm and cold spots simultaneously. It appears to be a fusion of warm and paradoxical cold effects, producing a strange, somewhat unpleasant sensation of "heat" that seems to be attended by uneasiness resembling that of pain. The sensation is sometimes called "psychological heat."

Sudden temperature contrasts can play tricks on the senses. If hot water is run over one hand, and cold water over the other long enough for both to adjust to the temperatures and both are then plunged into lukewarm water, the resulting sensation will be that the cold hand will feel warm and the hot, cold. It would seem that in plunging the cold-adapted hand, nerve cells for perceiving cold were suddenly inhibited and those for perceiving hot were suddenly stimulated, while in the hot-adapted hand the reverse took place.

A single pencil may be felt as if it were two when it is held between crossed fingers in such a way that the skin is simultaneously stimulated at two points that would usually require two separate objects to produce such a sensation.

Anchor-effect phenomena. Sensitivity in any sense (sight, hearing, touch, and so on) may be measured as the just-perceptible intensity (threshold, or limen) of the appropriate stimulus. A light too dim to see is said to be subliminal (or below the visual threshold); a sound stimulus that is loud enough to hear is described as being above the auditory threshold (as being supraliminal). The smallest detectable stimulus is called the absolute threshold; the smallest detectable change in the intensity of a stimulus is called the difference threshold. Since such thresholds fluctuate within the same individual under different conditions, they are recognized as statistical (average) values rather than as fixed personal characteristics. At any rate, such thresholds can serve as points of reference, or anchors, against which subsequent stimuli are judged or perceived.

Other colour illusions. The normal human eye can function to detect about 130 gradations of colour in the visible spectrum (as in the rainbow), about 20 barely noticeable differences within a given colour, and about 500 of brightness. When two spots of equally bright light are observed in close succession, the first intensity may seem brighter. (This aftereffect is another example of time error.) The first light may be said to serve the function of brightness adaptation (or adjustment) in the eye, causing the second light to fall on a partly adapted and therefore less sensitive retina. In a brief time such excitement in the retina (or even in the brain) tends to subside, or fade. As a result of what theorists call "fading traces" of excitement (held to produce such aftereffect phenomena), various hues of a given colour may appear to be lighter or darker when looked at successively or in contrast to the background.

Weight illusions. The felt perception of differences in weights received experimental attention in 1899, which experiments indicated that a second weight is felt to be either heavier or lighter than an immediately preceding first one, partially as a result of the expectancy of the person doing the lifting. Having lifted the first weight, the subject is "set" for a certain effort on the next try. If the second weight is lifted quickly and easily, it will feel lighter than the first; if it comes up more slowly, it will feel heavier. Expectancy, or set, is often invoked in efforts to explain the size-weight illusion, in which a large cardboard box feels lighter than a smaller box even though both weigh the same.

Weight perception

Olfactory intensity. Smell (olfactory) discrimination is influenced by any odour to which the olfactory structures already have adapted. Receptors in the nose, however, adapt quickly and cease to respond to a particular stimulus producing so-called olfactory fatigue. Thus, an odour

that is strong at first will gradually become imperceptible, as happens when one becomes unaware of the smell of his own body. There may also be present the phenomenon of masking; this is a decrease in sensitivity to one odour after exposure to another (for example, a strong-smelling disinfectant).

Loudness illusions. The human ear typically serves to distinguish between about 1,500 levels of pitch. For loudness, differential-threshold studies reveal about 325 separately perceived levels in the region of greatest auditory sensitivity (about 1,000 to 4,000 cycles per second). The number of discriminable tones is in the hundred thousands. When two sounds are heard in close succession (time error once more) the intensity or loudness of the second is judged by comparing it with the first. Thus, a murmur may sound loud when compared to a whisper, or a "deafening" noise may make all other sounds inaudible. The hum of an electric fan may help to diffuse the street noises of traffic and thus improve the discrimination of sounds in the room.

Intersensory effects. Normally, the senses combine to produce a kind of common, unitary, or integrated perceptual experience. In dining, for example, the visual array on the table, the conversational tones or background music, the tactile sensations, aromas, and taste of the food, all combine to enhance the gustatory experience with each sense (*e.g.*, sight, hearing) contributing to it. Physiologically, taste and smell appear to be particularly subject to intersensory effects (interdependent). In other situations, seeing, hearing, touching, and often smelling and tasting are all employed in an intersensory way in object identification or location.

Synesthesia. A "crossing" of the senses, analogous to a short circuit in a radio and called synesthesia, has been reported. "Colour-hearing," in which people say that specific sounds evoke in them the actual experience of certain colours, is relatively frequent. Some musicians and others report that they see particular colours whenever they hear given tones and musical passages; poets sometimes claim to hear sounds or musical tones when they see words, images, and colours. Synesthesia may be induced with drugs, and in rare psychiatric disorders the sufferer may not be able to tell whether he is seeing or hearing.

Sounds can evoke visual experiences

Intersensory rivalry and facilitation. Stimulation through one sense may enhance the function of another. Seeing a boat rocked by waves may activate the sense of balance in an observer on a pier to the point at which it causes seasickness. A painting of an Arctic scene of frost and snow may evoke the sensation of cold or a shiver that produces gooseflesh. An explosion or gunshots may give a bystander the illusion of being struck; and a picture of appetizing food may evoke sensations of taste and smell.

Sensory rivalry, in which one stimulus inhibits the perception of another, may result from a conflict of cues if sensory information is ambiguous or discrepant, as in the tilted-room experiment, during which the visual sense conflicts with cues from the sense of equilibrium. States of pain, panic, monotony, or fatigue may create conditions in which various senses mask or inhibit each other. A witness of a terrifying sight, for example, may become oblivious to all sounds around him. Distraction can elevate the pain threshold, as in the case of wounded soldiers whose injuries become painful only after the stress of combat subsides. In a similar way, some dentists use auditory analgesia (a "masking" of pain by sound).

Illusions of psychiatric significance. Illusions called pseudohallucinations occur at times when feelings of anxiety or fear are projected on external objects, as when a child perceives threatening faces or monsters in shadows at night or sees hobgoblins in trees. A soldier tense with apprehension may, in his fear, perceive inanimate objects as an attacking enemy or one of his own comrades as the foe. In literature, the character Don Quixote perceived windmills as enemy knights. Psychiatric patients have perceived other people as machines, or as teddy bears, or as devils.

Déjà vu

The déjà vu phenomenon is a feeling that a past episode is repeating itself in the present; there is a fusion of past and present to create an illusion that one is reliving an experience and that he therefore knows its outcome. It might be called a "hallucination" of familiarity; some theorists interpret the experience as being based on reactivation of old memory traces by stimuli resembling those experienced in the past in the way that a pressed rose discovered in a long-forgotten dance program, for example, may trigger a flood of old memories (see also MEMORY).

Emotions, compelling associations, or strong expectations frequently cause illusional misperceptions in everyday life. The hostile listener hears someone say "wire" and thinks he is being called a "liar"; the self-consciously obese girl misinterprets the word "fate" as "weight." There is also the mistaken identification of strangers as friends in the street. It is as if John thinks Tom is Dick because he is hoping to see Dick or because he definitely wants to avoid Tom.

THE DEVELOPMENT OF MODERN UNDERSTANDING OF HALLUCINATIONS

Traditional psychiatric sources define hallucinations as "perceptions without corresponding stimuli from without" (Swiss psychiatrist E. Bleuler), or as the "apparent perception of an external object when no such object is present" (L.E. Hinsie and J. Shatzky, *Psychiatric Dictionary*, 1940). A historical survey of the subject of hallucinations clearly reflects the development of scientific thought in psychiatry, psychology, and neurobiology. By 1838 the significant relationship between the content of dreams and of hallucinations had been pointed out. In the 1840s the occurrence of hallucinations under a wide variety of conditions (including psychological and physical stress) as well as their genesis through the effects of such drugs as stramonium and hashish had been described.

A.J.F. Brierre de Boismont, a French physician, in 1845 described many instances of hallucinations associated with intense concentration, or with musing, or simply occurring in the course of psychiatric disorder. In the last half of the 19th century, studies of hallucinations continued. Investigators in France were particularly oriented toward abnormal psychological function, and from this came descriptions of hallucinosis during sleepwalking and related reactions. Perhaps the most simple and yet enduring conceptions were those evolved by Sir Francis Galton in the 1880s, and English neurologist John Hughlings Jackson's formulation of the hallucination as being released or triggered by the nervous system was a milestone along the way.

During the first three decades of the 20th century, a spirited interest in hallucinations continued. Freud's concepts of conscious and unconscious activities added new significance to the content of dreams and hallucinations. It was theorized that infants normally hallucinate the objects and processes that give them gratification. Although the notion has recently been disputed, the "regression" hypothesis (*i.e.*, that hallucinating is a regression, or return, to infantile ways) is still widely employed, especially by those who find it clinically useful. During the same period, others put forth theories that were more broadly biological than Freud's but that had more points in common with Freud than with each other.

Infantile hallucinations

The medical and scientific literature has continued to contain many references to hallucinatory phenomena, though for 20 years after 1932 there was a surprising decrease of interest. Attention has been revived by the recent upsurge of work on hallucinogenic drugs.

THE NATURE OF HALLUCINATIONS

The general theory of hallucinations here delineated rests upon two fundamental assumptions. One is that life experiences influence the brain in such a way as to leave in that organ enduring physical changes that have variously been called neural traces, templates, or engrams. Ideas and images are held to derive from the incorporation and activation of these engrams in complex circuits involving nerve cells. Such circuits in the cortex (outer layers) of the brain appear to subserve the neurophysiology of memory, thought, imagination, and fantasy. The emotions associated with these intellectual and perceptual functions seem to be mediated through cortex connections with the deeper parts of the brain (the limbic system or "visceral brain,"

for example), thus permitting a dynamic interplay between perception and emotion through transactions that appear to take place largely at unconscious levels.

Insofar as conscious awareness can be interpreted neurophysiologically, it is found to be regulated through a general arousal process the influence of which is mediated by the ascending midbrain reticular activating system (a network of nerve cells in the brainstem). Analyses of hallucinations reported by sufferers of neurological disorders and by neurosurgical patients in whom the brain is stimulated electrically have shown the importance of the temporal lobes (at the sides of the brain) to auditory hallucinations, for example, and of other functionally relevant parts of the brain in this process.

Interaction of internal and external stimuli

A second basic assumption is that the total human personality is best understood in terms of the constant interplay of forces that continually emanate from inside (as internal physiological activity) and from outside the individual (as sensory stimuli). Such transactions between the environment and the individual may be said to exert an integrating and organizing influence upon memory traces stored in the nervous system and to affect the patterns in which sensory engrams are activated to produce experiences called images, fantasies, dreams, or hallucinations, as well as the emotions associated with these patterns. If such a constantly shifting balance exists between internal and external environmental forces, both physiological considerations (*e.g.,* brain function) and cultural, experiential factors emerge as major determinants of the content and meaning of hallucinations.

The brain is bombarded constantly by sensory impulses, but most of these are excluded from consciousness in a dynamically shifting, selective fashion. The exclusion seems to be accomplished through the exercise of integrative inner mechanisms that permit the limited field of one's awareness to hold selected parts of potential experience in clear focus. (The sound of a ticking clock, for example, fades in and out of awareness.) These mechanisms somehow must simultaneously survey previously stored information within the brain, select tiny samples needed to give adaptive significance to the incoming flow of information from the environment, and bring forth only a few items for actual recall from the brain's enormous "memory banks." (L.J.W.)

Attention. Thus, the work of concentration or attention may be defined as a scanning and screening process, tending to keep out of consciousness what is not needed or wanted. Theoretically, this work involves an activity akin to information retrieval in digital computers (called "secondary process" thinking by Freud), which employs mainly the associative memory functions of the most recently evolved parts of the brain cortex. It may be contrasted with the way in which unconscious, instinctual, and emotional (analogue) information processing (Freud's "primary process" thinking) primarily involves functions of the parts of the brain that appeared much earlier in evolution. In conditions of psychological health, these primary and secondary functions seem highly integrated with each other. There is evidence that in dreams, delirium, or major psychiatric disorder (*e.g.,* schizophrenia), images, feelings, thoughts, and perceptions become dissociated from each other, and fragments of information arising within the brain itself are confused with information coming in from the external environment.

Primary and secondary functions

During normal wakefulness, the input of information through the sensory pathways serves a basic function in maintaining the organization of scanning and screening activity. As long as it is working well, the brain mechanism of scanning and screening seems to exclude from awareness not only information from the internal and external environments that is undesired or has low priority but also the vast bulk of information already stored within the brain in the form of neural traces, their derivations, and interrelations. Some of this information can be brought deliberately into awareness. (One readily recalls his address when asked, for example.) Many children and a few adults can screen in and scan perceptual memory traces with great clarity, thereby permitting eidetic ("photographic") or near-eidetic imagery.

Reduced sensory activity. When, as in sleep, sensory input is diminished, distorted, or impaired, its organizing effect upon the screening and scanning mechanism tends to decrease. Simultaneously, as a rule, there is a decrease in the stimulating effect of sensory input on the ascending midbrain reticular activating system (through connections from the major incoming sensory pathways, such as tactual structures, passing through the brainstem), and, as a result, arousal and awareness diminish. Under a variety of circumstances, however, great reduction or impairment of sensory input may be accompanied by a residual awareness of considerable degree. In such instances, when the usual information-input level no longer suffices completely to inhibit their emergence, ordinarily quiescent perceptual traces stored in the brain may be activated or "released" and re-experienced. Released perceptions of this sort do not ordinarily become conscious with hallucinatory vividness. In order for them to emerge even into clear awareness, there must be a sufficient general level of arousal for awareness to occur, and the particular perception-bearing brain circuits must trigger and reverberate sufficiently to command awareness.

A sustained level and variety of sensory input normally contributes to the process that inhibits the emergence of reactivated memory traces from within the brain. When effective (attention-commanding) sensory input decreases sufficiently, there may be a release into awareness of previously recorded perceptions through disinhibition of the brain circuits that represent them. If a sufficient level of cortical arousal persists or is stimulated in any one of several ways, these released perceptions can enter awareness and be experienced as fantasies, illusions, visions, dreams, or hallucinations. The greater the level of arousal, the more vivid the hallucinations will be.

INDUCTION OF HALLUCINATIONS

Direct brain stimulation. Hallucinatory experiences can be brought about through direct stimulation of the exposed brain, as in cases of local cortical arousal under the neurosurgeon's stimulating electrode. Under circumstances other than surgery or such situations as, for example, some forms of epilepsy, however, there must be a decrease in the forces that ordinarily dominate consciousness (and inhibit the release of recorded percepts) before hallucinations occur. These inhibiting forces require for their maintenance a relatively high level of sensory input of appropriate quality and frequency.

Sleep. The ways in which the reticular network of cells in the brainstem acts as a regulatory and integrating system for these relationships remain under intensive study. Since levels of brain arousal during sleep and wakefulness also are mediated via reticular formation activity, sleeping and dreaming merit consideration as hallucinatory activities. As a person falls asleep, he passes through a period of "partial sleep" in which awareness of the environment drops rapidly but in which the level of cortical arousal (which falls less rapidly) remains sufficiently high to permit some appreciation of external stimulation. Thus, the so-called hypnagogic phenomena occur.

Dreams and hallucinations

Hypnagogic hallucinations. Common hypnagogic hallucinations may be visual (*e.g.,* scenes from the previous few hours appear) or auditory (*e.g.,* one seems to hear his name called). A frequently occurring hypnagogic hallucination is the sensation of loss of support or balance, perhaps accompanied by a fragmentary "dream" of falling, followed immediately by a jerking reflex recovery movement (the myoclonic jerk) that may jolt the sleeper back into wakefulness.

Dreams. The first stages of sleep appear to be dream free, the sleeper producing large, slow brain waves that may be recorded with a device called the electroencephalograph (EEG.) Sensory stimuli from without (*e.g.,* noise, cold) or stimuli from within the body (*e.g.,* dyspepsia, anxiety), plus a somewhat regular spontaneous fluctuation in the depth of sleep, periodically (perhaps every 90 minutes or so) bring the sleeper into a state that favours perceptual release, in which case dreaming tends to take place. This state (with specific EEG signs and rapid movements of the closed eyes) probably occurs several times every night, so

that more than 20 percent of an average person's sleep is taken up with several dreams, each of about 10 to 15 minutes' duration. At the time of awakening, the typical sleeper again passes through a period of perceptual release, often experiencing dreams that increase in intensity, and perhaps may have the hallucinations of "partial sleep," these now being called hypnopompic experiences.

A simplified but perhaps helpful way of characterizing these conditions might be to imagine a person standing at a window opposite a fireplace, looking out at a garden in the sunset. He is so absorbed by the view of the outside world that he fails to visualize the interior of the room at all. As it becomes darker outside, images of the objects in the room behind him are reflected dimly in the window glass. For a time he may see either the garden (if he gazes into the distance) or the reflection of the room's interior (if he focusses on the glass). Night falls, but the fire still illuminates the room. The watcher now sees in the glass a reflection of the interior of the room behind him, which appears to be outside the window. This illusion becomes dimmer as the fire dies down, and, finally, when it is dark both outside and within, nothing more is seen. If the fire flares up from time to time, the visions in the glass reappear.

In an analogous way, hallucinatory experiences such as those of normal dreams occur when the "daylight" (sensory input) is reduced while the "interior illumination" (general level of brain arousal) remains "bright," and images originating within the "rooms" of our brains may be perceived (hallucinated) as though they came from outside the "windows" of our senses.

Another analogy might be that dreams, like the stars, are shining all the time. Though the stars are not often seen by day, since the sun shines too brightly, if, during the day, there is an eclipse of the sun, or if a viewer chooses to be watchful awhile after sunset or awhile before sunrise, or if he is awakened from time to time on a clear night to look at the sky, then the stars, like dreams, though often forgotten, may always be seen.

A more brain-related concept is that of a continuous information-processing activity (a kind of "preconscious stream") that is influenced continually by both conscious and unconscious forces and that constitutes the potential supply of dream content. The dream is an experience during which, for a few minutes, the individual has some awareness of the stream of data being processed. Hallucinations in the waking state also would involve the same phenomenon, produced by a somewhat different set of psychological or physiological circumstances.

Excessive excitation. It is valuable to consider the probable relationship between the level of physiological arousal in the brain and information processing during the waking state. The functions of consciousness apparently reach an optimal point in relation to level of arousal, beyond which they disorganize progressively as arousal increases. The presence of marked arousal (produced, for example, by extreme anxiety or by chemical stimulation of the brain) is accompanied by marked disturbance of concentration. Again, contact with external stimuli is impaired, this time by excessive input that "jams the circuits," in which case spontaneous dissociative experiences may occur. As arousal increases further, the hallucinations of full-blown delirium or psychotic excitement may appear with frightening vividness, intensity, and emotional accompaniment. Greater brain arousal might result in generalized seizure phenomena, as in epilepsy.

Sensory deprivation. When people are kept in isolation (sensory deprivation), information input via the senses (*e.g.,* hearing and sight) is depatterned or reduced. If such a person remains alert, he is likely to experience vivid fantasies and perhaps hallucinations. A slight amount of stimulation of the hallucinated sense may enhance the likelihood of the hallucination's appearance. If stimuli are markedly reduced and the level of arousal is high, the hallucinations can be especially vivid and emotionally charged.

Loss of sleep. Progressive sleep loss appears to decrease one's capacity for integrating perceptions of the external environment. Hallucinations probably will occur in anyone if wakefulness is sufficiently prolonged; anxiety is likely to hasten or to enhance hallucinatory production. The disorganizing effect of excessive wakefulness has been exploited in extorting confessions from prisoners. Observations suggest that fleeting hallucinations typically begin after two or three days without sleep, and that after 100 to 200 sleepless hours a progressive personality disorganization will develop, marked by periods of hallucinosis or, in some cases, by the reappearance of a previously existing psychiatric disorder.

Hypnosis and trance states. The mystic achieves hallucinations by gaining control of his own dissociative mechanisms; perhaps this is a form of self-hypnosis. Such individuals can accomplish an astonishing withdrawal from the environment by prolonged intense concentration (*e.g.,* by gazing at some object). The hallucinations may be of the type in which the person perceives his "inner self" to leave his body to view himself (autoscopic hallucination) or to be transported to new surroundings. Alternatively, the hallucinations may take the form of unique visual imagery; for example, the *yantra* is a visual hallucination of a coloured, geometrical image that appears at a level of trance of the sort experienced by practitioners of Yoga. The recurrence of certain designs and patterns in human hallucinatory experience is probably related to structural aspects of the visual system.

The *yantra*

Ordinary experimental hypnotic and posthypnotic suggestions of hallucinations are well known. The hypnotic subject (who can be described as a person in a controlled dissociative state) may on occasion also experience spontaneous hallucinations in the absence of specific suggestions.

Prolonged monotony or fixation of attention may lead to diminished responsiveness to the environment with a general effect similar to that of absolute reduction of stimulation or of hypnotic trance. Under these conditions such dissociative phenomena as "highway hypnosis" among drivers of motor vehicles may occur. Similar phenomena that occur among aviators have been called fascination or fixation. During prolonged, monotonous flight, pilots may experience visual, auditory, and bodily (kinesthetic) hallucinations; for example, one may suddenly feel that his plane is in a spin or a dive or that it is upside down, even though it is flying level. A kinesthetic hallucination such as this can be so vivid that the pilot will attempt "corrective" manoeuvring of the aircraft, with potentially tragic results.

Sensory defects. Many other examples of hallucinations related to decrease or impairment of sensory input are known in clinical medicine. Visual hallucinations may occur in cases of cataract (opaque lens in the eye) and have been compared with phantom limb experiences (*e.g.,* "pain" in the toes of a missing foot), since there is an absence of normal stimuli from the environment in both cases. Individuals who suffer a progressive loss of hearing may experience auditory hallucinations. A case of combined visual and auditory hallucinations in a patient with progressive blindness (from cataracts) and with deafness from ear disease (otosclerosis) has been observed. Hallucinations of the phantom limb probably arise as the projection of an experientially established set of brain engrams in the absence of long-accustomed sensory input from the missing part. Differences in nerve impulses from those once produced by a missing limb may cause the phantom to be perceived as being distorted in proportion or in size.

Hallucination of a phantom limb

Psychological factors. Although the role of expectation (mental set) continues to be studied in relation to perception, there can be no doubt of the significance of psychological factors in determining the nature of hallucinated objects. It may be that the psychophysiologic basis for recognition requires the unconscious preparation of a perceptual engram (the physically stored memory of a previously seen object, for example) against which to match incoming sensory information for identification, significance, and meaning in terms of past experience. If some external object is present but inadequately recognized, an incorrect perceptual engram may be activated to be experienced as an illusion; in the absence of an external stimulus, such an engram is perceived as a hallucination. This may account for the specificity of collective visions

(*i.e.,* those shared by more than one person). Among lifeboat survivors at sea, for example, several people who share similar expectancies (mental sets) may see a nonexistent ship projected against the blank screen of empty sea and sky. Such an experience may persist in some of the people even after a logical belief in its impossibility has been communicated to all.

Multiple factors undoubtedly combine in bringing about the psychiatric symptoms of the psychoses; these symptoms often resemble the waking dreams in which hallucinations (usually auditory) may figure prominently. Such additive effects can also be demonstrated among "normal" people in the laboratory; for example, one may readily produce signs of hallucinations among sleep-deprived subjects or among subjects in a state of sensory isolation by administering otherwise subhallucinatory doses of drugs such as LSD or mescaline. In hospital cases of acute psychotic reactions with hallucinosis, combinations of factors clearly can be inferred to be at work: hereditary and cultural predispositions; excessive arousal in anxiety or panic; autointoxication (self-poisoning via deranged body physiology) through stress, exhaustion, sleep loss, and dehydration (water loss); and dissociative mechanisms that impair or distort the reception of information from a frightening or threatening social environment.

Chemical factors. Hallucinations may be produced by chemical changes deriving from internal metabolic disturbances that are otherwise engendered inside of the body, or that originate from outside of the body. Some chemicals that produce hallucinatory experiences seem to act by reducing sensory input; for example, dramatic hallucinatory recall of intense experiences from the recent past can be brought about by injections of anesthetic drugs such as sodium amybarbital, which favours the conditions for perceptual release. Hallucinations during induction of (and emergence from) general surgical anesthesia induced by a variety of other chemicals are well-known and can be explained on the same basis.

Such hallucinogenic chemicals seem to impair sensory input by decreasing the transmission of nerve impulses by raising the resistance of the nervous system to their passage. Other hallucinogens increase nerve transmission, disrupting the orderly input of information and "jamming the circuits." Many botanically derived hallucinogens seem to function this way—*e.g.,* LSD (lysergic acid diethylamide) and the ergot (a fungus) that grows on rye, psilocybin from mushrooms, mescaline from the peyote cactus, and tetrahydrocannabinol (THC) from marijuana. Hallucinations also can be induced by jamming the circuits through input overload produced mechanically, bombarding several sensory systems with intense stimuli simultaneously (*e.g.,* with bright flashing lights and loud noises).

Hallucinogenic drugs are substances that, administered in pharmacological doses (not toxic overdoses), create gross distortions in perception without causing loss of consciousness. These distortions frequently include hallucinations. Such compounds also are likely to exert profound effects on mood, thought, and observable behaviour. These resemble (or mimic) the disturbances generated in spontaneously occurring psychoses; indeed some hallucinogens have been termed "psychotomimetic" or "psychotogenic" on this account.

Research scientists and clinicians such as psychiatrists have sometimes deliberately taken these compounds in efforts to understand how it feels to be a severely psychiatric patient. It has been hoped that the study of such chemically induced "model psychoses" would lead to improved methods of treatment. In addition, some psychiatric workers speak of "psychedelic" (mind-manifesting) substances, controversially held to expand perceptual horizons and insight among a variety of people under treatment for such disorders as alcoholism, rigid personality patterns, and sexual frigidity.

Self-experimentation

The potentially dangerous psychological changes produced by psychedelic chemicals have sometimes been interpreted as "loosening ego structures," "dissolving ego boundaries," or "disrupting ego defenses." Such changes may include the experiencing of thoughts, feelings, and perceptions that are usually outside the individual's awareness ("unconscious" or "repressed"). Persons who take such drugs (*e.g.,* LSD) may become hypersuggestible, emotionally labile (unstable), and unusually aware of their own reactions and those of others. Feelings of transcendence of ordinary experience, distortions in time perception (*e.g.,* time may seem to slow down), and hallucinations have also been reported.

It appears that all human behaviour and experience (normal as well as abnormal) is well attended by illusory and hallucinatory phenomena. While the relationship of these phenomena to mental illness has been well documented, their role in everyday life has perhaps not been considered enough. Greater understanding of illusions and hallucinations among normal people may provide explanations for experiences otherwise relegated to the uncanny, "extrasensory," or supernatural. Such understanding may also illuminate the remarkable certainty that individuals express in their contrary interpretations of the same basic information. "Reality," like beauty, lies in the eye of the beholder.

(L.J.W./Ed.)

BIBLIOGRAPHY. An excellent general text for undergraduates is ROBERT SEKULER and RANDOLPH BLAKE, *Perception,* 3rd ed. (1994), focusing on perception through the senses. IRVIN ROCK, *An Introduction to Perception* (1975), is a standard introductory text for experimental psychology students.

A clear presentation of Gestalt theory and its case against structuralism and behaviourism is found in WOLFGANG KÖHLER, *Gestalt Psychology* (1947, reissued 1992). A comprehensive historical overview is available in EDWIN G. BORING, *Sensation and Perception in the History of Experimental Psychology* (1942, reissued 1977). Illustrations and discussions of Gestalt principles of organization are provided in WILLIAM N. DEMBER and JOEL S. WARM, *Psychology of Perception,* 2nd ed. (1979); and JULIAN E. HOCHBERG, *Perception,* 2nd ed. (1978). Also of interest are studies by HERMANN VON HELMHOLTZ, *Helmholz's Treatise on Physiological Optics,* ed. by JAMES P. SOUTHALL, 3 vol. (1924, reissued 1962; originally published in German, 3rd ed., 1909–11), the classic work on visual perception and its physiological basis; SHIMON ULLMAN, *The Interpretation of Visual Motion* (1979), an original and accessible account of how we connect successive views of a moving object; and DAVID MARR, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (1982), the book that triggered the computer revolution in vision science.

Implications of research on early experience for perceptual and intellectual development are spelled out in J. MCVICKER HUNT, *Intelligence and Experience* (1961). Two excellent collections of technical articles are RALPH NORMAN HABER (compiler), *Contemporary Theory and Research in Visual Perception* (1968), and *Information-Processing Approaches to Visual Perception* (1969). A scholarly discussion of depth perception is contained in BELA JULESZ, *Foundations of Cyclopean Perception* (1971).

PAUL FRAISSE, *The Psychology of Time* (1963, reprinted 1975; originally published in French, 1957), provides the most complete synthesis of the psychological works on time. The collective works edited by J.T. FRASER, *The Voices of Time,* 2nd ed. (1981); and R. FISHER, "Interdisciplinary Perspectives of Time," *Annals of the New York Academy of Sciences,* vol. 138, article 2 (1967), present the most varied aspects of the problem of time. ERWIN BÜNNING, *The Physiological Clock,* trans. from German, rev. 3rd ed. (1973), studies the problems of the temporal regulation of organisms. The point of view of genetic epistemology is developed by JEAN PIAGET, *The Child's Conception of Time* (1969; originally published in French, 1946).

WILLIAM EPSTEIN, *Varieties of Perceptual Learning* (1967), extensively reviews and evaluates diverse phenomena and theoretical formulations. ELEANOR J. GIBSON, *Principles of Perceptual Learning and Development* (1969), a major publication, examines the field from the viewpoint of the discovery theory of perceptual learning. More advanced reading on the perception of objects and space is presented in JAMES E. CUTTING, *Perception with an Eye for Motion* (1986). NED BLOCK (ed.), *Imagery* (1981), collects views on the nature of mental images. RONALD A. FINKE, "Theories Relating Mental Imagery to Perception," *Psychological Bulletin,* 98(2):236–259 (September 1985), is a review article.

Works on the psychology of illusion and hallucination include R.L. GREGORY, *The Intelligent Eye* (1970), a fascinating account of how we see and are sometimes deceived by appearances; S. TOLANSKY, *Optical Illusions* (1964), with most of the known visual illusions illustrated, one of the few modern books on this subject; LOUIS JOLYON WEST (ed.), *Hallucinations* (1962), an analysis of the subject by contributors from several scientific disciplines; and PETER D. SLADE and RICHARD P. BENTALL, *Sensory Deception: A Scientific Analysis of Hallucination* (1988), which discusses therapeutic approaches.

(L.J.W./Ed.)

Personality

The term personality has been defined in many ways, but as a psychological concept two main meanings have evolved. The first pertains to the consistent differences that exist between people: in this sense, the study of personality focuses on classifying and explaining relatively stable human psychological characteristics. The second meaning emphasizes those qualities that make all people alike and that distinguish psychological man from other species; it directs the personality theorist to search for those regularities among all people that define the nature of man as well as the factors that influence the course of lives. This duality may help explain the two directions that personality studies have taken: on the one hand, the study of ever more specific qualities in people, and, on the other, the search for the organized totality of psychological functions that emphasizes the interplay between organic and psychological events within people and those social and biological events that surround them. The dual definition of personality is interwoven in most of the topics discussed below. It should be emphasized, however, that no definition of personality has found universal acceptance within the field.

The study of personality can be said to have its origins in the fundamental idea that people are distinguished by their characteristic individual patterns of behaviour—the distinctive ways in which they walk, talk, furnish their living quarters, or express their urges. Whatever the behaviour, personologists—as those who systematically study personality are called—examine how people differ in the ways they express themselves and attempt to determine the causes of these differences. Although other fields of psychology examine many of the same functions and processes, such as attention, thinking, or motivation, the personologist places emphasis on how these different processes fit together and become integrated so as to give each person a distinctive identity, or personality. The systematic psychological study of personality has emerged from a number of different sources, including psychiatric case studies that focused on lives in distress, from philosophy, which explores the nature of man, and from physiology, anthropology, and social psychology.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 436, and the *Index*.

This article is divided into the following sections:

Personality theories 503
  Physiological type theories 503
    Humoral theories
    Morphological (body type) theories
  Psychoanalytic theories 504
    Freud
    Jung
    Adler
    Erikson
    The impact of psychoanalysis
  Trait theories 505
    Stability of traits
    Deviation from trait theory
  Modern trends in personality studies 506
    Sex differences
    Aggression
    Genetic aspects
    Cognitive controls and styles
  Origins of personality study 507
Personality assessment 508
  Measuring constructs 508
  Assessment methods 508
    The interview
    Rating scales
    Self-report tests
    Projective techniques
    Behavioral assessment
    Cognitive assessment
    Bodily assessment
    Personal facts
  Reliability and validity of assessment methods 513
    Evaluation techniques
    Clinical versus statistical prediction
Bibliography 514

Personality theories

The systematic study of personality as a recognizable and separate discipline within psychology may be said to have begun in the 1930s with the publication in the United States of two textbooks, *Psychology of Personality* (1937) by Ross Stagner and *Personality: A Psychological Interpretation* (1937) by Gordon W. Allport, followed by Henry A. Murray's *Explorations in Personality* (1938), which contained a set of experimental and clinical studies, and by Gardner Murphy's integrative and comprehensive text, *Personality: A Biosocial Approach to Origins and Structure* (1947). Yet personology can trace its ancestry to the ancient Greeks, who proposed a kind of biochemical theory of personality.

PHYSIOLOGICAL TYPE THEORIES

The idea that people fall into certain personality type categories in relation to bodily characteristics has intrigued numerous modern psychologists as well as their counterparts among the ancients. The idea that people must fall into one or another rigid personality class, however, has been largely dismissed. Two general sets of theories are considered here, the humoral and the morphological.

Humoral theories. Perhaps the oldest personality theory known is contained in the cosmological writings of the Greek philosopher and physiologist Empedocles and in related speculations of the physician Hippocrates. Empedocles' cosmic elements—air (with its associated qualities, warm and moist), earth (cold and dry), fire (warm and dry), and water (cold and moist)—were related to health and corresponded (in the above order) to Hippocrates' physical humours, which were associated with variations in temperament: blood (sanguine temperament), black bile (melancholic), yellow bile (choleric), and phlegm (phlegmatic). This theory, with its view that body chemistry determines temperament, has survived in some form for more than 2,500 years. According to these early theorists, emotional stability as well as general health depend on an appropriate balance among the four bodily humours; an excess of one may produce a particular bodily illness or an exaggerated personality trait. Thus, a person with an excess of blood would be expected to have a sanguine temperament—that is, to be optimistic, enthusiastic, and excitable. Too much black bile (dark blood perhaps mixed with other secretions) was believed to produce a melancholic temperament. An oversupply of yellow bile (secreted by the liver) would result in anger, irritability, and a "jaundiced" view of life. An abundance of phlegm (secreted in the respiratory passages) was alleged to make people stolid, apathetic, and undemonstrative. As biological science has progressed, these primitive ideas about body chemistry have been replaced by more complex ideas and by contemporary studies of hormones, neurotransmitters, and substances produced within the central nervous system, such as endorphins.

Hippocrates' physical humours

Morphological (body type) theories. Related to the biochemical theories are those that distinguish types of per-

sonalities on the basis of body shape (somatotype). Such a morphological theory was developed by the German psychiatrist Ernst Kretschmer. In his book *Physique and Character*, first published in 1921, he wrote that among his patients a frail, rather weak (asthenic) body build as well as a muscular (athletic) physique were frequently characteristic of schizophrenic patients, while a short, rotund (pyknic) build was often found among manic-depressive patients. Kretschmer extended his findings and assertions in a theory that related body build and personality in all people and wrote that slim and delicate physiques are associated with introversion, while those with rounded heavier and shorter bodies tend to be cyclothymic—that is, moody but often extroverted and jovial.

Despite early hopes that body types might be useful in classifying personality characteristics or in identifying psychiatric syndromes, the relations observed by Kretschmer were not found to be strongly supported by empirical studies. In the 1930s more elaborate studies by William H. Sheldon in the United States developed a system for assigning a three-digit somatotype number to people, each digit with a range from 1 to 7. Each of the three digits applies to one of Sheldon's three components of body build: the first to the soft, round endomorph, the second to the square, muscular mesomorph; and the third to the linear, fine-boned ectomorph. Thus, an extreme endomorph would be 711, an extreme ectomorph 117, and an average person 444. Sheldon then developed a 20-item list of traits that differentiated three separate categories of behaviours or temperaments. The three-digit temperament scale appeared to be significantly related to the somatotype profile, an association that failed to excite personologists.

Sheldon's three-digit somatotype system

Also during the 1930s, personality studies began to consider the broader social context in which a person lived. The American anthropologist Margaret Mead studied the patterns of cooperation and competition in 13 primitive societies and was able to document wide variations in those behaviours in different societies. In her book *Sex and Temperament in Three Primitive Societies* (1935), she showed that masculinity is not necessarily expressed through aggressiveness and that femininity is not necessarily expressed through passivity and acquiescence. These demonstrated variations raised questions about the relative roles of biology, learning, and cultural pressures in personality characteristics.

PSYCHOANALYTIC THEORIES

Freud. Perhaps the most influential integrative theory of personality is that of psychoanalysis, which was largely promulgated during the first four decades of the 20th century by the Austrian neurologist Sigmund Freud. Although its beginnings were based in studies of psychopathology, psychoanalysis became a more general perspective on normal personality development and functioning. The field of investigation began with case studies of so-called neurotic conditions, which included hysteria, obsessive-compulsive disorders, and phobic conditions. Patients with hysterical symptoms complained of acute shortness of breath, paralyses, and contractures of limbs for which no physical cause could be found. In the course of interviews, Freud and his early coworker and mentor, the Austrian physician Josef Breuer, noted that many of their patients were unsure of how or when their symptoms developed and even seemed indifferent to the enormous inconvenience the symptoms caused them. It was as if the ideas associated with the symptoms were quarantined from the consciousness and lay neglected by normal curiosity. To explain this strange pattern Breuer and Freud made two assumptions. The first was based on the general scientific position of determinism, which was quite prevalent in 19th-century science: although no apparent physical causes could be implicated, these neurotic symptoms were nevertheless caused, or determined, perhaps not by one but by multiple factors, some of which were psychologically motivated. The second assumption entailed unconscious psychological processes; that is, ideas continue to be active, to change, and to influence behaviour even when they are outside of awareness. One source for this assumption was the observation of posthypnotic suggestion, which seemed to imply that past experiences, surviving outside of consciousness as latent memories, could be activated by a signal from the environment and could then influence behaviour even though the hypnotized person was unaware of the reasons for his behaviour.

Influence of unconscious processes

Breuer and Freud believed that the specific motivation for these neurotic symptoms lay in the patient's desire to obliterate from memory profoundly distressing events that were incompatible with the patient's moral standards and therefore in conflict with them. These events were considered to have been sexual in nature, and further exploration convinced Freud that his patients had had even earlier troublesome sexual experiences—usually seductions—the memories of which had lain dormant until awakened by a more recent sexual encounter. Freud reasoned that the earlier seduction experience imparted to the later one its pathogenic force. Freud at first accepted many of the experiences reported by his young, impressionable patients as actual seductions. He later came to believe that many, though not all, of the narrations were fantasies. Based upon this conviction Freud formulated a theory indicating that personality is shaped by such experiences as well as by other traumatic or frustrating events. He postulated that the fantasies about sexual traumas were expressions of a sexual drive. Thereafter in Freud's therapeutic method, the search for actual sexual trauma was replaced by an exploration of the ways in which patients' sexual inclinations, already present in childhood, were expressed in behaviour. Neurosis and personality in general came to be viewed as outcomes of conflict between sexual motivations and defenses against them, the conflict being rooted in early child development.

Freud assumed that his patients were motivated to ward off those fantasies that had an exciting as well as a repelling quality about them. Freud described various psychological devices (defense mechanisms) by which people tried to make the fantasies bearable. For example, in the obsessive-compulsive condition, which refers to persistent unwelcome ideas or recurrent irresistible urges to perform certain acts, such as incessant hand washing, the defense maneuvers are called isolation and displacement. They consist in separating (isolating) a fantasy from its corresponding emotion, and then attaching (displacing) the emotion to another, previously trivial idea; for instance, to the hand washer it is the hands that are dirty rather than the desires. Freud also noted that people who rely on isolation and displacement are otherwise characterized by nonpathological personality qualities such as perfectionism, indecisiveness, and formality in interpersonal contacts. To Freud the fantasies were the mental representations of basic drives, among which sex, aggression, and self-preservation were paramount. These drives, moreover, required taming as the child matured into an adult, and the taming process involved blocking out of consciousness some of the ideas associated with the expression of those drives. Other methods of defense include repression, a kind of withholding of conflicting ideas from recall; projection, the attribution to others of one's own rejected tendencies; and reaction formation, turning into its opposite a tendency rejected in oneself—as in excessive generosity as a defense against avarice. The basic conflict between drives and control processes, which Freud believed to be the basis of several neuroses, was also invoked to explain both dream content and the "psychopathology of everyday life"—the ordinary slips of the tongue (sometimes called Freudian slips) and errors such as forgetting intentions or misplacing objects.

Isolation and displacement

These primary human drives, moreover, were seen to undergo transformations as part of psychological and physical growth. This formulation widened the realm of sexuality beyond reproduction, by proposing that genital activity does not encompass all of sexuality, because sexual activity can be observed long before biological maturity and can occur without leading to reproduction. The theory further proposed that sexual maturation develops in a sequence of stages as parts of the body successively yield sensual pleasure to the child, beginning with the mouth, followed by the anus, and then the genitals. Social demands for inhibition and control of the drives centre about the functions

of these zones, and it is from this process of socialization that personality is said to emerge. For example, the extent to which the personality expresses power, responsibility, compliance, and defiance seems to coincide with anal expressions of the sexual drive and is related to the process of obtaining control over anal functions.

The conflict between the drives—conceptualized as a wholly unconscious structure called the id—and the drive control processes—conceptualized as a largely unconscious structure called the ego—results in the creation of a characteristic style for mediating conflicts, which is assumed to be formed prior to adolescence. While learning and experience are recognized as conspicuous factors in the shaping of these behaviours, the theory also gives prominence to possibly inborn differences in the strength of drives and of the control processes.

Among the controlling functions of the ego are identifications and defenses. Children are inclined to behave like the significant adult models in their environment, Freud postulated. These identifications give identity and individuality to the maturing child. Moreover, the process of self-criticism is part of the ego controls (Freud called it the superego) and acts as an internal and often unconscious conscience that influences moral values.

Role model identification

Jung. The Swiss psychiatrist Carl Gustav Jung, an early adherent of Freud's theories, questioned the degree of emphasis that Freud gave to sexual motivations in personality development. Jung accepted the significant effect of the unconscious processes, but unlike Freud he preferred to emphasize that behaviour is motivated more by abstract, even spiritual, processes than by sexual drives. Jung also focused more on individual differences; in particular he developed a typology of reaction styles, distinguishing between two basic means of modulating basic drives, introversion and extroversion. Introversion was defined as preoccupation with one's inner world at the expense of social interactions and extroversion as a preference for social interplay for living out inner drives (collectively termed libido). The existence of these two types receives empirical support from most studies of traits (see below).

Adler. The Austrian psychiatrist Alfred Adler, another of Freud's early followers, also disputed the importance of sexual motives. Adler described a coping strategy that he called compensation, which he felt was an important influence on behaviour. In his view people compensated for a behavioral deficiency by exaggerating some other behaviour: a process analogous to organic processes called hypertrophy, in which, for example, if one eye is injured, the other eye may compensate by becoming more acute. In Adler's view, a person with a feeling of inferiority related to a physical or mental inadequacy would also develop compensating behaviours or symptoms. Shortness of stature, for example, could lead to the development of domineering, controlling behaviours. Adler assigned a prominent place to family dynamics in personality development. Children's position in their family—their birth order—was seen as determining significant character traits.

Erikson's eight stages of development

Erikson. Freud's emphasis on the developmental unfolding of the sexual, aggressive, and self-preservative motives in personality was modified by the American psychoanalyst Erik H. Erikson, who integrated psychological, social, and biological factors. Erikson's scheme proposed eight stages of the development of drives, which continue past Freud's five stages of childhood (oral, anal, phallic, latency, and genital) and through three stages of adulthood. The stages proceed in leaps according to what is called an epigenetic process. The term epigenesis, borrowed from embryology, refers to the predetermined developmental sequence of parts of an organism. Each part has a special time for its emergence and for its progressive integration within the functioning whole. Each phase of emergence depends upon the successful completion of the preceding phase. According to Erikson, environmental forces exercise their greatest effect on development at the earliest stages of growth, because anything that disturbs one stage affects all of the following stages. As if controlled by a biological timetable, each given stage must be superseded by a new one, receding in significance as the new stage assumes dominance. A constant interleaving at critical periods—in which some parts emerge while others are suppressed—must proceed smoothly if personality problems are to be avoided.

The Freudian theory of development with Erikson's modifications provides for a succession of drive-control (inner and environmental) interactions. These can be fit into a schema of polar attitudes that develop in progressive stages of a person's life, creating a conflict at each stage which should be resolved to avoid extremes of personality development. Erikson thus evolved his eight stages of development, which he described as: (1) infancy: trust versus mistrust; (2) early childhood: autonomy versus shame and doubt; (3) preschool: initiative versus guilt; (4) school age: industry versus inferiority; (5) puberty: identity versus identity confusion; (6) young adulthood: intimacy versus isolation; (7) middle adulthood: generativity versus stagnation; and (8) late adulthood: integrity versus despair.

The impact of psychoanalysis. There is little doubt that psychoanalysis had a profound influence on personality theory during the 20th century. It turned attention from mere description of types of people to an interest in how people become what they are. Psychoanalytic theory emphasizes that the human organism is constantly, though slowly, changing through perpetual interactions, and that, therefore, the human personality can be conceived of as a locus of change with fragile and indefinite boundaries. It suggests that research should focus not only on studies of traits, attitudes, and motives but also on studies that reflect the psychoanalytic view that personality never ceases to develop and that even the rate of personality modification changes during the course of a life. Although the theory holds that conflict and such basic drives as sex and aggression figure prominently in personality development and functioning, their presence may be neither recognizable nor comprehensible to persons untrained to look for those motives. However, personality characteristics are relatively stable over time and across situations, so that a person remains recognizable despite change. Another feature of psychoanalytic theory is the insistence that personality is affected by both biological and psychosocial forces that operate principally within the family, with the major foundations being laid early in life.

The basis of psychoanalytic theory

The data on which psychoanalytic theory rests came from the psychoanalysts' consulting rooms, where patients in conflict told their life stories to their analysts. No provision is made in that setting for experimental manipulation, for independent observation, or for testing the generality of the formulations. As a consequence, although much of the theory has found its way into accepted doctrine, psychoanalysis cannot claim a body of experimentally tested evidence. Nevertheless, psychoanalytic theory provides at least a preliminary framework for much of personality research involving motives and development.

TRAIT THEORIES

Contemporary personality studies are generally empirical and based on experiments. While they are more precise, and thus may be more valid than much of psychoanalytic theory, experiments perforce have a narrower scope than the grand sweep of psychoanalysis. In the 1940s many investigators focused on intensive studies of individual traits and of combinations of traits that seemed to define personality types, such as the "authoritarian personality." Others, like the American psychologists David C. McClelland and John W. Atkinson, studied the characteristic presence of certain needs identified by Murray, such as the need for achievement or affiliation. The method used to measure these needs was to examine the fantasy productions of Murray's Thematic Apperception Test (TAT) and to relate the motive score to other behavioral indexes such as personal history, occupational choice, speed of learning, and persistence of behaviour following failure.

Stability of traits. Traits such as sociability, impulsiveness, meticulousness, truthfulness, and deceit are assumed to be more or less stable over time and across situations. Traits refer not to single instances of a behaviour, such as lying, but to persistent although not unvarying behaviour that, according to some personologists, implies a disposition to respond in a particular, identifiable way. According

to Allport's 1937 textbook, traits represent structures or habits within a person and are not the construction of observers; they are the product of both genetic predispositions and experience. It can be generally stated that traits are merely names for observed regularities in behaviour, but do not explain them. Nevertheless, the study of how traits arise and are integrated within a person forms a major area of personality studies.

In the English language there are several thousand words representing traits, many of them close in meaning to others (for example, meticulous, careful, conscientious). Most of the measurement studies employ self-report (personality) inventories that require people to describe themselves by checking relevant adjectives or answering questions about typical behaviours that they are conscious of displaying. In some measurements, observers rate the behaviour of others. Psychologists such as Hans J. Eysenck in the United Kingdom and Raymond B. Cattell in the United States have attempted to reduce the list to what they could consider to be the smallest possible number of trait clusters. The statistical technique of factor analysis has been favoured for this task, since it explores the correlations among all of the trait names and identifies clusters of correlations among traits that appear to be independent of (uncorrelated with) each other. Common to almost all the trait systems are variables related to emotional stability, energy level, dominance, and sociability, although different investigators choose different names for these factors. Eysenck, for example, has reduced the trait names to but three higher-order factors—introversion–extroversion, neuroticism, and psychoticism—and has attempted to explore the biological roots of each factor.

Eysenck's three basic traits

Deviation from trait theory. The idea that traits represent relatively stable behaviours has received criticism from psychologists who point out that behavioral consistency across situations and across time is not the rule. For example, in a study of children's moral development, the American psychologists Hugh Hartshorne and Mark A. May in 1928 placed 10- to 13-year-old children in situations that gave them the opportunity to lie, steal, or cheat; to spend money on themselves or on other children; and to yield to or resist distractions. The predictive power of personal and educational background was low, and children were not found consistently honest or dishonest, distractible, or altruistic. The most powerful predictor of children's behaviour was what other children around them were doing.

In the 1960s and '70s some psychologists, including Walter Mischel and Albert Bandura in the United States, recalled the Hartshorne and May study and variations of it to support their view that behaviour is controlled not by hypothetical traits but according to the degree of regularity of external stimuli. That is, they believe that personality traits are only consistent if the situation is consistent and that they vary once the situation changes. In their view, behavioral consistency does not reflect stable personality traits. Rather, it is the environment that evokes and shapes the illusion of such traits. This would be in keeping with the view of social learning theorists that personality, like other elements of a person's psychological makeup, is largely a learning phenomenon related to such factors as the imitation of role models. Social learning theory would also contend that personality is more susceptible to change than would trait theory.

Although it has been demonstrated that behaviour is seldom entirely consistent, it also has been shown that it reflects considerable consistency. Even in the Hartshorne and May study some children showed consistently honest or dishonest behaviour, and behavioral consistency was found to increase with age.

Support for personal consistency is bolstered by studies of what has been called the fundamental attribution error. The investigators, most of them social psychologists, report that, in observing the behaviour of others, people exaggerate the role of internal causes and invoke traits as a primary cause (*e.g.*, "John acted the way he did because he is honest"). In assigning cause to their own behaviour, however, people more often cite external causes such as the particular situation. These tendencies are accompanied by another discovered regularity: in seeking sources for their own behaviour, people are likely to favour internal causes (and thus agree with an observer's judgment) when they consider a behaviour to be desirable (*e.g.*, success, as in "I was successful because I am skillful"), and they invoke external, situational causes in judging a behaviour they deem undesirable (*e.g.*, failure, as in "I failed because the test was unfair"). There are, of course, limits to the regularity with which these generalizations hold. Because people tend to know their own characteristics better than observers do, they are generally more aware than observers are of any divergences from their usual behaviour.

Variations in trait identification

Although people may assume the existence of traits in themselves, they do not, in analyzing a specific situation, see themselves as a mere collection of trait names. Consequently, they are not for the most part perplexed by, and often do not recognize, cross-situational inconsistency in their own behaviour. But in observing another person's behaviour, most people attribute high consistency to that person, as if many positive traits could be inferred from the attribution or observation of one positive trait. For example, the American social psychologist Solomon Asch has shown that a physically attractive person will tend to be judged as having many other desirable qualities. Asch also demonstrated that, in forming impressions of the personal characteristics of others, observers are most influenced by their first impression. The reason first impressions seem to be almost indelible is that they carry an excessive amount of new information, which has a high degree of unpredictability. That is, the more new information contained in an event, the more attention it attracts. Since impressions about a person tend to be integrated into a single characterization, an observer may be jarred by recognizing an undesirable fact about an attractive person and may try either to ignore that fact or to mitigate (rationalize) it. These propensities make up a "common sense psychology," in the words of Fritz Heider, an American psychologist. This "naive" psychology, as he called it, consists of a set of rules that guide most people's impressions of other people and of social situations. These rules are used constantly to interpret one's own and other people's behaviour and to predict behaviour under certain conditions. The psychoanalytic view, however, seriously challenges this common sense psychology. Psychoanalysis has no problem explaining that those who put to death countless people in the Nazi death chambers, for example, could also be devoted parents, whereas common sense psychology would have difficulty with this. For the psychoanalyst, a personality may be integrated, but it is rarely seamless and regular. People generally make two types of errors in judging personality: they impute more personality consistency to others than the actors themselves would allow, and they often ignore the operation of unconscious psychological processes that can explain at least some of the inconsistencies.

Much work on trait structure and impression formation has concerned adjectival words that describe traits, and the fact that these studies have been carried out principally in the United States and western Europe has led some anthropologists, such as the American Robert LeVine, to remark that modern personality trait theory is ethnocentric. For example, the folk-psychological concepts and the trait matrices derived from factor analyses include culture-specific assumptions about personal experiences, such as the distinctions between mind and body, natural and supernatural, and intellect and morality, which do not exist in the folk traditions of many non-Western peoples. Unlike most other cultures, Western thought assumes that a high degree of personal autonomy is desirable and that the most important emotional and personal relations are with a marital partner. For some psychologists these cultural differences point to the need for a less culture-bound approach to personality trait theory.

Ethnocentric trait theory

MODERN TRENDS IN PERSONALITY STUDIES

Sex differences. Despite the physical differences between males and females the finding of behavioral differences between the sexes is controversial. Behaviours associated with sex roles depend heavily on the social

and cultural context, and studies of stereotypic male and female roles are therefore understandably ambiguous. Yet some findings indicate small but consistent differences. While there are no differences in measured IQ, itself regarded as a culture-bound assessment, females do better than males on verbal tasks. Girls generally begin to speak earlier than boys and have fewer language problems in school and in the course of maturation. Males generally exhibit greater skill in understanding spatial relations and in solving problems that involve mathematical reasoning. Beginning at the toddler stage, the activity level of males is generally higher than that of females. A related finding is that boys are more likely to be irritable and aggressive than girls and more often behave like bullies. Men usually outscore women in antisocial personality disorders, which consist of persistent lying, stealing, vandalism, and fighting, although these differences do not appear until after about the age of three. A study by the American anthropologists Beatrice B. Whiting and Carolyn P. Edwards found that males were consistently more aggressive than females in seven cultures, suggesting that there is a predisposition in males to respond aggressively to provocative situations, although how and whether the attacking response occurs depends on the social and cultural setting.

The tendency of man to kill his own kind

Aggression. Humans are perhaps the only species of animal that does not have an internal inhibition against slaughtering other members of the species. It has been theorized that man, like other animals, is motivated by an aggressive drive, which has significant survival value, but lacks internal inhibitions against killing his fellow men. Inhibitions, therefore, must be imposed externally by society. Social learning theorists emphasize the decisive effects of situations in triggering and controlling aggression. They account for the poor predictability of aggressive behaviour in man by noting that the environmental context is generally unpredictable. Yet research has shown that an aggressive act is most likely to be produced by a person with a history of aggressive behaviour.

Genetic aspects. While social learning theorists emphasize the active shaping of personality by external social influences, experimental evidence has accumulated that genetic factors play a prominent role, if not in the transmission of specific behaviour patterns, then in the readiness of people to respond to environmental pressures in particular ways. In observations of animals, it is commonplace to find in different breeds of dogs wide divergences in behaviour that are attributed to genetic differences: some are friendly, others aggressive; some are timid, others bold (of course there may also be wide variations within a given breed). Among human infants observed in a neonatal nursery, there are also clearly observable differences in activity, passivity, fussiness, cuddliness, and responsiveness. These patterns, which some authorities say may be genetically influenced, shape the ways in which the infant will interact with the environment and can be considered an expression of personality.

Personality studies of twins

In systematic studies of humans, studies of twins and adopted children have been used to try to evaluate environmental and genetic factors as determinants of a number of behaviour patterns. These studies have shown that genetic factors account for about 50 percent of the range of differences found in a given population. Most of the remaining differences are attributable not to the environment that is common to members of a family but to the environment that is unique to each member of the family or that results from interactions of family members with one another. In the United States, behaviour geneticists such as Robert Plomin report that, in behaviours describable as sociability, impulsiveness, altruism, aggression, and emotional sensitivity, the similarities among monozygotic (identical) twins is twice that among dizygotic (fraternal) twins, with the common environment contributing practically nothing to the similarities. Similar findings are reported for twins reared together or separately.

The study of the genetic aspects of personality is a relatively new undertaking. Almost all populations studied have been from industrialized Western nations whose rearing environments are more nearly alike than different. It is known that the more homogeneous the environment, the stronger the genetic contribution will appear. As with the psychology of traits, cross-cultural studies are required to test the validity of the claims of behaviour genetics.

Cognitive controls and styles. Psychologists have long been aware that people differ in the consistent way in which they receive and respond to information. Some make careful distinctions between stimuli, whereas others blur distinctions, and some may typically prefer to make broad categories, whereas others prefer narrow ones for grouping objects. These consistencies in an individual seem to be fairly stable across time and even across situations. They have been referred to as cognitive controls. Combinations of several cognitive controls within a person have been referred to as cognitive style, of which there can be numerous variations.

Cognitive control studies explore constraints within a person that limit the influence of both environment and motivation, and as such they are expressions of personality. In the 1940s and '50s several studies explored the extent to which personal needs or drives determine what one perceives. In one study, children from rich and poor families were asked to adjust a circle of light to the size of several coins of increasing value and to the size of cardboard disks. All of the children overestimated the size of the coins, although not of the neutral disks, but the poor children overestimated the sizes more than did the rich children. The assumption that need influences such judgments has been widely held. Even Shakespeare, in *A Midsummer Night's Dream*, noted, "Or in the night, imagining some fear, / how easy is a bush supposed a bear."

Limits of psychological drive distortion

But there are limits to the distorting power of drives, and the experimental demonstration of the influence of motives has been difficult to confirm, perhaps because the formal components of cognition—the workings, for example, of attention, judgment, or perception—and individual difference in their expression have been neglected by personologists. Investigators of cognitive controls examine the psychological limits on the distorting effects of needs and of external reality. For example, in estimating the size of a disk, some people are more exact than others, and the extent to which a need can distort size judgments will consequently be limited by the perceiver's preference for strict or relaxed standards of comparison.

The American psychologists George S. Klein and Herman Witkin in the 1940s and '50s were able to show that several cognitive controls were relatively stable over a class of situations and intentions. For example, the psychologists found a stable tendency in some people to blur distinctions between successively appearing stimuli so that elements tended to lose their individuality (leveling) and an equally stable tendency in other individuals to highlight differences (sharpening). This organizing principle is apparent in judgments of the size of a series of objects, as well as in memory, where it may manifest itself in a blurring of elements in the recall of a story.

Another much studied cognitive control is called field dependence-field independence. It pertains to the extent to which people are influenced by inner (field-independent) or environmental (field-dependent) cues in orienting themselves in space and the extent to which they make fine differentiations in the environment. The more field-independent people are, the greater is their ability to articulate a field. There are no general intellectual capacity differences between field-dependent and field-independent people, but there is a tendency for field-dependent people to favour careers that include working with other people, such as teaching or social work. Field-independent people are more often found in careers that involve abstract issues such as mathematics. Cultural differences have also been found. Some Eskimo live and hunt in an environment with little variation, and a high degree of articulation of the field (field independence) would favour survival; some farmers of Sierra Leone, however, who inhabit an area of lush vegetation and many varieties of shape, require less differentiation of the field.

Characteristics of field dependence–field independence

ORIGINS OF PERSONALITY STUDY

In general, information about human personality has come from three different sources of study. The first is biolog-

ical, conceived to have genetic as well as environmental origins. The second is that of the social realm, including the impact of social forces on the growing child that shape such personal responses as motives, traits, behaviours, and attitudes. The third is the examination of clinical contacts with people who have suffered adaptive and adjustive failures. Some authorities have suggested that a greater degree of integration of all three sources of information and the methods derived from them would accelerate the growth of valid information about personality. (P.S.H.)

Personality assessment

The measurement of personal characteristics is the subject matter of the field of psychology today called personality assessment. Assessment is an end result of gathering information intended to advance psychological theory and research and to increase the probability that wise decisions will be made in applied settings (*e.g.,* in selecting the most promising people from a group of job applicants). The approach taken by the specialist in personality assessment is based on the assumption that much of the observable variability in behaviour from one person to another results from differences in the extent to which individuals possess particular underlying personal characteristics (traits). The assessment specialist seeks to define these traits, to measure them objectively, and to relate them to socially significant aspects of behaviour.

A distinctive feature of the scientific approach to personality measurement is the effort, wherever possible, to describe human characteristics in quantitative terms. How much of a trait manifests itself in an individual? How many traits are present? Quantitative personality measurement is especially useful in comparing groups of people as well as individuals. Do groups of people from different cultural and economic backgrounds differ when considered in the light of their particular personality attributes or traits? How large are the group differences?

Overt behaviour is a reflection of interactions among a wide range of underlying factors, including the bodily state of the individual and the effects of that person's past personal experiences. Hence, a narrowly focused approach is inadequate to do justice to the complex human behaviour that occurs under the constantly changing set of challenges, pleasures, demands, and stresses of everyday life. The sophisticated measurement of human personality inescapably depends on the use of a variety of concepts to provide trait definitions and entails the application of various methods of observation and evaluation. Personality theorists and researchers seek to define and to understand the diversity of human traits, the many ways people have of thinking and perceiving and learning and emoting. Such nonmaterial human dimensions, types, and attributes are constructs—in this case, inferences drawn from observed behaviour. Widely studied personality constructs include anxiety, hostility, emotionality, motivation, and introversion-extroversion. Anxiety, for example, is a concept, or construct, inferred in people from what they say, their facial expressions, and their body movements.

Personality constructs

Personality is interactional in two senses. As indicated above, personal characteristics can be thought of as products of interactions among underlying psychological factors; for example, an individual may experience tension because he or she is both shy and desirous of social success. These products, in turn, interact with the types of situations people confront in their daily lives. A person who is anxious about being evaluated might show debilitated performance in evaluative situations (for example, taking tests), but function well in other situations in which an evaluative emphasis is not present. Personality makeup can be either an asset or a liability depending on the situation. For example, some people approach evaluative situations with fear and foreboding, while others seem to be motivated in a desirable direction by competitive pressures associated with performance.

MEASURING CONSTRUCTS

Efforts to measure personality constructs stem from a variety of sources. Frequently they grow out of theories of personality; anxiety and repression (the forgetting of unpleasant experiences), for example, are among the central concepts of the theory of psychoanalysis. It is understandable that efforts would be made to quantify one's degree of anxiety, for example, and to use the score thus obtained in the assessment of and in the prediction of future behaviour. Among the major issues in the study of personality measurement is the question of which of the many personality constructs that have been quantified are basic or fundamental and which can be expected to involve wasted effort in their measurement because they represent poorly defined combinations of more elemental constructs; which measurement techniques are most effective and convenient for the purpose of assessment; and whether it is better to interview people in measuring personality, or to ask them to say, for example, what an inkblot or a cloud in the sky reminds them of.

Problems in measuring personality constructs

Efforts to measure any given personality construct can fail as a result of inadequacies in formulating or defining the trait to be measured and weaknesses in the assessment methods employed. An investigator might desire to specify quantitatively the degree to which individuals are submissive in social and competitive situations. His effectiveness will depend on the particular theory of submissiveness he brings to bear on the problem; on the actual procedures he selects or devises to measure submissiveness; and on the adequacy of the research he performs to demonstrate the usefulness of the measure. Each of these tasks must be considered carefully in evaluating efforts to measure personality attributes.

The methods used in personality description and measurement fall into several categories that differ with regard to the type of information gathered and the methods by which it is obtained. While all should rely on data that come from direct observations of human behaviour if they are to have at least the semblance of scientific value, all may vary with regard to underlying assumptions, validity, and reliability (consistency, in this case).

ASSESSMENT METHODS

Personality tests provide measures of such characteristics as feelings and emotional states, preoccupations, motivations, attitudes, and approaches to interpersonal relations. There is a diversity of approaches to personality assessment, and controversy surrounds many aspects of the widely used methods and techniques. These include such assessments as the interview, rating scales, self-reports, personality inventories, projective techniques, and behavioral observation.

The interview. In an interview the individual under assessment must be given considerable latitude in "telling his story." Interviews have both verbal and nonverbal (*e.g.,* gestural) components. The aim of the interview is to gather information, and the adequacy of the data gathered depends in large part on the questions asked by the interviewer. In an employment interview the focus of the interviewer is generally on the job candidate's work experiences, general and specific attitudes, and occupational goals. In a diagnostic medical or psychiatric interview considerable attention would be paid to the patient's physical health and to any symptoms of behavioral disorder that may have occurred over the years.

Two broad types of interview may be delineated. In the interview designed for use in research, face-to-face contact between an interviewer and interviewee is directed toward eliciting information that may be relevant to particular practical applications under general study or to those personality theories (or hypotheses) being investigated. Another type, the clinical interview, is focused on assessing the status of a particular individual (*e.g.,* a psychiatric patient); such an interview is action-oriented (*i.e.,* it may indicate appropriate treatment). Both research and clinical interviews frequently may be conducted to obtain an individual's life history and biographical information (*e.g.,* identifying facts, family relationships), but they differ in the uses to which the information is put.

The clinical interview

Although it is not feasible to quantify all of the events occurring in an interview, personality researchers have devised ways of categorizing many aspects of the content of

what a person has said. In this approach, called content analysis, the particular categories used depend upon the researchers' interests and ingenuity, but the method of content analysis is quite general and involves the construction of a system of categories that, it is hoped, can be used reliably by an analyst or scorer. The categories may be straightforward (*e.g.,* the number of words uttered by the interviewee during designated time periods), or they may rest on inferences (*e.g.,* the degree of personal unhappiness the interviewee appears to express). The value of content analysis is that it provides the possibility of using frequencies of uttered response to describe verbal behaviour and defines behavioral variables for more-or-less precise study in experimental research. Content analysis has been used, for example, to gauge changes in attitude as they occur within a person with the passage of time. Changes in the frequency of hostile reference a neurotic makes toward his parents during a sequence of psychotherapeutic interviews, for example, may be detected and assessed, as may the changing self-evaluations of psychiatric hospital inmates in relation to the length of their hospitalization.

Sources of erroneous conclusions that may be drawn from face-to-face encounters stem from the complexity of the interview situation, the attitudes, fears, and expectations of the interviewee, and the interviewer's manner and training. Research has been conducted to identify, control, and, if possible, eliminate these sources of interview invalidity and unreliability. By conducting more than one interview with the same interviewee and by using more than one interviewer to evaluate the subject's behaviour, light can be shed on the reliability of the information derived and may reveal differences in influence among individual interviewers. Standardization of interview format tends to increase the reliability of the information gathered; for example, all interviewers may use the same set of questions. Such standardization, however, may restrict the scope of information elicited, and even a perfectly reliable (consistent) interview technique can lead to incorrect inferences.

Rating scales. The rating scale is one of the oldest and most versatile of assessment techniques. Rating scales present users with an item and ask them to select from a number of choices. The rating scale is similar in some respects to a multiple choice test, but its options represent degrees of a particular characteristic.

Uses of rating scales

Rating scales are used by observers and also by individuals for self-reporting (see below). They permit convenient characterization of other people and their behaviour. Some observations do not lend themselves to quantification as readily as do simple counts of motor behaviour (such as the number of times a worker leaves his lathe to go to the restroom). It is difficult, for example, to quantify how charming an office receptionist is. In such cases, one may fall back on relatively subjective judgments, inferences, and relatively imprecise estimates, as in deciding how disrespectful a child is. The rating scale is one approach to securing such judgments. Rating scales present an observer with scalar dimensions along which those who are observed are to be placed. A teacher, for example, might be asked to rate students on the degree to which the behaviour of each reflects leadership capacity, shyness, or creativity. Peers might rate each other along dimensions such as friendliness, trustworthiness, and social skills. Several standardized, printed rating scales are available for describing the behaviour of psychiatric hospital patients. Relatively objective rating scales have also been devised for use with other groups. Rating scales often take a graphic form:

To what degree is John shy?

not at all	slightly	moderately	very	extremely

A number of requirements should be met to maximize the usefulness of rating scales. One is that they be reliable: the ratings of the same person by different observers should be consistent. Other requirements are reduction of sources of inaccuracy in personality measurement; the so-called halo effect results in an observer's rating someone favourably on a specific characteristic because the observer has a generally favourable reaction to the person being rated. One's tendency to say only nice things about others or one's proneness to think of all people as average (to use the midrange of scales) represents other methodological problems that arise when rating scales are used.

Self-report tests. The success that attended the use of convenient intelligence tests in providing reliable, quantitative (numerical) indexes of individual ability has stimulated interest in the possibility of devising similar tests for measuring personality. Procedures now available vary in the degree to which they achieve score reliability and convenience. These desirable attributes can be partly achieved by restricting in designated ways the kinds of responses a subject is free to make. Self-report instruments follow this strategy. For example, a test that restricts the subject to true-false answers is likely to be convenient to give and easy to score. So-called personality inventories (see below) tend to have these characteristics, in that they are relatively restrictive, can be scored objectively, and are convenient to administer. Other techniques (such as inkblot tests) for evaluating personality possess these characteristics to a lesser degree.

Self-report personality tests

Self-report personality tests are used in clinical settings in making diagnoses, in deciding whether treatment is required, and in planning the treatment to be used. A second major use is as an aid in selecting employees, and a third is in psychological research. An example of the latter case would be where scores on a measure of test anxiety—that is, the feeling of tenseness and worry that people experience before an exam—might be used to divide people into groups according to how upset they get while taking exams. Researchers have investigated whether the more test-anxious students behave differently than the less anxious ones in an experimental situation.

Personality inventories. Among the most common of self-report tests are personality inventories. Their origins lie in the early history of personality measurement, when most tests were constructed on the basis of so-called face validity; that is, they simply appeared to be valid. Items were included simply because, in the fallible judgment of the person who constructed or devised the test, they were indicative of certain personality attributes. In other words, face validity need not be defined by careful, quantitative study; rather, it typically reflects one's more-or-less imprecise, possibly erroneous, impressions. Personal judgment, even that of an expert, is no guarantee that a particular collection of test items will prove to be reliable and meaningful in actual practice.

A widely used early self-report inventory, the so-called Woodworth Personal Data Sheet, was developed during World War I to detect soldiers who were emotionally unfit for combat. Among its ostensibly face-valid items were these: Does the sight of blood make you sick or dizzy? Are you happy most of the time? Do you sometimes wish you had never been born? Recruits who answered these kinds of questions in a way that could be taken to mean that they suffered psychiatric disturbance were detained for further questioning and evaluation. Clearly, however, symptoms revealed by such answers are exhibited by many people who are relatively free of emotional disorder.

Rather than testing general knowledge or specific skills, personality inventories ask people questions about themselves. These questions may take a variety of forms. When taking such a test, the subject might have to decide whether each of a series of statements is accurate as a self-description or respond to a series of true-false questions about personal beliefs.

Several inventories require that each of a series of statements be placed on a rating scale in terms of the frequency or adequacy with which the statements are judged by the individual to reflect his tendencies and attitudes. Regardless of the way in which the subject responds, most inventories yield several scores, each intended to identify a distinctive aspect of personality.

The Minnesota Multiphasic Personality Inventory

One of these, the Minnesota Multiphasic Personality Inventory (MMPI), is probably the personality inventory in widest use in the English-speaking world. Also available in other languages, it consists in one version of 550 items (*e.g.,* "I like tall women") to which subjects are to respond

"true," "false," or "cannot say." Work on this inventory began in the 1930s, when its construction was motivated by the need for a practical, economical means of describing and predicting the behaviour of psychiatric patients. In its development efforts were made to achieve convenience in administration and scoring and to overcome many of the known defects of earlier personality inventories. Varied types of items were included and emphasis was placed on making these printed statements (presented either on small cards or in a booklet) intelligible even to persons with limited reading ability.

Most earlier inventories lacked subtlety; many people were able to fake or bias their answers since the items presented were easily seen to reflect gross disturbances; indeed, in many of these inventories maladaptive tendencies would be reflected in either all true or all false answers. Perhaps the most significant methodological advance to be found in the MMPI was the attempt on the part of its developers to measure tendencies to respond, rather than actual behaviour, and to rely but little on assumptions of face validity. The true-false item "I hear strange voices all the time" has face validity for most people in that to answer "true" to it seems to provide a strong indication of abnormal hallucinatory experiences. But some psychiatric patients who "hear strange voices" can still appreciate the socially undesirable implications of a "true" answer and may therefore try to conceal their abnormality by answering "false." A major difficulty in placing great reliance on face validity in test construction is that the subject may be as aware of the significance of certain responses as is the test constructor and thus may be able to mislead the tester. Nevertheless, the person who hears strange voices and yet answers the item "false" clearly is responding to something—the answer still is a reflection of personality, even though it may not be the aspect of personality to which the item seems to refer; thus, careful study of responses beyond their mere face validity often proves to be profitable.

Reactions to the testing environment

Much study has been given to the ways in which response sets and test-taking attitudes influence behaviour on the MMPI and other personality measures. The response set called acquiescence, for example, refers to one's tendency to respond with "true" or "yes" answers to questionnaire items regardless of what the item content is. It is conceivable that two people might be quite similar in all respects except for their tendency toward acquiescence. This difference in response set can lead to misleadingly different scores on personality tests. One person might be a "yea-sayer" (someone who tends to answer true to test items); another might be a "nay-sayer"; a third individual might not have a pronounced acquiescence tendency in either direction.

Acquiescence is not the only response set; there are other test-taking attiudes that are capable of influencing personality profiles. One of these, already suggested by the example of the person who hears strange voices, is social desirability. A person who has convulsions might say "false" to the item "I have convulsions" because he believes that others will think less of him if they know he has convulsions. The intrusive potentially deceiving effects of the subjects' response sets and test-taking attitudes on scores derived from personality measures can sometimes be circumvented by varying the content and wording of test items. Nevertheless, users of questionnaires have not yet completely solved problems of bias such as those arising from response sets. Indeed, many of these problems first received widespread attention in research on the MMPI, and research on this and similar inventories has significantly advanced understanding of the whole discipline of personality testing.

The MMPI as originally published consists of nine clinical scales (or sets of items), each scale having been found in practice to discriminate a particular clinical group, such as people suffering from schizophrenia, depression, or paranoia (see MENTAL DISORDERS). Each of these scales (or others produced later) was developed by determining patterns of response to the inventory that were observed to be distinctive of groups of individuals who had been psychiatrically classified by other means (*e.g.*, by long-term observation). The responses of apparently normal subjects were compared with those of hospital patients with a particular psychiatric diagnosis—for example, with symptoms of schizophrenia. Items to which the greatest percentage of "normals" gave answers that differed from those more typically given by patients came to constitute each clinical scale.

In addition to the nine clinical scales and many specially developed scales, there are four so-called control scales on the inventory. One of these is simply the number of items placed by the subject in the "cannot say" category. The L (or lie) scale was devised to measure the tendency of the test taker to attribute socially desirable attributes to himself. In response to "I get angry sometimes" he should tend to mark false; extreme L scorers in the other direction appear to be too good, too virtuous. Another so-called F scale was included to provide a reflection of the subjects' carelessness and confusion in taking the inventory (*e.g.*, "Everything tastes the same" tends to be answered true by careless or confused people). More subtle than either the L or F scales is what is called the K scale. Its construction was based on the observation that some persons tend to exaggerate their symptoms because of excessive openness and frankness and may obtain high scores on the clinical scales; others may exhibit unusually low scores because of defensiveness. On the K-scale item "I think nearly anyone would tell a lie to keep out of trouble," the defensive person is apt to answer false, giving the same response to "I certainly feel useless at times." The K scale was designed to reduce these biasing factors; by weighting clinical-scale scores with K scores, the distorting effect of test-taking defensiveness may be reduced.

Types of control scales

In general, it has been found that the greater the number and magnitude of one's unusually high scores on the MMPI, the more likely it is that one is in need of psychiatric attention. Most professionals who use the device refuse to make assumptions about the factualness of the subject's answers and about his personal interpretations of the meanings of the items. Their approach does not depend heavily on theoretical predilections and hypotheses. For this reason the inventory has proved particularly popular with those who have strong doubts about the eventual validity that many theoretical formulations will show in connection with personality measurement after they have been tested through painstaking research. The MMPI also appeals to those who demand firm experimental evidence that any personality assessment method can make valid discriminations among individuals.

In recent years there has been growing interest in actuarial personality description—that is, in personality description based on traits shared in common by groups of people. Actuarial description studies yield rules by which persons may be classified according to their personal attributes as revealed by their behaviour (on tests, for example). Computer programs are now available for diagnosing such disorders as hysteria, schizophrenia, and paranoia on the basis of typical group profiles of MMPI responses. Computerized methods for integrating large amounts of personal data are not limited to this inventory and are applicable to other inventories, personality tests (*e.g.*, inkblots), and life-history information. Computerized classification of MMPI profiles, however, has been explored most intensively.

The MMPI has been considered in some detail here because of its wide usage and because it illustrates a number of important problems confronting those who attempt to assess personality characteristics. Many other omnibus personality inventories are also used in applied settings and in research. The California Psychological Inventory (CPI), for example, is keyed for several personality variables that include sociability, self-control, flexibility, and tolerance. Unlike the MMPI, it was developed specifically for use with "normal" groups of people. Whereas the judgments of experts (usually psychiatric workers) were used in categorizing subjects given the MMPI during the early item-writing phase of its development, nominations by peers (such as respondents or friends) of the subjects were relied upon in work with the CPI. Its technical development has been evaluated by test authorities to be of high

The California Psychological Inventory

order, in part because its developers profited from lessons learned in the construction and use of the MMPI. It also provides measures of response sets and has been subjected to considerable research study.

From time to time, most personality inventories are revised for a variety of reasons, including the need to take account of cultural and social changes and to improve them. For example, a revision of the CPI was published in 1987. In the revision, the inventory itself was modified to improve clarity, update content, and delete items that might be objectionable to some respondents. Because the item pool remained largely unchanged, data from the original samples were used in computing norms and in evaluating reliability and validity for new scales and new composite scores. The descriptions of high and low scorers on each scale have been refined and sharpened, and correlations of scale scores with other personality tests have been reported.

Other self-report techniques. Beyond personality inventories, there are other self-report approaches to personality measurement available for research and applied purposes. Mention was made earlier of the use of rating scales. The rating-scale technique permits quantification of an individual's reactions to himself, to others, and, in fact, to any object or concept in terms of a standard set of semantic (word) polarities such as "hot-cold" or "good-bad." It is a general method for assessing the meanings of these semantic concepts to individuals.

Another method of self-report called the Q-sort is devised for problems similar to those for which rating scales are used. In a Q-sort a person is given a set of sentences, phrases, or words (usually presented individually on cards) and is asked to use them to describe himself (as he thinks he is or as he would like to be) or someone else. This description is carried out by having the subject sort the items on the cards in terms of their degree of relevance so that they can be distributed along what amounts to a rating scale. Examples of descriptive items that might be included in a Q-sort are "worries a lot," "works hard," and "is cheerful."

Typical paper-and-pencil instruments such as personality inventories involve verbal stimuli (words) intended to call forth designated types of responses from the individual. There are clearly stated ground rules under which he makes his responses. Paper-and-pencil devices are relatively easy and economical to administer and can be scored accurately and reliably by relatively inexperienced clerical workers. They are generally regarded by professional personality evaluators as especially valuable assessment tools in screening large numbers of people, as in military or industrial personnel selection. Assessment specialists do not assume that self-reports are accurate indicators of personality traits. They are accepted, rather, as samples of behaviour for which validity in predicting one's everyday activities or traits must be established empirically (*i.e.*, by direct observation or experiment). Paper-and-pencil techniques have moved from their early stage of assumed (face) validity to more advanced notions in which improvements in conceptualization and methodology are clearly recognized as basic to the determination of empirical validity.

Evaluation of self-reports

Projective techniques. One group of assessment specialists believes that the more freedom people have in picking their responses, the more meaningful the description and classification that can be obtained. Because personality inventories do not permit much freedom of choice, some researchers and clinicians prefer to use projective techniques, in which a person is shown ambiguous stimuli (such as shapes or pictures) and asked to interpret them in some way. (Such stimuli allow relative freedom in projecting one's own interests and feelings into them, reacting in any way that seems appropriate.) Projective techniques are believed to be sensitive to unconscious dimensions of personality. Defense mechanisms, latent impulses, and anxieties have all been inferred from data gathered in projective situations.

Personality inventories and projective techniques do have some elements in common; inkblots, for example, are ambiguous, but so also are many of the statements on inventories such as the MMPI. These techniques differ in that the subject is given substantially free rein in responding to projective stimuli rather than merely answering true or false, for example. Another similarity between projective and questionnaire or inventory approaches is that all involve the use of relatively standardized testing situations.

While projective techniques are often lumped together as one general methodology, in actual practice there are several approaches to assessment from a projective point of view. Although projective techniques share the common characteristic that they permit the subject wide latitude in responding, they still may be distinguished broadly as follows: (1) associative techniques, in which the subject is asked to react to words, to inkblots, or to other stimuli with the first associated thoughts that come to mind; (2) construction techniques, in which the subject is asked to create something—for example, make up a story or draw a self-portrait; (3) completion techniques, in which the subject is asked to finish a partially developed stimulus, such as adding the last words to an incomplete sentence; (4) choice or ordering techniques, in which the subject is asked to choose from among or to give some orderly sequence to stimuli—for example, to choose from or arrange a set of pictures or inkblots; (5) expressive techniques, in which the subject is asked to use free expression in some manner, such as in finger painting.

Hidden personality defense mechanisms, latent emotional impulses, and inner anxieties all have been attributed to test takers by making theoretical inferences from data gathered as they responded in projective situations. While projective stimuli are ambiguous, they are usually administered under fairly standardized conditions. Quantitative (numerical) measures can be derived from subjects' responses to them. These include the number of responses one makes to a series of inkblots and the number of responses to the blots in which the subject perceives what seem to him to be moving animals.

Inkblot techniques

The Rorschach Inkblot Test. The Rorschach inkblots were developed by a Swiss psychiatrist, Hermann Rorschach, in an effort to reduce the time required in psychiatric diagnosis. His test consists of 10 cards, half of which are in colour and half in black and white. The test is administered by showing the subject the 10 blots one at a time; the subject's task is to describe what he sees in the blots or what they remind him of. The subject is usually told that the inkblots are not a test of the kind he took in school and that there are no right or wrong answers.

Rorschach's work was stimulated by his interest in the relationship between perception and personality. He held that a person's perceptual responses to inkblots could serve as clues to basic personality tendencies. Despite Rorschach's original claims for the validity of his test, subsequent negative research findings have led many users of projective techniques to become dubious about the role assigned the inkblots in delineating relationships between perception and personality. In recent years, emphasis has tended to shift to the analysis of nuances of the subject's social behaviour during the test and to the content of his verbal responses to the examiner—whether, for example, he seeks to obtain the assistance of the examiner in "solving" the inkblots presented to him, sees "angry lions" or "meek lambs" in the inkblots, or is apologetic or combative about his responses.

Over the years, considerable research has been carried out on Rorschach's inkblots; important statistical problems in analyzing data gathered with projective techniques have been identified, and researchers have continued in their largely unsuccessful efforts to overcome them. There is a vast experimental literature to suggest that the Rorschach technique lacks empirical validity. Recently, researchers have sought to put the Rorschach on a sounder psychometric (mental testing) basis. New comprehensive scoring systems have been developed, and there have been improvements in standardization and norms. These developments have injected new life into the Rorschach as a psychometric instrument.

The Holtzman Inkblot Test

A similar method, the Holtzman Inkblot Test, has been developed in an effort to eliminate some of the statistical problems that beset the Rorschach test. It involves the

administration of a series of 45 inkblots, the subject being permitted to make only one response per card. The Holtzman has the desirable feature that it provides an alternate series of 45 additional cards for use in retesting the same person.

Research with the Rorschach and Holtzman has proceeded in a number of directions; many studies have compared psychiatric patients and other groups of special interest (delinquents, underachieving students) with ostensibly normal people. Some investigators have sought to derive indexes or predictions of future behaviour from responses to inkblots and have checked, for example, to see if anxiety and hostility (as inferred from content analyses of verbal responses) are related to favourable or unfavourable response to psychotherapy. A sizable area of exploration concerns the effects of special conditions (*e.g.,* experimentally induced anxiety or hostility) on the inkblot perceptions reported by the subject and the content of his speech.

Thematic Apperception Test (TAT). There are other personality assessment devices, which, like the Rorschach, are based on the idea that an individual will project something of himself into his description of an ambiguous stimulus.

The TAT, for example, presents the subject with pictures of persons engaged in a variety of activities (*e.g.,* someone with a violin). While the pictures leave much to one's imagination, they are more highly specific, organized visual stimuli than are inkblots. The test consists of 30 black and white pictures and one blank card (to test imagination under very limited stimulation). The cards are presented to the subject one at a time, and he is asked to make up a story that describes each picture and that indicates the events that led to the scene and the events that will grow out of it. He is also asked to describe the thoughts and feelings of the persons in his story.

Although some content-analysis scoring systems have been developed for the TAT, attempts to score it in a standardized quantitative fashion tend to be limited to research and have been fewer than has been the case for the Rorschach. This is especially the state of affairs in applied settings in which the test is often used as a basis for conducting a kind of clinical interview; the pictures are used to elicit a sample of verbal behaviour on the basis of which inferences are drawn by the clinician.

In one popular approach, interpretation of a TAT story usually begins with an effort to determine who is the hero (*i.e.,* to identify the character with whom the subject seems to have identified himself). The content of the stories is often analyzed in terms of a so-called need-press system. Needs are defined as the internal motivations of the hero. Press refers to environmental forces that may facilitate or interfere with the satisfaction of needs (*e.g.,* in the story the hero may be physically attacked, frustrated by poverty, or suffer the effects of rumours being spread about him). In assessing the importance or strength of a particular inferred need or press for the individual who takes the test, special attention is given to signs of its pervasiveness and consistency in different stories. Analysis of the test may depend considerably on the subjective, personal characteristics of the evaluator, who usually seeks to interpret the subjects' behaviour in the testing situation; the characteristics of his utterances; the emotional tone of the stories; the kinds of fantasies he offers; the outcomes of the stories; and the conscious and unconscious needs speculatively inferred from the stories.

Word-association techniques. The list of projective approaches to personality assessment is long, one of the most venerable being the so-called word-association test. Jung used associations to groups of related words as a basis for inferring personality traits (*e.g.,* the inferiority "complex"). Administering a word-association test is relatively uncomplicated; a list of words is presented one at a time to the subject who is asked to respond with the first word or idea that comes to mind. Many of the stimulus words may appear to be emotionally neutral (*e.g.,* building, first, tree); of special interest are words that tend to elicit personalized reactions (*e.g.,* mother, hit, love). The amount of time the subject takes before beginning each response and the response itself are used in efforts to analyze a word association test. The idiosyncratic, or unusual, nature of one's word-association responses may be gauged by comparing them to standard published tables of the specific associations given by large groups of other people.

Word stimulus responses

Sentence-completion techniques. The sentence-completion technique may be considered a logical extension of word-association methods. In administering a sentence-completion test, the evaluator presents the subject with a series of partial sentences that he is asked to finish in his own words (*e.g.,* "I feel upset when . . . "; "What burns me up is . . . "). Users of sentence-completion methods in assessing personality typically analyze them in terms of what they judge to be recurring attitudes, conflicts, and motives reflected in them. Such analyses, like those of TAT, contain a subjective element.

Behavioral assessment. Objective observation of a subject's behaviour is a technique that falls in the category of behavioral assessment. A variety of assessments could be considered, for example, in the case of a seven-year-old boy who, according to his teacher, is doing poorly in his schoolwork and, according to his parents, is difficult to manage at home and does not get along with other children. The following types of assessment might be considered: (1) a measure of the boy's general intelligence, which might help explain his poor schoolwork; (2) an interview with him to provide insights into his view of his problem; (3) personality tests, which might reveal trends that are related to his inadequate social relationships; (4) observations of his activities and response patterns in school; (5) observations of his behaviour in a specially created situation, such as a playroom with many interesting toys and games; (6) an interview with his parents, since the boy's poor behaviour in school may by symptomatic of problems at home; and (7) direct observation of his behaviour at home.

Types of assessment

Making all of these assessments would be a major undertaking. Because of the variety of data that are potentially available, the assessor must decide which types of information are most feasible and desirable under a given set of circumstances. In most cases, the clinician is interested in both subjective and objective information. Subjective information includes what clients think about, the emotions they experience, and their worries and preoccupations. Interviews, personality inventories, and projective techniques provide indications of subjective experience, although considerable clinical judgment is needed to infer what is going on within the client from test responses. Objective information includes the person's observable behaviour and usually does not require the assessor to draw complex inferences about such topics as attitudes toward parents, unconscious wishes, and deep-seated conflicts. Such objective information is measured by behavioral assessment. It is often used to identify behavioral problems, which are then treated in some appropriate way. Behavioral observations are used to get information that cannot be obtained by other means. Examples of such observations include the frequency of a particular type of response, such as physical attacks on others or observations by ward attendants of certain behaviours of psychiatric patients. In either case, observational data must meet the same standards of reliability as data obtained by more formal measures.

The value of behavioral assessment depends on the behaviours selected for observation. For example, if the goal of assessment is to detect a tendency toward depression, the responses recorded should be those that are relevant to that tendency, such as degrees of smiling, motor activity, and talking.

Baseline observations

A type of behavioral assessment called baseline observations is becoming increasingly popular. These are recordings of response frequencies in particular situations before any treatment or intervention has been made. They can be used in several ways. Observations might be made simply to describe a person's response repertoire at a given time. For example, the number of aggressive responses made by children of different ages might be recorded. Such observations also provide a baseline for judging the effectiveness of behaviour modification techniques. A similar

set of observations, made after behaviour modification procedures have been used, could be compared with the baseline measurement as a way of determining how well the therapy worked.

Behavioral observations can be treated in different ways. One of these is to keep track of the frequency with which people make designated responses during a given period of time (*e.g.,* the number of times a psychiatric patient makes his own bed or the number of times a child asks for help in a novel situation). Another approach involves asking raters to support their judgments of others by citing specific behaviour (critical incidents); a shop foreman, for example, may rate a worker as depressed by citing incidents when the worker burst into tears. Critical incidents not only add validity to ordinary ratings, but they also suggest behavioral details that might be promising predictors of success on the job, response to psychiatric treatment, or level of academic achievement.

Behavioral observations are widely made in interviews and in a variety of workaday settings. Employers, supervisors, and teachers—either formally or informally—make use of behavioral observations in making decisions about people for whom they have responsibility. Unfortunately the subject may know he is being studied or evaluated and, therefore, may behave atypically (*e.g.,* by working harder than usual or by growing tense). The observer may be a source of error by being biased in favour of or against the subject. Disinterested observers clearly are to be preferred (other things being equal) for research and clinical purposes. The greater the care taken to control such contributions to error, the greater the likelihood that observations will prove to be reliable.

Cognitive assessment. The types of thoughts experienced by individuals are reflective of their personalities. Just as it is important to know what people do and how their behaviour affects others, it is also necessary to assess the thoughts that may lie behind the behaviour. Cognitive assessment provides information about thoughts that precede, accompany, and follow maladaptive behaviour. It also provides information about the effects of procedures that are intended to modify both how subjects think about a problem and how they behave.

Cognitive assessment can be carried out in a variety of ways. For example, questionnaires have been developed to sample people's thoughts after an upsetting event. Beepers (electronic pagers) have been used to signal subjects to record their thoughts at certain times of the day. There are also questionnaires to assess the directions people give themselves while working on a task and their theories about why things happen as they do.

Methods of recording people's thoughts

The assessment of thoughts and ideas is a relatively new development. It has received impetus from the growing evidence that thought processes and the content of thoughts are related to emotions and behaviour. Cognitive assessment provides information about adaptive and maladaptive aspects of people's thoughts and the role their thoughts play in the processes of planning, making decisions, and interpreting reality.

Bodily assessment. Bodily responses may reveal a person's feelings and motivations, and clinicians pay particular attention to these nonverbal messages. Bodily functions may also reflect motivations and concerns, and some clinicians also pay attention to these. Sophisticated devices have been developed to measure such physiological changes as pupil dilation, blood pressure, and electrical skin responses under specific conditions. These changes are related to periodic ratings of mood and to other physiological states that provide measures of stability and change within the individual. Technological advances are making it possible to monitor an individual's physiological state on a continuous basis. Sweat, heartbeat, blood volume, substances in the bloodstream, and blood pressure can all be recorded and correlated with the presence or absence of certain psychological conditions such as stress.

Techniques for gathering personal information

Personal facts. One type of information that is sometimes overlooked because of its very simplicity consists of the subject's life history and present status. Much of this information may be gathered through direct interviews with a subject or with an informant through questionnaires and through searches of records and archives. The information might also be gathered by examining the subject's personal documents (*e.g.,* letters, autobiographies) and medical, educational, or psychiatric case histories. The information might concern the individual's social and occupational history, his cultural background, his present economic status, and his past and present physical characteristics. Life-history data can provide clues to the precursors and correlates of present behaviour. This information may help the investigator avoid needlessly speculative or complex hypotheses about the causation of personality traits when simple explanations might be superior. Failure on the part of a personality evaluator to be aware of the fact that someone had spent two years during World War II in a concentration camp could result in misleading inferences and conjectures about the subject's present behaviour.

RELIABILITY AND VALIDITY OF ASSESSMENT METHODS

Assessment, whether it is carried out with interviews, behavioral observations, physiological measures, or tests, is intended to permit the evaluator to make meaningful, valid, and reliable statements about individuals. What makes John Doe tick? What makes Mary Doe the unique individual that she is? Whether these questions can be answered depends upon the reliability and validity of the assessment methods used. The fact that a test is intended to measure a particular attribute is in no way a guarantee that it really accomplishes this goal. Assessment techniques must themselves be assessed.

Evaluation techniques. Personality instruments measure samples of behaviour. Their evaluation involves primarily the determination of reliability and validity. Reliability often refers to consistency of scores obtained by the same persons when retested. Validity provides a check on how well the test fulfills its function. The determination of validity usually requires independent, external criteria of whatever the test is designed to measure. An objective of research in personality measurement is to delineate the conditions under which the methods do or do not make trustworthy descriptive and predictive contributions. One approach to this problem is to compare groups of people known through careful observation to differ in a particular way. It is helpful to consider, for example, whether the MMPI or TAT discriminates significantly between those who show progress in psychotherapy and those who do not, whether they distinguish between law violators of record and apparent nonviolators. Experimental investigations that systematically vary the conditions under which subjects perform also make contributions.

Although much progress has been made in efforts to measure personality, all available instruments and methods have defects and limitations that must be borne in mind when using them; responses to tests or interview questions, for example, often are easily controlled or manipulated by the subject and thus are readily "fakeable." Some tests, while useful as group screening devices, exhibit only limited predictive value in individual cases, yielding frequent (sometimes tragic) errors. These caveats are especially poignant when significant decisions about people are made on the basis of their personality measures. Institutionalization or discharge, and hiring or firing, are weighty personal matters and can wreak great injustice when based on faulty assessment. In addition, many personality assessment techniques require the probing of private areas of the individual's thought and action. Those who seek to measure personality for descriptive and predictive reasons must concern themselves with the ethical and legal implications of their work.

Limitations in personality measurement methods

A major methodological stumbling block in the way of establishing the validity of any method of personality measurement is that there always is an element of subjective judgment in selecting or formulating criteria against which measures may be validated. This is not so serious a problem when popular, socially valued, fairly obvious criteria are available that permit ready comparisons between such groups as convicted criminals and ostensible noncriminals, or psychiatric hospital patients and noninstitutionalized individuals. Many personality characteristics, however, cannot be validated in such directly observable

ways (*e.g.,* inner, private experiences such as anxiety or depression). When such straightforward empirical validation of an untested measure hopefully designed to measure any personality attribute is not possible, efforts at establishing a less impressive kind of validity (so-called construct validity) may be pursued. A construct is a theoretical statement concerning some underlying, unobservable aspect of an individual's characteristics or of his internal state. ("Intelligence," for example, is a construct; one cannot hold "it" in one's hand, or weigh "it," or put "it" in a bag, or even look at "it.") Constructs thus refer to private events inferred or imagined to contribute to the shaping of specific public events (observed behaviour). The explanatory value of any construct has been considered by some theorists to represent its validity. Construct validity, therefore, refers to evidence that endorses the usefulness of a theoretical conception of personality. A test designed to measure an unobservable construct (such as "intelligence" or "need to achieve") is said to accrue construct validity if it usefully predicts the kinds of empirical criteria one would expect it to—*e.g.,* achievement in academic subjects.

Predictiveness of personality measures

The degree to which a measure of personality is empirically related to or predictive of any aspect of behaviour observed independently of that measure contributes to its validity in general. A most desirable step in establishing the usefulness of a measure is called cross-validation. The mere fact that one research study yields positive evidence of validity is no guarantee that the measure will work as well the next time; indeed, often it does not. It is thus important to conduct additional, cross-validation studies to establish the stability of the results obtained in the first investigation. Failure to cross-validate is viewed by most testing authorities as a serious omission in the validation process. Evidence for the validity of a full test should not be sought from the same sample of people that was used for the initial selection of individual test items. Clearly this will tend to exaggerate the effect of traits that are unique to that particular sample of people and can lead to spuriously high (unrealistic) estimates of validity that will not be borne out when other people are studied. Cross-validation studies permit assessment of the amount of "shrinkage" in empirical effectiveness when a new sample of subjects is employed. When evidence of validity holds up under cross-validation, confidence in the general usefulness of test norms and research findings is enhanced. Establishment of reliability, validity, and cross-validation are major steps in determining the usefulness of any psychological test (including personality measures).

Clinical versus statistical prediction. Another measure of assessment research has to do with the role of the assessor himself as an evaluator and predictor of the behaviour of others. In most applied settings he subjectively (even intuitively) weighs, evaluates, and interprets the various assessment data that are available. How successful he is in carrying out his interpretive task is critical, as is knowledge of the kinds of conditions under which he is effective in processing such diverse data as impressions gathered in an interview, test scores, and life-history data. The typical clinician usually does not use a statistical formula that weighs and combines test scores and other data at his disposal. Rather, he integrates the data using impressions and hunches based on his past clinical experience and on his understanding of psychological theory and research. The result of this interpretive process usually includes some form of personality description of the person under study and specific predictions or advice for that person.

Evaluating the evaluator

The degree of success an assessor has when he responds to the diverse information that may be available about a particular person is the subject of research that has been carried out on the issue of clinical versus statistical prediction. It is reasonable to ask whether a clinician will do as good a job in predicting behaviour as does a statistical formula or "cookbook"—*i.e.,* a manual that provides the empirical, statistically predictive aspects of test responses or scores based on the study of large numbers of people.

An example would be a book or table of typical intelligence test norms (typical scores) used to predict how well children perform in school. Another book might offer specific personality diagnoses (*e.g.,* neurotic or psychotic) based on scores such as those yielded by the different scales of the MMPI. Many issues must be settled before the deceptively simple question of clinical versus statistical prediction can be answered definitively.

When statistical prediction formulas (well-supported by research) are available for combining clinical information, however, experimental evidence clearly indicates that they will be more valid and less time-consuming than will a clinician (who may be subject to human error in trying to simultaneously consider and weigh all of the factors in a given case). The clinician's chief contributions to diagnosis and prediction are in situations for which satisfactory formulas and quantified information (*e.g.,* test scores) are not available. A clinician's work is especially important when evaluations are required for rare and idiosyncratic personality characteristics that have escaped rigorous, systematic empirical study. The greatest confidence results when both statistical and subjective clinical methods simultaneously converge (agree) in the solution of specific clinical problems. (I.G.S.)

BIBLIOGRAPHY. Definitions of personality as a psychological concept, with treatment of related issues, are found in the following reference works: ROM HARRÉ and ROGER LAMB (eds.), *The Dictionary of Personality and Social Psychology* (1986); RICHARD L. GREGORY (ed.), *The Oxford Companion to the Mind* (1987); RAYMOND J. CORSINI *et al.* (eds.), *Concise Encyclopedia of Psychology* (1987); and JONATHAN L. FREEDMAN, *Introductory Psychology,* 2nd ed. (1982).

Major theories of personality are surveyed in CALVIN S. HALL and GARDNER LINDZEY, *Theories of Personality,* 3rd ed. (1978); and in NATHAN BRODY, *Personality: In Search of Individuality* (1988). ROGER BROWN, *Social Psychology, the Second Edition* (1986); and IRWIN G. SARASON, *Personality: An Objective Approach,* 2nd ed. (1972), review and evaluate contemporary personality studies and research methods. The best presentation of psychoanalytic ideas remains Freud's own, available in many translated selections. The following two, edited and translated by JAMES STRACHEY, can be recommended: SIGMUND FREUD, *Introductory Lectures on Psychoanalysis* (1977; originally published in German, 1916–17), and *New Introductory Lectures on Psychoanalysis* (1974; originally published in German, 1933). GARDNER MURPHY, *Personality: A Biosocial Approach to Origins and Structure* (1947, reissued 1966), integrates views and data that reach across historical epochs and many intellectual disciplines. The theory of cognitive controls and styles is carefully analyzed in GEORGE S. KLEIN, *Perception, Motives, and Personality* (1970).

Discussions of the variety of currently held views on personality, demonstrating a broad interdisciplinary approach, include JAMES F. MASTERSON, *The Search for the Real Self: Unmasking the Personality Disorders of Our Age* (1988); and ARTHUR PEACOCKE and GRANT GILLETT (eds.), *Persons and Personality: Contemporary Inquiry* (1987). Adaptation of personality in social interaction is explored in JOEL ARONOFF and JOHN P. WILSON, *Personality in the Social Process* (1985); ALAN S. WATERMAN, *The Psychology of Individualism* (1984); ROBERT A. LEVINE, *Culture, Behavior, and Personality: An Introduction to the Comparative Study of Psychosocial Adaptation,* 2nd ed. (1982); NANCY CANTOR and JOHN F. KIHLSTROM, *Personality, Cognition, and Social Interaction* (1981); and NANCY CANTOR and JOHN F. KIHLSTROM (eds.), *Personality and Social Intelligence* (1987). Issues of personality development from childhood through adulthood are interpreted in ROBERT L. LEAHY (ed.), *The Development of the Self* (1985); LAURENCE R. SIMON, *Cognition and Affect: A Developmental Psychology of the Individual* (1986); LAWRENCE S. WRIGHTSMAN, *Personality Development in Adulthood* (1988); and ERIK H. ERIKSON, *Childhood and Society,* new anniversary ed. (1985).

General approaches to personality assessment are surveyed in ANNE ANASTASI, *Psychological Testing,* 6th ed. (1988); and LEE J. CRONBACH, *Essentials of Psychological Testing,* 4th ed. (1984). JAMES V. MITCHELL, JR. (ed.), *The Ninth Mental Measurement Yearbook,* 2 vol. (1985), describes numerous psychometric tests, including personality assessment instruments. DOUGLAS N. JACKSON and SAMUEL MESSICK (eds.), *Problems in Human Assessment* (1967, reprinted 1978), deals with selected aspects of personality assessment. E. JERRY PHARES, *Introduction to Personality,* 2nd ed. (1988), reviews the personality field with a focus on assessment; and IRWIN G. SARASON and BARBARA R. SARASON, *Abnormal Psychology: The Problem of Maladaptive Behavior,* 6th ed. (1989), provides examples of personality assessment of individuals who display deviant behaviour.

(P.S.H./I.G.S.)

Peru

A country that borders the Pacific Ocean, straddles the great Andes Mountain Ranges, and contains the headwaters of the Amazon River, Peru is the third largest nation in South America after Brazil and Argentina. Officially it is called the Republic of Peru (República del Perú). Its name is derived from a Quechua Indian word implying land of abundance, a reference to the economic wealth produced by the highly organized Inca civilization that ruled the region for centuries. The nation's vast mineral, agricultural, and marine resources long have served as the economic foundation of the country.

Except for the Lake Titicaca basin in the southeast, Peru's borders run through sparsely populated zones. The boundaries with Ecuador to the northwest, Bolivia to the southeast, and Chile to the south run across the high Andes, whereas the borders with Colombia to the northeast and Brazil to the east traverse lower ranges or tropical forests. Peru's land area of 496,225 square miles (1,285,216 square kilometres) is supplemented by territorial waters, reaching 200 miles (320 kilometres) into the Pacific, that are claimed by Peru.

Peru is essentially a tropical country, with its northern tip nearly touching the Equator. Despite its tropical location, a great diversity of climate, of way of life, and of economic activity is brought about by the country's extremes of altitude and by the southwest winds that sweep in across the cold Peru, or Humboldt, Current, which flows along its Pacific shoreline. The immense difficulties of travel posed by the Andes have long impeded national unity. Iquitos, on the upper Amazon, lies only about 600 miles northeast of Lima, the capital, but before the airplane, travelers between the cities often chose a 7,000-mile trip via the Amazon, the Atlantic and Caribbean, the Isthmus of Panama, and the Pacific, rather than the shorter mountain route.

This article is divided into the following sections:

Physical and human geography 515
- The land 515
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 519
  - Pre-Hispanic groups
  - Ethnic diversity
  - Religion
  - Demography
- The economy 520
  - Resources
  - Agriculture
  - Fishing and forestry
  - Industry
  - Finance
  - Trade
  - Transportation
- Administration and social conditions 521
  - Government
  - Education
  - Health and welfare
- Cultural life 522
  - Heritage
  - The arts
  - Cultural institutions
  - Recreation
  - Press and broadcasting

History 523
- The Inca 523
- Discovery and exploration by Europeans 523
- Colonial period 523
- Achievement of independence 524
- Peru from 1824 to 1884 524
  - Struggle for power
  - Establishment of order
  - The War of the Pacific (1879–83)
- Peru from 1884 to 1930 525
  - Social reforms and economic development
  - Formation of the Aprista movement
- Peru from 1930 to 1968 525
  - Troubled democracy
  - The dictatorship of Manuel Odría
  - Return to elected government
- Military rule (1968–80) 526
  - Economic nationalism
  - New junta
- Return to civilian rule 526

Bibliography 527

Physical and human geography

THE LAND

Relief. Peru is traditionally described in terms of three broad, longitudinal regions: the arid Costa on the west; the rugged Sierra, or Andes, system in the centre; and, on the east, the wet and forested Montaña of the tropical Amazon Basin.

The Costa. The coastal plain can be readily divided into three parts—north, central, and south—based on the amount of level land and the distance between the Andean ranges and the sea. Generally speaking, the amount of level coastal land diminishes from north to south. In the northern region, from Ecuador to Chimbote, the plain is typically some 20 to 30 miles wide, with a maximum width of more than 90 miles in the Sechura Desert south of Piura. The central coastal region, which stretches from Chimbote to Nazca, is narrower than the northern region and is characterized by areas of rough hills that extend from the Andes to the shores of the ocean. From Nazca southward to the Chilean border the coast is for the most part lined by low mountains; the southern valleys are narrow and only in scattered spots are level lands found near the ocean.

The Sierra, or Andean, region. Along the western edge of South America, the Andes Mountains were created by tectonic activity in which the South American Plate overrode the Nazca Plate. The Peruvian Andes are typical of mountain regions of the Pacific Rim: they are young in geologic terms, and their continuing uplift is manifested by frequent earthquakes and much instability. Three main backbones protrude from the Peruvian Andes; they are commonly called the cordilleras Occidental, Central, and Oriental, although these designations are not used within Peru.

Slopes are relatively gentle in northern Peru, and maximum elevations seldom exceed 16,000 feet (about 5,000 metres). The Andes in central Peru are higher and more rugged. The ranges of the central zone form particularly difficult barriers to movement. The main pass east of Lima, for instance, is at an elevation of more than 15,000 feet—higher than many of the peaks in the north. Many of the mountains of central Peru are snowcapped and are a popular attraction for climbers and tourists. Of particular fame is the Cordillera Blanca, with the country's highest peak, Mount Huascarán, at 22,205 feet. In southern Peru the character of the Andes changes to that of a high plateau region; this is the Puna, with vast tablelands and elevations between 13,000 and 16,000 feet. Scattered peaks, with elevations of up to about 21,000 feet, protrude

Mount Huascarán

above the broad southern plateaus. Beginning with Misti Volcano, near Arequipa, many of the southern peaks form a volcanic chain that stretches into northern Chile.

The Montaña. The lower slopes of the western Andes merge with the heavily forested tropical lowlands of the Amazon Basin to form the Montaña, which occupies more than three-fifths of the area of Peru. The term *selva* is also used to refer to this area. The physiography of the region is characterized by rolling hills and level plains that extend eastward to the borders with Colombia, Brazil, and Bolivia. Elevations are uniformly low, ranging from about 1,300 feet at the eastern edge of the Andes to about 300 feet above sea level along the Amazon River at the Peru–Brazil border.

Drainage. Distinctive drainage patterns dissect the Costa, Sierra, and Montaña. Of the more than 50 rivers that flow west from the Andes across the Costa, most are short (usually less than 200 miles long) and precipitous, with highly seasonal rates of flow. Most have a period of peak flow (usually during the December to March rainy season) followed by a long dry period; only the largest of the Costa rivers, such as the Santa, have dependable year-round flows.

The Sierra not only contains the headwaters of the streams that flow to both the Pacific and the Amazon but also has a large area of internal drainage. In the south several rivers cross the altiplano in Peru to empty into Lake Titicaca, which is shared with Bolivia and is—at an altitude of 12,500 feet—the world's highest navigable body of water.

The Amazon River

The Montaña region is characterized by great rivers. The Amazon, with the largest volume of flow of any river in the world, has headwaters that rise in several places in the Peruvian Andes; one of the main branches, the Ucayali, originates in southern Peru some 1,700 miles from its juncture with the main branch. The Amazon is navigable, but such large tributaries as the Marañón, Huallaga, and Ucayali can be navigated only for relatively short distances west of the port of Iquitos. These rivers flow northward in long deep valleys before turning east to join the Amazon, forming mostly hindrances to transportation rather than important trade routes.

Soils. Peru has little fertile soil. In the Costa region most of the river valleys have rich soils, derived from silts carried to the coastal plain by rivers flowing out of the Andes. In some areas, however, improper use of the land has led to deposition of salts, thus reducing soil fertility. The soils between valleys, derived largely from windblown sands, are also poorly developed. Sierra soils are fertile in some of the highland basins, but soils on the mountain slopes are often thin and of poor quality. Soils of low fertility covered by heavy forest growth typify the Montaña; they quickly erode when the forests are cut.

Climate. Three broad climatic regions can be readily distinguished in Peru paralleling the three main topographic regions: the Costa, the Sierra, and the Montaña.

Coastal aridity

The coastal desert. From the Peruvian–Ecuadorian border south to northern Chile, the west coast of South America has one of the Earth's driest climates. This region is dry for three reasons: (1) The Andes block rain-bearing winds from the Amazon Basin, (2) air masses moving toward the coast out of the South Pacific high pressure system produce little rainfall, and (3) northward-flowing cold water off the coast (the Peru, or Humboldt, Current) contributes little moisture to surface air masses. This is not a hot desert, however; average temperatures of the Costa range from 66° F (19° C) in winter to 72° F (22° C) in summer. Despite its dryness some parts of the Costa receive sufficient moisture from winter fogs (locally known as *garúa*) to support some vegetation.

The mountain climates. Within the Sierra are a wide range of climates that vary according to such factors as latitude, altitude, local winds, and rain shadow effects. In general, temperatures decrease as elevation increases and rainfall decreases from north to south and from east to west. During the December–March rainy season, the heaviest precipitation is in the north and along the eastern flanks of the Andes. Temperatures vary little seasonally, but there is a tremendous diurnal range (between daily highs and lows). In Cuzco at 11,152 feet of altitude, for example, the January average temperature is 52° F (11° C), and the July average 47° F (8° C). The diurnal range, however, is frequently more than 40° F (22° C) between the midday maximum and the predawn minimum. Snow falls in the Sierra at higher altitudes, and many peaks have permanent snow.

The tropical forest climates. Hot humid conditions characterize the Montaña climate of eastern Peru. Rainfall throughout the region is high (Iquitos averages more than 90 inches [2,286 millimetres] annually), with precipitation common throughout the year, although it is somewhat heavier from December to March. There is little seasonal variation of temperatures, but the diurnal range again is relatively large. Daytime highs at Iquitos sometimes extend into the mid-90's F (mid-30's C), whereas at night temperatures may fall into the 60's F (below 15° C).

El Niño. The most severe variation in Peruvian weather patterns occurs irregularly, at intervals of about a decade or so. This change, usually called El Niño ("The Child," because it usually begins around Christmas time), is but a small part of what is known as the Southern Oscillation, a pan-Pacific reversal of atmospheric and sea conditions. Although the causes of this phenomenon are not completely understood, the effects in Peru are quite clear: (1) warm water replaces the cold water of the Peru Current, (2) heavy rains fall in the coastal desert, and (3) drought occurs in the southern highlands. Severe occurrences of the El Niño phenomenon, such as those that took place in 1925 and 1983, cause ecological disasters, including widespread loss of bird and fish life and great economic damage.

Plant and animal life. Peruvian plant and animal life can be classified according to the three main physiographic regions: the Costa, the Sierra, and the Montaña.

The Costa. Evidence of plant life is relatively rare in the barren desert of coastal Peru. Where coastal fog is heavy, *lomas* (a mix of grasses and other herbaceous species) are common. In the north coast region, some parts of the desert are covered by epiphytes or by stands of sapote or algarroba (mesquite). The most important feature of the coast, however, is the enormous amount of bird, marine mammal, and fish life that abounds in the coastal waters. The biomass includes such small fish as anchovies and such larger types as corbina (sea bass), tuna, swordfish, and marlin. Sea lions thrive in isolated parts of the coast. Bird life is heavy on islands off the coast. Among the most important bird species are pelicans, cormorants, gannets, and various gulls. Humboldt penguins, an endangered species, are found as far north as the Ballestas Islands near the Paracas Peninsula.

The Sierra. Two plant communities characterize the Peruvian highlands: puna grasslands at elevations from about 13,000 to 16,000 feet, and, at lower elevations, a mixture of native and introduced species. The Puna has an abundance of forage grasses and is home to the llama, alpaca, vicuña, and guanaco, which are native to the region. At lower elevations grow such domesticates as potatoes, quinoa, and corn (maize). Several species of eucalyptus have replaced native tree species.

Amazonian wildlife

The Montaña. The eastern slopes of the Andes and the Amazon plains are covered by a heavy growth of tropical forest. In its woods and waters live thousands of plant and animal species. Interesting mammals of this region include the jaguar, capybara, tapir, and several species of monkey. Of special note is the wide and colourful variety of bird and fish life. Reptiles and insects abound. The forests have a broad assortment of hardwood and softwood species that produce a variety of forest products. Scattered in isolated fields in the eastern foothills of the Andes, too, are plantations of coca, the plant from which cocaine is illegally produced.

Settlement patterns. The nature of Peruvian life, whether urban or rural, varies by physiographic region. Modern patterns of settlement also reflect three major influences: (1) pan-Andean cultures of pre-Hispanic Peru; (2) colonial settlement of the Costa and the Sierra; and (3) migration to the cities and colonization of the Montaña.

Pre-Hispanic patterns. Diverse groups of indigenous In-

The early migrants

dians occupied Peru during the pre-Hispanic period. When the first migrants arrived in the Andean area, probably more than 13,000 years ago, they were at a hunting and gathering stage of cultural development. Over a long period of time, however, varied and more sophisticated ways of life were developed. Along the coast, groups became specialized in fishing and shellfish collecting. In the Puna, hunting of vicuña and guanaco was replaced by herding of their related species, the llama and alpaca. Finally, in many parts of Peru agriculture was developed—including the domestication of numerous species of plants, such as beans, quinoa, and potatoes.

At the time of the Spanish arrival, the population of Peru was largely resident in rural areas, with society organized around villages (called by the Incas *ayllus*). The most densely settled areas were the coastal river valleys and some fertile basins in the highlands—for example, those of Cajamarca, the Mantaro Valley near Huancayo, and Cusco, as well as the region around Lake Titicaca. Some urban centres had developed as the capitals of kingdoms or empires—such as the Chimú's Chan Chán near Trujillo and the Inca's Cuzco—or as religious centres—such as the pre-Incan Pachacamac, south of Lima.

The Spanish influence on settlement

Colonial patterns. The Spanish conquest of the Incas in 1532 was accompanied by several dramatic changes in Andean settlement patterns. First, the Spanish were oriented toward their mother country in Europe. Thus, Spanish cities such as Piura (1532), Lima (1535), and Trujillo (1534) were established near ports that were the sea links to Spain. Second, Spanish settlements focused on the extraction of resources, leading to the establishment of mining centres in Huancavelica and at Potosí, in modern Bolivia. Third, after a period of rapid population decline caused mainly by the introduction of European diseases, the Spanish established new towns that brought together the remnants of the surviving rural population. Finally, the Spanish divided the country into encomiendas, which later formed the basis for haciendas and kept the best agricultural land in the hands of a few wealthy owners. They established feudal systems based on peasant labour that lasted until the sweeping land reforms of the mid-20th century.

MAP INDEX

Political subdivisions

Amazonas ..... 05 00 s 78 00 w
Ancash ..... 09 30 s 77 45 w
Apurímac ..... 14 00 s 73 00 w
Arequipa ..... 16 00 s 72 15 w
Ayacucho ..... 14 00 s 74 00 w
Cajamarca ..... 06 15 s 78 50 w
Callao ..... 12 04 s 77 09 w
Cuzco ..... 12 30 s 72 30 w
Huancavelica ... 13 00 s 75 00 w
Huánuco ..... 09 30 s 75 50 w
Ica ..... 14 20 s 75 30 w
Junín ..... 11 30 s 75 00 w
La Libertad ..... 08 00 s 78 30 w
Lambayeque .... 06 20 s 80 00 w
Lima ..... 12 00 s 76 35 w
Loreto ..... 05 00 s 75 00 w
Madre de Dios .. 12 00 s 70 15 w
Moquegua ..... 16 50 s 70 55 w
Pasco ..... 10 30 s 75 15 w
Piura ..... 05 00 s 80 20 w
Puno ..... 15 00 s 70 00 w
San Martín ..... 07 00 s 76 50 w
Tacna ..... 17 40 s 70 20 w
Tumbes ..... 03 50 s 80 30 w
Ucayali ..... 09 30 s 73 20 w

Cities and towns

Abancay ..... 13 35 s 72 55 w
Acomayo ..... 13 55 s 71 41 w
Ancón ..... 11 47 s 77 11 w
Andahuaylas .... 13 39 s 73 23 w
Aplao ..... 16 05 s 72 31 w
Arequipa ..... 16 24 s 71 33 w
Ascope ..... 07 43 s 79 07 w
Ayabaca ..... 04 38 s 79 43 w
Ayacucho ..... 13 07 s 74 13 w
Ayaviri ..... 14 52 s 70 35 w
Bagua ..... 05 40 s 78 31 w
Bambamarca .... 06 41 s 78 32 w
Barranca ..... 10 45 s 77 46 w
Bolívar ..... 07 18 s 77 48 w
Cajabamba ..... 07 37 s 78 03 w
Cajamarca ..... 07 10 s 78 31 w
Calca ..... 13 20 s 71 57 w
Callao ..... 12 04 s 77 09 w
Camaná ..... 16 37 s 72 42 w
Canta ..... 11 25 s 76 38 w
Caraveli ..... 15 46 s 73 22 w
Caraz ..... 09 04 s 77 49 w
Casma ..... 09 28 s 78 19 w
Castilla ..... 05 12 s 80 38 w
Celendín ..... 06 52 s 78 09 w
Cerro de Pasco .. 10 41 s 76 16 w
Chachapoyas ... 06 12 s 77 51 w
Chaupimarca ... 10 26 s 76 32 w
Chicla ..... 11 45 s 76 18 w
Chiclayo ..... 06 46 s 79 51 w
Chimbote ..... 09 05 s 78 36 w
Chincha Alta .... 13 27 s 76 08 w
Chota ..... 06 33 s 78 39 w
Chulucanas ..... 05 06 s 80 10 w
Concepción ..... 11 55 s 75 17 w
Contamana ..... 07 15 s 74 54 w
Coracora ..... 15 02 s 73 47 w
Cotahuasi ..... 15 12 s 72 56 w
Cutervo ..... 06 22 s 78 51 w
Cuzco ..... 13 31 s 71 59 w
Espinar ..... 14 47 s 71 29 w
Ferreñafe ..... 06 38 s 79 48 w
Huacho ..... 11 07 s 77 37 w
Huacrachuco .... 08 39 s 77 05 w
Huallanca ..... 09 51 s 76 56 w
Huamachuco .... 07 48 s 78 04 w
Huancapi ..... 13 41 s 74 04 w
Huancavelica ... 12 46 s 75 02 w
Huancayo ..... 12 04 s 75 14 w
Huanta ..... 12 56 s 74 15 w
Huánuco ..... 09 55 s 76 14 w
Huaral ..... 11 30 s 77 12 w
Huaraz ..... 09 32 s 77 32 w
Huari ..... 09 20 s 77 14 w
Huarmey ..... 10 04 s 78 10 w
Ica ..... 14 04 s 75 42 w
Ilabaya ..... 17 25 s 70 31 w
Ilave ..... 16 05 s 69 40 w
Ilo ..... 17 38 s 71 20 w
Imperial ..... 13 04 s 76 21 w
Iñapari ..... 10 57 s 69 35 w
Iquitos ..... 03 46 s 73 15 w
Jaén ..... 05 42 s 78 47 w
Jauja ..... 11 48 s 75 30 w
Juanjuí ..... 07 11 s 76 45 w
Juli ..... 16 13 s 69 27 w
Juliaca ..... 15 30 s 70 08 w
Junín ..... 11 10 s 76 00 w
La Oroya ..... 11 32 s 75 54 w
Lagunas ..... 05 14 s 75 38 w
Lamas ..... 06 25 s 76 35 w
Lambayeque .... 06 42 s 79 55 w
Lampa ..... 15 21 s 70 22 w
Lima ..... 12 03 s 77 03 w
Lircay ..... 12 56 s 74 43 w
Macusani ..... 14 05 s 70 26 w
Manú ..... 12 15 s 70 50 w
Matucana ..... 11 51 s 76 24 w
Miraflores ..... 12 07 s 77 02 w
Mollendo ..... 17 02 s 72 01 w
Moquegua ..... 17 12 s 70 56 w
Motupe ..... 06 09 s 79 44 w
Moyobamba ..... 06 03 s 76 58 w
Nauta ..... 04 32 s 73 33 w
Nazca ..... 14 50 s 74 57 w
Olmos ..... 05 59 s 79 46 w
Otuzco ..... 07 54 s 78 35 w
Oxapampa ..... 10 34 s 75 24 w
Paita ..... 05 06 s 81 07 w
Pampas ..... 12 24 s 74 54 w
Panao ..... 09 49 s 76 00 w
Paucará ..... 12 42 s 74 41 w
Paucartambo .... 13 18 s 71 40 w
Pisco ..... 13 42 s 76 13 w
Piura ..... 05 12 s 80 38 w
Pucallpa ..... 08 23 s 74 32 w
Puerto Chicama . 07 42 s 79 27 w
Puerto Maldonado .... 12 36 s 69 11 w
Puno ..... 15 50 s 70 02 w
Puquio ..... 14 42 s 74 08 w
Quillabamba ..... 12 49 s 72 43 w
Requena ..... 04 58 s 73 50 w
Rímac ..... 12 03 s 77 03 w
Rioja ..... 06 05 s 77 09 w
San Ignacio ..... 05 08 s 78 59 w
San Juan ..... 15 21 s 75 10 w
San Pedro de Lloc ..... 07 26 s 79 31 w
San Ramón ..... 11 08 s 75 20 w
San Vicente de Cañete ..... 13 05 s 76 24 w
Saña ..... 06 55 s 79 35 w
Santiago de Cao . 07 58 s 79 15 w
Santiago de Chuco ..... 08 09 s 78 11 w
Santo Tomás ..... 14 29 s 72 06 w
Saposoa ..... 06 56 s 76 48 w
Satipo ..... 11 16 s 74 37 w
Sicuani ..... 14 16 s 71 13 w
Socabaya ..... 16 28 s 71 31 w
Sullana ..... 04 53 s 80 41 w
Tacna ..... 18 01 s 70 15 w
Talara ..... 04 34 s 81 17 w
Tarapoto ..... 06 30 s 76 25 w
Tarata ..... 17 28 s 70 02 w
Tarma ..... 11 25 s 75 42 w
Tayabamba ..... 08 17 s 77 18 w
Tingo María ..... 09 09 s 75 56 w
Trujillo ..... 08 07 s 79 02 w
Tumbes ..... 03 34 s 80 28 w
Urcos ..... 13 42 s 71 38 w
Virú ..... 08 25 s 78 45 w
Yungay ..... 09 09 s 77 44 w
Yurimaguas ..... 05 54 s 76 06 w
Zumilla ..... 03 30 s 80 16 w
Zorritos ..... 03 40 s 80 40 w

Physical features and points of interest

Altipiano (Puna), *upland* ..... 18 00 s 68 00 w
Amazon, *river* ... 04 21 s 70 02 w
Andes Mountains . 08 00 s 77 00 w
Anticona, see Ticlio Pass
Apurímac, *river* .. 11 48 s 74 03 w
Arequipa Volcano, see Misti Volcano
Blanco, Cordillera . 09 10 s 77 35 w
Central, Cordillera ..... 00 09 s 77 00 w
Chavín de Huántar, *historical site* .. 08 06 s 79 05 w
Chicama, *river* ... 07 56 s 79 17 w
Chincha Islands . 13 38 s 76 25 w
Chira, *river* ..... 04 54 s 81 08 w
Colca Canyon ... 15 25 s 72 00 w
Cóndor, Cordillera del ... 04 00 s 78 30 w
Coropuna, Mount . 15 31 s 72 42 w
Corrientes, *river* . 03 43 s 74 45 w
Costa, *region* .... 11 00 s 77 30 w
Curaray, *river* .... 02 22 s 74 05 w
Ene, *river* ..... 11 09 s 74 19 w
Gran Pajatén, *historical site* ... 07 29 s 77 22 w
Huallaga, *river* ... 05 10 s 75 32 w
Huari, *historical site* .. 12 57 s 74 10 w
Huascarán, Mount ..... 09 07 s 77 37 w
Huascarán National Park ... 10 00 s 77 30 w
Huaura, *river* .... 11 06 s 77 39 w
Huayhuash, Cordillera ..... 10 30 s 76 45 w
Ica, *river* ..... 14 54 s 75 34 w
Inambari, *river* ... 12 41 s 69 44 w
La Montaña, *lowlands* ..... 10 00 s 73 00 w
Las Ballestas, *islands* ..... 13 44 s 76 25 w
Machu Picchu, *historical site* ... 13 07 s 72 34 w
Madre de Dios, *river* ..... 12 35 s 69 35 w
Mantaro, *river* .... 11 48 s 74 03 w
Manú National Park ..... 12 00 s 71 30 w
Marañón, *river* ... 04 30 s 73 27 w
Misti (Arequipa) Volcano ..... 16 18 s 71 24 w
Morona, *river* .... 04 45 s 77 04 w
Nanay, *river* ..... 03 42 s 73 16 w
Napo, *river* ..... 03 20 s 72 40 w
Negra, Cordillera . 09 25 s 77 40 w
Occidental, Cordillera ..... 13 00 s 75 00 w
Pacaya, *river* .... 05 22 s 74 18 w
Pachitea, *river* ... 08 46 s 74 32 w
Pacific Ocean ... 10 00 s 80 00 w
Paracas Peninsula ..... 13 48 s 76 24 w
Pastaza, *river* .... 04 55 s 76 24 w
Piedras, Río de las ..... 12 30 s 69 14 w
Pisco, *river* ..... 13 42 s 76 15 w
Puna, see Altiplano
Purús, *river* ..... 10 00 s 71 00 w
Putumayo, *river* .. 02 15 s 73 00 w
Santa, *river* ..... 08 58 s 78 39 w
Santiago, *river* ... 04 27 s 77 38 w
Sechura, Bay of .. 05 42 s 81 00 w
Sechura Desert .. 05 50 s 80 40 w
Tambopata, *river* . 12 36 s 69 11 w
Tapiche, *river* ... 04 59 s 73 51 w
Ticlio Pass (Anticona) ..... 11 35 s 76 21 w
Tigre, *river* ..... 04 26 s 74 05 w
Titicaca, Lake ... 15 48 s 69 24 w
Ucayali, *river* .... 04 30 s 73 27 w
Urubamba, *river* . 12 56 s 74 26 w
Utcubamba, *river* . 05 30 s 78 32 w
Vilcabamba, Cordillera de ... 13 00 s 73 00 w
Yaguas, *river* .... 02 45 s 70 04 w
Yavarí, *river* ..... 04 21 s 70 02 w
Yerupaja, Mount . 10 16 s 76 54 w
Yuruá, *river* ..... 09 30 s 72 44 w


79°
75°
71°
0°
4°
8°
12°
16°
Equator
Quito
COLOMBIA
ECUADOR
Portoviejo
Ambato
Guayaquil
Cuenca
Gulf of Guayaquil
Machala
Tumbes
Zorritos
Zarumilla
TUMBES
Talara
Ayabaca
Sullana
Chulucanas
San Ignacio
Paita
Piura
Castilla
PIURA
Bay of Sechura
SECHURA DESERT
Olmos
Motupe
Jaén
Bagua
AMAZONAS
Rioja
Moyobamba
Jaramillo
Iquitos
LORETO
Nauta
Requena
Yurimaguas
Lagunas
Tarapoto
Juanjui
Cutervo
Chota
Chachapoyas
Jamas
Pomahuaca
Ferreñafe
LAMBAYEQUE
Lambayeque
Chiclayo
Saposoa
SAN MARTIN
Colasay
Bambamarca
Celendín
CAJAMARCA
Cajamarca
Bolívar
Contamana
San Pedro de Lloc
Pacasmayo
Puerto Chicama
Chilete
Cascas
Contumazá
Huamachuco
Santiago de Cao
Chan Chan
Trujillo
LA LIBERTAD
Santiago de Chuco
Tayabamba
Pataz
Pucallpa
Virú
Huacrachuco
Chimbote
Carhuaz
Yungay
Mount Huascarán 6768
Cosmia
Huaraz
ANCASH
HUASCARAN NATIONAL PARK
Chavín de Huantar
Huallanca
Huari
Timon Maria
Pomos
HUANUCO
UCAYALI
Huánuco
Mount Yanahuanca 4634
PASCO
Huamalies
Cerro de Pasco
Oxapampa
Chaglla
Huacho
Barranca
Junin
San Ramón
Tarma
Satipo
Tingo Pass
La Oroya
Chicla
Jauja
Mancicaya
JUNIN
Huaral
Canta
Ancón
Callao
Lima
Miraflores
Chosica
Concepción
Huancayo
Paucara
Río de las Piedras
Ucayali
Iñapari
MADRE DE DIOS
MANU NATIONAL PARK
Mano
Madre de Dios
Puerto Maldonado
CUZCO
LIMA
Huancavelica
Lircay
Huanta
Quillabamba
Machu Picchu
Urubamba
Cusco
Paucartambo
Urcos
San Vicente de Cañete
Chincha Alta
HUANCAVELICA
Huon
Ayacucho
Andahuaylas
Abancay
Chumbivilcas
CHINCHA ISLANDS
LAS BALLESTAS
PARACAS PENINSULA
Pisco
AYACUCHO
Huancasancos
APURIMAC
Aco
Acomayo
Sicuani
Ica
ICA
Santo Tomás
PUNO
Ayaviri
BOLIVIA
Nazca
Puquio
Ocaña
Caraveli
Mount Coropuna 6426
Lampa
Juliaca
COLCA CANYON
AREQUIPA
Aplao
Puno
Lake Titicaca
Arequipa 2823
Mount Pichupichu 5821
Socabaya
Camana
Mollendo
Moquegua
Ilo
Ilabaya
Tarata
TACNA
Tacna
Arica
CHILE
PACIFIC OCEAN
Cities over 250,000
Cities 50,000 to 250,000
Cities under 50,000
National capitals
Political subdivision capitals
LORETO Political subdivision names
International boundaries
Political subdivision boundaries
Intermittent rivers
Dams
Glaciers
Sand areas
National parks
Historical sites
Spot elevations in metres (1 m = 3.28 ft)
Transverse Mercator Projection
Scale 1: 9,410,000
1 inch equals approx. 150 miles
0 50 100 150 200 mi
0 50 100 150 200 250 300 km
©Encyclopaedia Britannica Inc.

Recent migrations. In Peru, as in most Latin-American countries, there has been a mass migration to the cities during the 20th century, especially since the end of World War II. Lima has been the principal destination during this rural exodus, but Trujillo in the north and Arequipa in the south have also received large numbers of migrants. The lack of opportunity in rural regions is usually cited as the reason for movement to the cities, where migrants hope they will find better health care and educational opportunities, as well as jobs. Some do improve their lot, but others end up in city slums or in squatter settlements at the edges of the cities, where conditions may be little improved over those in the rural areas.

A second focus of migration in Peru has been eastward, to the Montaña. At the end of the 19th century the world rubber boom caused many people to move into the Amazon Basin; more recently the Peruvian government has developed a program to improve the economy of the Montaña, a main purpose of which is to divert migrants away from the already crowded coastal urban centres. The completion of roads from Chiclayo on the north coast to Tarapoto in the northern Montaña and from Lima to Pucallpa in the central Montaña region have stimulated this eastward movement. Further development along the eastern side of the Andes is designed to open new settlements in this region. Nevertheless, the Montaña remains the least populated of the three regions.

Urban Peru. The massive 20th-century migration from the countryside has brought rapid growth to Peruvian urban centres. Lima has become the urban giant, more than 10 times larger than the next largest city, but other cities, particularly Trujillo and Chimbote in the north and Arequipa in the south, have also grown rapidly. Since World War II Peru has changed from a country with a predominantly rural population to one that has more than two-thirds of its people living in cities; more than one-fourth of the nation's population lives within the greater Lima metropolitan area.

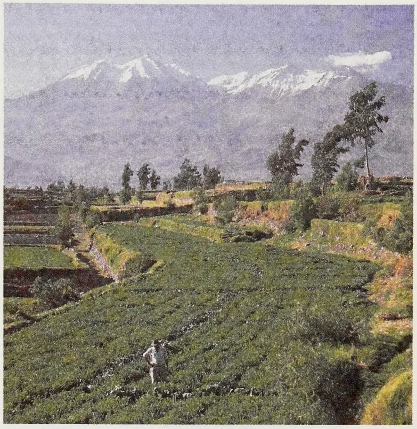
Growth of cities

Ornate colonial architecture contrasts with modern high-rise buildings in Lima, which is the heart of Peru's commerce and industry. Large factories are located in the city, but much of the industrial production takes place in the small workshops of the squatter settlements that surround the city. A difficult problem in Lima has been that of matching the urban infrastructure to the city's growth rate. Lima has only one freeway and lacks an up-to-date mass transit system. Basic public services are, in many neighbourhoods, rudimentary at best. (See LIMA.)

Arequipa in the Sierra and Trujillo in the Costa are other major urban centres. Arequipa is the largest city in southern Peru. Founded in 1540, it is often called the White City because most of its buildings are constructed out of white volcanic rock. Agriculture around Arequipa has improved with the completion of several important irrigation projects, and the area has become a major wool-processing and milk-producing region. Trujillo is a major centre in northern Peru but does not dominate the north as Arequipa does the south. That is because other cities, notably Chiclayo, Chimbote, and Piura, share power in the north, whereas Arequipa is rivaled only by Cuzco, which is in the mountains to the east. Trujillo is the historic power centre in northern Peru, however, and it has become an important commercial centre. Its industries include tractor and diesel motor factories as well as food-processing plants. Chimbote, Peru's best harbour, has a steel mill and numerous fish-processing plants. Chiclayo and Piura mainly serve as regional political and commercial centres.

Most highland cities are small. In the north the principal city, Cajamarca, is noted chiefly as the place where the Spanish conquistador Francisco Pizarro captured the Inca emperor Atahuallpa. Huaraz, 200 miles north of Lima, is a rapidly growing city that was connected to Lima by a paved road in the mid-1970s. To the south, Cerro de Pasco, the nation's largest mining centre, is, at more than 14,200 feet, one of the world's highest cities. Huancayo, about 100 miles due east of Lima, is a farming centre famous for its colourful Sunday market, where Indians sell such handicrafts as llama-wool blankets, ponchos, and sweaters. The best-known Andean centre is the ancient city of Cuzco, once the Inca capital. Tourists from all parts of the world visit Inca remains in Cuzco and its environs, as well as its many colonial churches. The Inca past is apparent in many places. Inca walls topped by Spanish-style structures stand along many streets around Cuzco's main plaza. The most monumental Inca ruins are those of the fortress/sanctuary of Sacsahuamán, built on a hill overlooking the city. The bygone world of Spanish colonial power is evident in the tile-roofed houses and churches of Cuzco. Among the most impressive is the cathedral, dating from around 1550.

Andean cities



Chachani volcano (highest peak) overlooking terraced fields near Arequipa in the southern Sierra region, Peru.
Chip and Rosa María de la Cueva Peterson

The major cities of eastern Peru are Iquitos and Pucallpa. Iquitos, on the upper Amazon, was a small jungle outpost until the rubber boom of the 1880s. When the boom ended, lumber became the major product of the area. More recently oil and tourism have contributed to its growth. Pucallpa, on the Ucayali River, is connected to Lima by road and to Iquitos by river vessels. The area around Pucallpa is a major colonization zone.

Growth of Iquitos

THE PEOPLE

Pre-Hispanic groups. Throughout the pre-Hispanic period, the peoples of Peru were largely isolated from one another by the rugged topography of the country. At least three times, however, a unifying culture spread across the Andes. Beginning *c.* 1000 BC, the Chavín culture permeated the region, emanating possibly from the northern ceremonial site of Chavín de Huántar. After AD 600, the Huari civilization, based at a site of the same name near modern Ayacucho, dominated most of the central Andean region. Finally, the Inca Empire developed, eventually to control all of the territory from northern Ecuador to central Chile. The Inca spread their language, Quechua, across the highlands and along the coast, although some groups near Lake Titicaca spoke Aymara at the time of the Spanish conquest. Quechua and Aymara are still prevalent and have official usage, with Spanish, in regions where they are heavily spoken. Tropical forest areas were outside Incan influence, and the numerous languages and dialects now spoken in the Amazon region reflect the diverse linguistic heritage of the tropical forest peoples.

Ethnic diversity. The Spanish conquerors dominated Peruvian society, including politics, religion, and economics. They brought their European culture and trans-

Spanish cultural influence

mitted their racial characteristics, Spanish language, and Roman Catholic religion to their descendants. The Spaniards introduced a few African slaves, but their number did not become significant. Following independence and the prohibition of slavery, Chinese immigrants were imported as farm labourers, and new groups of Spaniards, northern Europeans, and Japanese were among other arrivals. These diverse racial and ethnic groups have tended to intermarry and produce a mix of racial types, which in modern Peru constitute a complex racial mosaic.

Differences in life-styles and attitudes are pronounced. Peruvians of Spanish descent and the mestizos (racially mixed people) live mainly along the coast and control most of the nation's wealth. Typically, a small group of white people hold the main power in government and industry. Mestizo culture is a blend of Indian and European ways known as *criollo*. The Spanish-speaking mestizos make up the middle class of Peruvian society. They hold managerial, administrative, and professional jobs, but some are also small landowners and labourers. The Indians of the Sierra, who make up about one-third of the population, live in extreme poverty in a harsh environment; they remain both indifferent to, and outside of, the mainstream affairs of the country. Land reforms have brought some improvement, but many highland Indians shepherd llama herds and work tiny plots of land to eke out a living. Like their Inca ancestors, the overwhelming number of Indians read neither their own nor another language. The lowland Indians of the Montaña occupy a social position similar to that of the highland Indians.

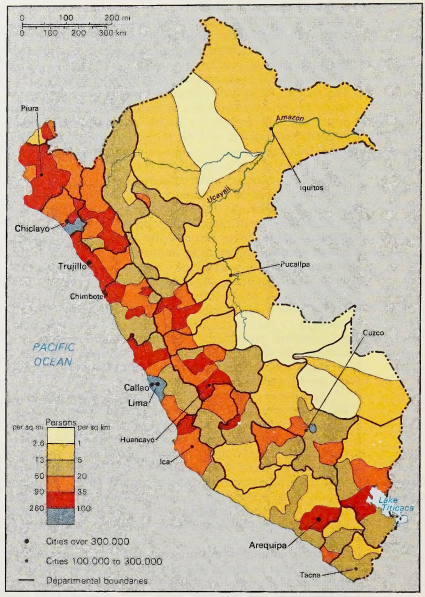
Religion. Ancient Peru had various polytheistic and pantheistic religions. The most important gods were Viracocha (lord, creator, and father of men) and Pacha Mama (Earth mother). The Sun, Moon, and such phenomena as lightning and mountains were also worshiped. Each culture raised temples to honour its local divinity.

Roman Catholic dominance

The Hispanic conquest of the Incas brought new religious traditions to the Andean area. The Spanish indoctrinated the Indians and spread Roman Catholicism, built hundreds of churches, and held fiestas for patron saints in each village. The people were not strict in their practices, however. Protestant sects have proliferated during the 20th century, and the Indians have mixed many pagan beliefs into the Roman Catholic rituals to produce a syncretic religion rich in traditions. Although most Peruvians are Roman Catholic, the constitution provides for freedom of religion.

Demography. The population of the Inca Empire at the time of the Spanish conquest in 1532 is commonly estimated to have been around 12,000,000, although estimates vary. Not all of these people, of course, lived within the boundaries of modern Peru, but it is clear that Peru was the most densely settled area in pre-Hispanic South America. During the first century of Spanish domination, the Indian population declined by almost 80 percent—due to overwork, malnutrition, and the introduction of such diseases as smallpox and measles. The country's first accurate census (1791) showed the impact of Hispanic dominance of the Inca: the population had declined to slightly more than 1,000,000 (which included Europeans, people of mixed ancestry, and black slaves). After independence the population gradually increased, mainly as a result of high birth rates. By the mid-1960s the population of Peru was about the same as that of the Inca society at its height—in other words, it had taken more than 300 years to replace the population lost in the first century of Spanish domination.

Since World War II, the population of Peru has grown rapidly and has become predominantly urban. There has been a sharp decline in death rates, caused mainly by improvements in sanitation practices and health care, while at the same time birth rates have remained very high. As a result, there has been a surplus of population in many areas, particularly in the Andean region. Overpopulation of the rural areas has been the root cause of the mass migration to the cities that has occurred in Peru since the 1950s. The high rate of population growth has also placed great pressure on Peru's educational and health care systems.



Population density of Peru.

THE ECONOMY

Peru is a developing country whose economy has long been dependent upon the export of raw materials to the more developed nations of the Northern Hemisphere. It is one of the world's leading fishing countries and ranks among the largest producers of bismuth, silver, and copper. In recent decades, the country has struggled to modernize its economy by developing nontraditional export industries as well as the manufacture of consumer items to meet local needs. Serious economic problems persist, however, in several areas. A shortage of investment capital is paralleled by a loss of university-trained technicians to offers of higher salaries abroad. Extensive destruction of transportation and agricultural systems occurs periodically from earthquakes, landslides, and other natural disasters. The limited agricultural areas do not meet the needs of the rapidly expanding population, resulting in continually rising imports of foodstuffs and difficult attempts to alter the nation's farming and eating habits. To remedy these and other economic deficiencies the government nationalized the petroleum and other industries in the late 1960s and early 1970s and made extensive efforts at agrarian reform. Nationalization, however, created additional economic problems, including massive government debt and a large trade deficit. This caused successive Peruvian governments to reassess the role of the state in the economy and to reopen some economic sectors to private entrepreneurs.

Major economic problems

Resources. *Minerals and mining.* Peru has a wealth of mineral resources. Copper, iron, lead, zinc, bismuth, phosphates, and manganese exist in great quantities of high-yield ores. Gold and silver are found extensively, as are other rare metals, and petroleum fields are located along the far north coast. Oil exploration in the eastern tropical forests indicates substantial reserves.

In spite of this potential mineral wealth, exploitation

has lagged for a number of reasons, including diminished foreign investment, world price fluctuations, lack of transportation facilities, a scarcity of processing plants, and the depletion of deposits. Difficulties of topography have also hindered development, because some of the most promising resources are located at elevations above 12,000 feet or in the Amazonian forests.

Hydroelectric development

Power. The hydroelectric potential of Peru is great, especially on the rivers that flow eastward out of the Andes Mountains to the Amazon Basin. Large power plants have been built on the Santa and Mantaro rivers, and other locations have been selected for future development. Most existing plants, both thermal and hydroelectric, have been connected to a coordinated national electric grid. More than three-fourths of the nation's electrical energy is produced from hydroelectric sources. Much of the nation's production and demand is in the Lima metropolitan area, where there is a heavy concentration of industry.

Agriculture. The primary economic activity in Peru has long been farming, although since World War II the importance of this sector of the national economy has declined sharply. About one-third of the work force, however, is still employed in agriculture. Despite the number of farm labourers, harvests have not kept pace with the country's rapid population growth, and since the 1960s the per capita production of food has been declining. Peru must import large amounts of grain (particularly wheat, rice, and corn), vegetable oils, dairy products, and meat to feed its population. Although ambitious development plans have been designed to improve output, the scarcity of arable land and water is a limiting factor in Peru.

The most productive agricultural areas are the irrigated valleys of the coastal region, where yields are uniformly high. Principal crops include sugarcane, cotton, rice, corn, grapes, olives, and vegetables. In the Sierra, cropland is limited and soil fertility low. The main crops in the Sierra region are potatoes and grains, especially wheat, corn, and quinoa, an extremely high-protein cereal. Numerous illegal coca plantations (for the cocaine trade) have been set up in the Amazon region of Peru, although many Indian groups have traditionally harvested coca leaves for local use.

Fishing and forestry. In the 1950s and 1960s, Peru's fishing industry expanded rapidly, based on the harvest of enormous schools of anchovy. These fish were converted into fish meal and oil for export as animal feed. By 1963 Peru was the world's leading fishing nation, measured in terms of tonnage caught. The anchovy catch peaked at more than 12,000,000 tons in 1970. Overfishing, combined with a severe occurrence of the El Niño current in 1971–72, sent the fishing industry into decline. Recovery took place during the late 20th century, although the catch did not approach earlier record levels. Increasing emphasis is put on fish for human consumption in the domestic and export markets. Forestry has been mainly concentrated in the Montaña. Many commercial woods are available, but they are often inaccessible and exploitation has been hampered by fears of ecological damage.

Decline of the fishing industry

Industry. Although the Peruvian government has tried to disperse industrial production, most Peruvian factories are located within the greater Lima area. Many products are imported as partially assembled kits, such as automobiles and appliances, with completion done in local factories using Peruvian labour and some Peruvian components. Much of the production of consumer goods is done in small workshops, outside government regulation and taxation—a disturbing factor for economic planners.

The government began promoting industrial growth in the late 20th century, and manufacturing became the fastest growing segment of the economy. To better utilize the nation's natural resources to achieve self-sustained growth, a strong push was given to industries such as those producing petroleum, textiles, processed food, steel, cement, fertilizer, and chemicals. Many of these industries either were nationalized or benefited from special tax incentives and trade protectionist policies during the 1970s; most were privatized again during the 1990s. In general, the government's efforts to promote industrial development have not met expected goals.

Finance. The main institutions dealing with finance in Peru are the large commercial banks. The Central Reserve Bank of Peru and the National Bank control such areas as currency regulation, bank regulation, and foreign exchange. Government financial policies are focused on dealing with inflation and foreign debt, which were serious problems in the 1970s and '80s.

Trade. Foreign trade has been a mainstay of the Peruvian economy since colonial times. The country has been historically dependent on imported manufactured products, a situation that has prompted some administrations to subsidize import substitution industries. The need for imported materials remains, however. Since the 1950s, Peru's imports have consisted primarily of foodstuffs, consumer goods, and component parts for Peruvian industries. Petroleum products formed an expensive share of Peru's imports in the early 1970s, but increased domestic production, particularly from the Amazon area, turned Peru into a net exporter of oil by 1980. Other important exports have been such primary commodities as ores and minerals (copper, silver, lead, and zinc, for example) and such agricultural products as cotton and sugar. Fish meal, too, has been a leading export since the 1960s.

Major trading partners

The United States is Peru's major trading partner, usually accounting for as much as one-fourth of imports and exports. Western European countries and Japan are the source for most other manufactured goods, and those countries also buy a large percentage of what Peru exports. Other recipients of Peru's exports include Japan, China, Germany, The Netherlands, Italy, and neighbouring South American countries such as Colombia and Brazil. In 1969 Peru became a charter member of the Andean Group, but economic problems have hampered implementation of some trade policies.

Transportation. Peru's transportation system faces the challenge of the Andes and of the complex Amazon River system. The only truly integrated networks are the roads and the airlines; the country's two railroad systems have not been interconnected, and maritime traffic is primarily dependent on calls from world shipping lines. River traffic in the Amazonian region is unimportant because of the vast distances and low population density of that area. Roadways cross the country from north to south, or they form penetration roads that run east–west over the Andes. The most important road is the Pan American Highway, which parallels the coast from Ecuador to Chile. Another main road connects with Chile through the Andes.

The major Peruvian railroad, the Central Railway, rises from the coast at Callao near Lima to cross the continental divide at about 15,700 feet. It connects with a branch line to Cerro de Pasco, making it of great importance to the mining industry of the central Andes. A longer line, the Southern Railway, serves Cuzco, Juliaca, Arequipa, and other cities; some of its traffic originates in Bolivia. Callao, on the Pacific Ocean, is the port of Lima and the largest of Peru's numerous ports. Iquitos, located on the Amazon some 2,300 miles from the river's mouth, is the major river port of eastern Peru.

The rough terrain of Peru compels the use of the airplane, but it also complicates flight. Air transport is especially important in hard-to-reach places of the heavily forested east. Commercial aviation began in 1928, and domestic lines operate in addition to numerous foreign airlines. The Jorge Chávez International Airport in Callao, which serves Lima, is the most important of Peru's airports.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Peru has a long history of unstable political life that has been punctuated by numerous military coups and changes of constitution. The 1993 Peruvian constitution decrees a government headed by a popularly elected president who serves as the chief of state and commander of the armed forces. The prime minister, appointed by the president, presides over the Council of Ministers (Cabinet), also appointed by the president. The council approves all presidential law decrees and draft bills sent to the legislature. The bicameral legislature consists of a Senate

The legislature

and a Chamber of Deputies. Members of the legislature are elected to five-year terms, running concurrently with the term of the president. All judges are appointed by the president from lists of nominees submitted by the National Justice Council. Appointments are confirmed by the Senate, and, once confirmed, judges may serve until age 70. The Supreme Court has nationwide jurisdiction and hears appeals from lower-court decisions; it also investigates the conduct of lower-court judges.

The country has been politically divided into 24 departments and one constitutional province, Callao; the departments are further divided into provinces and districts. The 1993 constitution calls for each department to be governed by an assembly. Considerable powers were given to local governments by the constitution. In 1987 legislation was passed that provided for the reorganization of departments into 12 regional governments, but the plan has not been put into effect. A wide spectrum of political parties, ranging from right-wing conservative to left-wing socialist and communist, participate in the political process, as do hastily formed coalitions.

Education. Peru's educational problems are complicated by the steadily increasing percentage of young people in its population. Thus the state must spend a disproportionate share of its resources on education, which is free and compulsory for all children between six and 15. Compulsory education is difficult to enforce, however, especially outside urban centres. Because of extremely large class sizes and limited resources, the quality of education received by children in public schools is regarded as low. As a result, most middle- and upper-class parents send their children to private schools. Universities in Peru include such large, high-calibre institutions as the Pontifical Catholic University of Peru, the University of Lima, and the National University of San Marcos, which was founded in 1551 and claims to be the oldest university in South America. There are also a number of provincial and private universities.

Health and welfare. Numerous public agencies in Peru are involved with national health and social security. Nevertheless, there is a shortage of doctors, nurses, and health care facilities, particularly outside the Lima metropolitan area, and the nation faces a difficult path to adequate health service for its population. Infant and young-child mortality has long been a major problem, and it is one of the major concerns of the government. Peru has a serious shortage of housing units, especially in the urban squatter settlements but also in the countryside. Sanitation is another major problem, with most cities lacking adequate sewerage as well as street lighting and paving.

Shortage of health care

CULTURAL LIFE

The complex ethnic and cultural mixture of Peru presents an entwining of aboriginal pantheism, Spanish mysticism, and African religious practices manifested in the nation's music, literature, textiles, handicrafts, gold and silver work, and bounteous cuisine.

Heritage. Peru's cultural past is best known in such Inca remains as Machu Picchu, a "lost city" discovered in 1911 by U.S. explorer Hiram Bingham northwest of Cuzco. Surrounded by lush green, forested hills and snow-capped peaks, Machu Picchu comprises hundreds of well-built agricultural terraces, a multitude of small stone houses, and several ceremonial temples constructed of carved rock. Research suggests that Machu Picchu was a royal estate of the Inca emperor Pachacuti. Hundreds of archaeological ruins dot the Peruvian countryside, but they are especially impressive around Cuzco, in the region known as the Sacred Valley. Archaeologists have uncovered thousands of decorated jugs and bowls and embroidered textiles; the weavings of the Paracas culture and the ceramics of the Mochica are especially distinguished, as are the extensive Nazca Lines. The dryness of the coast has preserved many pre-Incan remains.

Machu Picchu

The arts. *Folk culture.* Peruvian folk culture is deeply tinged with inheritances from the ancestral peoples. In both town and countryside notable examples of pre-Hispanic and mestizo lore abound in myths, song, superstitions, and dances. Handicrafts, popular with tourists and collectors, provide a close link with such pre-Hispanic crafts as weaving, ceramics, and metalworking.

Westernized arts. The arts have long occupied positions of esteem among Peru's educated minority. Since the late 19th century, most writers have felt a ceaseless duty to analyze their society. Ricardo Palma was among the first to utilize Peruvian themes. *Aves sin nido* (1889; *Birds Without a Nest*), by Clorinda Matto de Turner, was the first of many books whose authors exposed the conditions of Indian life. César Vallejo is often regarded as Peru's finest poet, and José María Arguedas and Mario Vargas Llosa have received high critical acclaim. José María Valle-Riestra's opera *Ollanta* and Vicente Stea's *Sinfonía autóctona* (*Aboriginal Symphony*) were the major musical works of 19th-century Peru. Later, Luis Duncker Lavalle incorporated Peruvian motifs into Western forms.

Literature

Painting reached its zenith with the famous Cuzco school during the 17th and 18th centuries. Most of the thousands of paintings and sculptures are anonymous, and the works show resemblances both to Byzantine and to Asian forms. Modern Peruvian art has followed an abstract course, notably in the work of the painter Fernando de Szyszlo and the sculptor Joaquín Roca Rey.

The oldest colonial work is the cathedral of Lima, and the most important architectural jewel is the convent and church of San Francisco in Lima. Contemporary architecture has been characterized by the so-called neo-Peruvian, or Peruvian Baroque, style and by modern concrete and steel structures.

The theatre is a popular institution in Peru, with a strong tradition dating to colonial times. National professional companies perform in major productions at the Municipal Theatre, which was built in Lima at the site of a colonial theatre dating to 1604. The concerts of the National Symphony Orchestra are also presented there, as are the performances of the main national and touring ballet and folk dance companies. Filmmaking in Peru is not well developed; most films produced there are short, the full-length features being mostly imports. Several Peruvian *telenovelas* (soap operas) have been distributed throughout Latin America.

Cultural institutions. Much of the country's cultural development is overseen by the National Institute of Culture, which seeks to make cultural activities available to all. The Peruvian museums are especially rich in their archaeological collections representing Peru's pre-Hispanic past. The most noteworthy of these are in Lima and include such institutions as the National Museum of Art, the Museum of Anthropology and Archaeology, and the Gold Museum. The main library collection is housed in the National Library in Lima and in the major university libraries.

Museums and libraries

Recreation. Recreational activities vary as widely in Peru as do the social classes, but for everyone there are the fiestas, which are held by numerous communities across the country. These colourful events often celebrate religious themes, but some are held for secular holidays and other events. The most popular spectator sport, as in most other Latin-American nations, is association football (soccer). Bullfights are not common in most cities but draw huge crowds to the renowned Plaza de Acho bullring in Lima. Basketball, horse racing, and cockfighting are also popular. Sporting and recreational facilities are most prevalent in the large metropolitan areas of the Costa.

Press and broadcasting. Although freedom of the press is guaranteed by the Peruvian constitution, the media have been periodically subjected to government control. The major dailies generally have a tradition of taking strong political stands in support of political parties of their choice. Most of the leading dailies, such as *El Comercio, Expreso,* and *Ojo,* are in Lima; others are published in Arequipa and Chiclayo. Lima's *El Peruano,* one of the oldest dailies in the Americas, was founded in 1825. Electronic media have sometimes been subjected to control, which became especially severe in the early 1970s when the national government assumed ownership of 51 percent of all television stations and 25 percent of all radio stations. Freedom of the press was curtailed in the 1990s under President Alberto Fujimori. Lima has several television channels, and there are stations in all of the major cities of the Costa.

For statistical data on the land and people of Peru, see

the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR. (J.P.-V./J.S.K.)

History

Humans have probably lived in Peru for more than 13,000 years. Beginning about 1250 BC, several advanced cultures, such as the Chavín, Chimú, Nazca, and Tiahuanaco, developed in different parts of Peru; however, the area was not unified politically until AD 1438, when the Inca set out from their base in Cuzco on a career of conquest that, during the next 50 years, brought under their control the area of present-day Peru, Bolivia, northern Argentina, Chile, and Ecuador. Within this area, the Inca established a totalitarian state that enabled the tribal ruler and a small minority of nobles to dominate an inert population. (R.N.Bu./T.M.D.)

THE INCA

Like the Aztec, the Inca came late upon the historical scene; even their legends do not predate AD 1200, with the first emperor. For, like Old World peoples, and unlike other aboriginal Americans, the Inca recounted their history by kingly reigns. Most of the accounts agree on 13 emperors (see PRE-COLUMBIAN CIVILIZATIONS). The first seven emperors were legendary, local, and of slight importance; their traditions are full of impossible or improbable events, especially those of the quasi-mythological founder, Manco Capac. In this period the Inca were a small tribe, one of many, whose domain did not extend many miles around their capital, Cuzco. They were almost constantly at war with neighbouring tribes.

Pachacuti

The incredibly rapid expansion of the Inca Empire began with Viracocha's son Pachacuti, one of the great conquerors, and one of the great men, in the history of the Americas. Also with his accession in 1438 reliable history began, almost all the chroniclers being in practical agreement. Pachacuti was called by the British geographer-historian Sir Clements Markham "the greatest man that the aboriginal race of America has produced." He and his son Topa Inca may be aptly compared to Philip and Alexander of Macedon. Pachacuti was evidently a great civic planner as well; tradition ascribes to him the city plan of Cuzco as well as the erection of many of the massive masonry buildings that still awe visitors to this ancient capital.

The sudden great expansion of the Inca Empire was one of the most extraordinary events of history. It covered a little less than a century, from the accession of Pachacuti in 1438 to the conquest by Francisco Pizarro in 1532, and most of it was apparently accomplished by Pachacuti and Topa Inca in the 30 years between 1463 and 1493. At its maximum the empire extended from the present Colombia–Ecuador border to central Chile, a coastal distance of more than 2,500 miles, encompassing approximately 380,000 square miles, about equal in area to France, Belgium, The Netherlands, Luxembourg, Switzerland, and Italy combined.

First the Aymara-speaking rivals in the region of Lake Titicaca, the Colla and Lupaca, were defeated, and then the Chanca to the west; the latter attacked and nearly captured Cuzco. After that there was little effective resistance. First the peoples to the north were subjugated as far as Quito, Ecuador, including the powerful and cultured "kingdom" of Chimú on the northern coast of Peru. Topa Inca then took over his father's role and turned southward, conquering all of northern Chile as far as the Maule River, the southernmost limit of the empire. His son, Huayna Capac, continued conquests in Ecuador to the Ancasmayo River, the present border between Ecuador and Colombia. (J.A.Mn./T.M.D.)

DISCOVERY AND EXPLORATION BY EUROPEANS

Spanish interest in the west coast of South America grew after Vasco Núñez de Balboa discovered the Pacific in 1513, but it was not until 1524 that Francisco Pizarro, aided by another soldier, Diego de Almagro, and a priest, Hernando de Luque, undertook explorations that led to the conquest of Peru. By 1527 they were convinced of the wealth of the Inca Empire. Failing to win further cooperation from Panama's governor, Pizarro returned to Spain, where he received authorization from Charles I to conquer and govern the area extending 600 miles south from Panama. When Pizarro, accompanied by his brothers, returned to Panama, Almagro was outraged by the vast powers Pizarro had acquired for himself. Nevertheless, he continued to collaborate. Pizarro embarked for Peru in late 1530 or early 1531 with 180 men. Establishing a base at San Miguel on the north coast of Peru, Pizarro crossed the mountains to make contact with the Inca Atahuallpa, who had recently been victorious in civil war against his half-brother Huascar, and who was then residing near Cajamarca with an army of about 30,000. Atahuallpa, scornful of the tiny band of invaders, accepted Pizarro's invitation to meet in Cajamarca. There, after Atahuallpa had refused to accept Spanish suzerainty, he was taken prisoner. After his agents had collected a large ransom for his promised release, Atahuallpa was executed for his presumed responsibility for the murder of Huascar. As a means of controlling the Indians, Pizarro then recognized Manco Capac, Huascar's brother, as emperor. In November 1533 the Spaniards occupied Cuzco, the Inca capital.

Pizarro's base at San Miguel

COLONIAL PERIOD

The consolidation of Spanish control proceeded. The city of Quito was subdued and Diego de Almagro left to conquer his domain of Chile. Pizarro organized a Spanish-type municipal government for Cuzco and in 1535 established a new city, Lima, on the coast, to facilitate communications with Panama. Lands were allotted to the conquerors, who were provided with a labour force by grants of encomiendas, which enabled them to collect tribute from the Indians.

Serious trouble then erupted. An unsuccessful Indian rebellion led by Manco Capac in 1536 was followed by open conflict among the conquerors over the division of the spoils. Almagro, disillusioned by Chile's poverty, sought to seize Cuzco from the Pizarros. Almagro was defeated and executed in 1538, but his adherents continued to conspire with his son, and they succeeded in assassinating Francisco Pizarro in 1541. However, an agent of the Spanish crown, sent to establish order, refused to recognize the younger Almagro, who was captured and executed in 1542.

Difficulties persisted nonetheless. The King of Spain, impelled by humanitarianism and by fear that the encomienda system might promote feudalism, promulgated in 1542 the New Laws. These laws threatened the existence of the encomiendas, which were so important to the conquerors. When Viceroy Blasco Núñez Vela arrived in Peru in 1544 to enforce the New Laws, the conquerors, led by Gonzalo Pizarro, revolted and executed the Viceroy. Pizarro maintained control for two years until Pedro de la Gasca, a Spanish agent, undermined his power.

The New Laws

It was nearly a decade before unruly conquerors were controlled under Viceroy Andrés Hurtado de Mendoza (1555–61), and not until the viceregal administration of Francisco de Toledo (1569–81) was systematic control of the huge Indian population attempted. Toledo adapted Indian institutions to the purposes of Spanish authority. Indian chieftains administered local Indian affairs according to custom but were responsible for collecting tribute and providing forced labour. Spanish agents (corregidors) were appointed to protect the interests of both crown and Indian in the Indian communities. Fearing that Manco Capac's son, Tupac Amaru, might prove dangerous to Spanish authority, Toledo ordered him executed in 1571.

By the end of Toledo's administration, the viceroyalty of Peru had assumed the form that it retained into the 18th century. Its territory included all of South America except Venezuela and Portuguese Brazil. Although ranching, agriculture, and commerce were carried on, the mining of precious metals, particularly silver, was the basic industry, making the colony the most important in the Spanish empire. The discovery of the fabulous Potosí mines in 1545 had been followed in 1563 by the opening of the Huancavelica mines, which produced the mercury essential to efficient processing of silver. Because the viceroyalty's mineral resources, except for the gold of New Granada

(Colombia), were in Peru proper and Upper Peru (Bolivia), these areas became the most highly developed and richest.

The centre of wealth and power for the entire region was the viceregal capital of Lima. There, during the 16th and 17th centuries, a series of viceroys ruled over most of Spanish South America. The elaborate viceregal court was the apex of a highly stratified society based upon forced Indian labour. It attracted not only the politically oriented but also the wealthy, the artistic, and the intellectual.

Lima was also important as seat of the audiencia, which administered royal justice, and as a religious, cultural, and commercial centre. The archbishop of Lima was head of the church in Peru. Many religious orders established monasteries and convents there, and the tribunal of the Inquisition worked to extirpate religious heresy. In Lima also was the capstone of the educational system—the University of San Marcos. Adding to the wealth and importance of Lima was the privileged position that its merchants enjoyed under the monopolistic Spanish trade system. Lima, with its port of Callao, was the entrepôt for trade between Europe and the commercial centres of South America, ranging from Quito to Chile on the Pacific coast and to Buenos Aires on the Atlantic. Under the Spanish system the bulk of legitimate trade to and from these areas had to pass through merchants in Lima.

During the later 17th century Peru experienced difficulties. Some of these, such as increasing contraband trade with non-Spanish merchants, attacks by pirates, and the growth of venality among government officials, reflected the internal decay of Spain and the decline of its international power. Contributing to Peruvian difficulties was the decline of its production of precious metals.

A series of governmental reforms complicated Peru's problems in the 18th century. The Bourbon dynasty, which in 1700 had replaced the Habsburgs as rulers of Spain, undertook a program of reform during the 18th century, seeking to promote the economic development of their colonies, improve colonial defenses, and provide more efficient government. The first to seriously affect Peru was the establishment of the new Viceroyalty of New Granada, ending Peru's control over northern South America and resulting in its loss to New Granada of the thriving port of Guayaquil (now in Ecuador). For the next few decades Bourbon reforms, together with overall expansion of the economy, improved conditions in Peru. In 1777–78, however, the Spanish government established another viceroyalty, that of Río de la Plata, this time depriving the Peruvian viceroy of authority over Upper Peru and the areas of present-day Argentina, Paraguay, and Uruguay. Chile was reconstituted as a virtually autonomous captaincy general. Following the disastrous loss of the silver mines of Upper Peru, the viceroyalty of Peru was still more weakened by reforms in the trade system, which permitted merchants in ports on the Atlantic and Pacific to trade directly with Spain.

Loss of Upper Peru

Internal strife created further complications. The Indians, who had from the time of the conquest suffered oppressive taxation and enforced labour, revolted in 1780 under Tupac Amaru II, a descendant of the last Inca emperor and a man of wealth and education. The revolt spread throughout Peru and into Upper Peru and Ecuador. Although Tupac Amaru II was captured and executed in 1781, the Indians continued to fight the Spaniards until 1783, causing considerable disruption.

Nevertheless, in the late 18th and 19th centuries, Peru experienced a period of intellectual development that was the result of the influence of the utilitarian ideas of the European Enlightenment, brought to Peru in books and by European participants in scientific expeditions in 1778 and 1793. Its chief manifestation was the establishment of a literary and scientific club in Lima, the Society of the Lovers of the Country.

ACHIEVEMENT OF INDEPENDENCE

The Napoleonic invasion of Spain in 1808 sparked the Creoles (those of European descent born in America) in other Spanish colonies to struggle for independence between 1810 and 1821. But Peru remained loyal because of the conservative attitude of the Peruvian aristocracy, the presence of many Spaniards in Peru, the concentration of Spanish military power in Lima, and the effective suppression of Indian uprisings. Peru's independence was, consequently, achieved primarily by outsiders.

José de San Martín

Among them was General José de San Martín of Argentina, whose aims were to secure Argentine control of Upper Peru's silver from the Spanish forces that had occupied Upper Peru and to assure Argentina's independence by destroying the remaining Spanish power in South America. Because Argentine forces had previously been defeated in Upper Peru, San Martín determined to surround the Spaniards by liberating Chile and using it as a base for a seaborne attack on Peru. Chile was freed in 1818 and a fleet was readied, which enabled San Martín to occupy the Peruvian port of Pisco in September 1820. When the viceroy withdrew his forces into the interior, San Martín entered Lima. Peruvian independence was declared on July 28, 1821.

Lacking power to attack the strong Spanish forces in the interior, San Martín sought aid from Simón Bolívar, who had liberated northern South America, but Bolívar declined, refusing to share leadership. San Martín then withdrew, and Bolívar assumed power in Peru to carry on the struggle for liberation. At the battles of Junín (Aug. 6, 1824) and Ayacucho (Dec. 9, 1824) Spanish power was broken and Peru's independence assured.

(R.N.Bu./T.M.D.)

PERU FROM 1824 TO 1884

The end of Spanish rule did not, however, provide a solution to the many political, social, and economic problems facing the country. The transition from a colonial dependency to a modern state proved difficult.

The caudillos

Struggle for power. At the outset of Peru's national existence, military leaders (caudillos) who had gained prominence in the struggle for independence sought to seize power. The departure of Simón Bolívar in 1826 removed a stabilizing influence. The aims of the caudillos were furthered by the absence of a tradition of self-government, by the prevalence of a feudal society of Creoles and Indians, and by the reluctance of civilians to assume political responsibility. Despite military influences, a liberal constitution was adopted in 1828. This did not prevent General Agustín Gamarra from taking government power by illegal means. He was succeeded in 1835 by another self-seeking caudillo, General Felipe Salaverry.

The ambitions of Gamarra and Salaverry were thwarted by Andrés de Santa Cruz, a military commander of Spanish-Indian descent who proposed a confederation of Peru and Bolivia. For three years Santa Cruz, though born in La Paz, was backed by influential groups in Peru and maintained the political union. But his hopes were shattered at the Battle of Yungay in 1839 by a joint force of nationalist-minded Peruvians and of Chileans fearing a threat to the balance of power in the Pacific.

Establishment of order. During the initial period of statehood in Peru, liberal and conservative parties with ill-defined programs emerged. Their rivalry enhanced the political instability of the country.

General Ramón Castilla assumed the presidency in 1845. Castilla dominated the political scene from 1845 to 1851 and from 1855 to 1862, in spite of his mestizo background. His greatest accomplishment for the promotion of national wealth was the exploitation of the guano deposits along the coast and offshore islands. Taxes on this industry, which was controlled by foreign corporations, furnished the principal source of government revenue for several decades. Castilla appealed to liberals by abolishing the payment of tribute by Indians and by emancipating the black slaves. Landowners on the coastal plantations, however, were permitted to import thousands of Chinese coolies in order to have a sufficient labour supply. As an additional concession to the liberals, Castilla established a system of state education at the primary and secondary levels. Through his influence, an assembly in 1860 adopted a constitution that lasted into the 1900s.

In the second half of the 19th century, Peru's history was characterized by many setbacks. In 1864 Spain dispatched a naval force to the Pacific, ostensibly to protect the

rights of Basque immigrants but in reality to reestablish domination over its former colony. In 1869, after meeting with the determined opposition of Peru and Chile, Spain withdrew and recognized Peru's independence for the first time, but the conflict was a heavy drain on Peru's treasury.

Dissatisfaction with military rule resulted in 1871 in the formation of the Civilian Party, representing an oligarchy of landowners and merchants. This party, headed by Manuel Pardo (president, 1872–76), approved a costly program of internal development, which included the construction of railroads across the Andes. Corruption on the part of government officials and contractors characterized the work, which decreased the isolation of the Peruvian interior but increased enormously the national debt.

The War of the Pacific (1879–83). Another untoward event was the War of the Pacific with Chile, caused mainly by rivalry over the exploitation of rich nitrate beds in the Atacama Desert (then in Peru, now in Chile). Chile's superior resources and military discipline brought overwhelming defeat to Peru and its ally Bolivia.

At the Battle of Iquique (then in Peru, now in Chile), on May 21, 1879, the Peruvians suffered the loss of one of their best warships, the *Independencia;* the *Huáscar* was then sunk on October 8, and this eventual surrender of control of the sea permitted a Chilean army to land on the Peruvian coast. On Jan. 17, 1881, Chilean forces captured the capital, Lima. Looting and pillaging followed and the National Library was destroyed. According to the terms of the Treaty of Ancón (Oct. 20, 1883), Peru turned over to Chile full possession of the province of Tarapacá and the administration for 10 years of the provinces of Tacna and Arica, after which a plebiscite was to determine their future sovereignty.

PERU FROM 1884 TO 1930

Economic recovery program

Expenditures for the war, and the consequent loss of revenue from the nitrate fields, created the possibility of imminent bankruptcy. To avert this disaster, the Civilian regime accepted in 1889 a plan proposed by the bondholders for handling the debt. The Peruvian Corporation, representing the creditors, with headquarters in London, was to control the railroads for 66 years, to mine up to 3,000,000 tons of guano, and to receive 33 annual payments of £80,000 each. The plan worked satisfactorily but was hated by the Peruvian people.

Social reforms and economic development. The decline in national prestige created an atmosphere conducive to political change. The Democratic Party was formed, and in 1895, under the leadership of Nicolás de Piérola, it won the presidential election. Having a broad, popular base, it championed direct suffrage and the restoration of municipal elections. Public education was fostered, but schools for the children of the poor were lacking.

An orderly political scene, marked by rivalry between the Democratic and Civilian groups, accelerated economic development. There was an increase in the production of minerals, notably copper, and of such agricultural commodities as cotton, sugar, and wool. In the mining of copper, U.S. capital acquired important interests.

Augusto Leguía y Salcedo

Augusto Leguía y Salcedo, chief spokesman for the Civilians, assumed the presidency in 1908. His first term in office (1908–12) was marked by the expansion of sugar and cotton production and the settlement of the boundary dispute with Brazil. During Leguía's second term (1919–30), he embarked upon expensive public works projects, financed by loans from U.S. banks. Rights to the oil fields of La Brea-Pariñas were given to the U.S.-owned International Petroleum Company, which built a refinery to supply the nation with gasoline and oil.

Leguía supported the adoption of a new constitution in 1920. Among its progressive provisions was Article 58, which protected the communal lands of the Indians from sale and seizure. Failure to implement this provision, however, gave rise to a significant development of Indianism. While most intellectuals urged gradual reforms, more radical measures were advocated by the Peruvian Communist Party and others.

Formation of the Aprista movement. The American Popular Revolutionary Alliance (APRA), known as the "Aprista movement," was formed in 1924 in Mexico City by Víctor Raúl Haya de la Torre, an intellectual then in exile. Internationally, it expressed the ideals of the unity of American Indians and the elimination of U.S. imperialism. Internally, it proclaimed the need to end the exploitation of the Peruvian masses through the institution of a planned economy and the nationalization of foreign-owned enterprises. Its anticapitalist and anti-imperialist stand appealed to intellectuals, to the Indians, and to the lower middle class.

By 1930 Leguía had experienced a definite loss in popularity. Final settlement of the long-standing Tacna–Arica dispute with Chile, by which Peru ceded the province of Arica, angered the extreme nationalists, while world depression cost Leguía the support of business groups.

PERU FROM 1930 TO 1968

In 1930 a military junta headed by Colonel Luis Sánchez Cerro overthrew Leguía, and Sánchez Cerro defeated Haya de la Torre, the APRA candidate, in the presidential elections of 1931. APRA claimed that the elections were fraudulent and instigated a campaign to discredit the regime. The threat from the left led to the emergence of a Fascist group, whose chief exponent was the historian José de la Riva Agüero. In July 1932 Apristas organized an uprising in Trujillo, on the northern coast, and Sánchez Cerro was assassinated by an Aprista in 1933. These incidents created an enduring enmity between the military establishment and APRA.

Troubled democracy. Sánchez Cerro's successor (1933–39) was General Oscar Benavides, who restored confidence in the economy. He also settled a dangerous boundary controversy with Colombia over the port of Leticia on the upper Amazon and a finger of land giving access to the river, both of which had been ceded to Colombia in a treaty of 1922. To avoid war Benavides returned the territory to Colombia. Benavides reduced the strength of APRA by declaring the party illegal, by a relentless persecution of its leaders, and by the adoption of social assistance projects. In the presidential election of 1939 the Apristas supported Manuel Prado, a banker and member of an aristocratic family of Lima.

Alignment with the United States

During World War II Peru cooperated with the United States, authorized Allied use of airfields and ports, and arranged to sell the Allies petroleum, cotton, and minerals. In 1942 Peru severed diplomatic relations with the Axis powers, and in 1945 it declared war on them. During the war Peru succeeded, with U.S. support, in getting a favourable settlement of a boundary dispute with Ecuador, which it had invaded.

World War II not only brought economic prosperity but also hope for real democracy. Prado, swayed by public opinion, approved the presidential candidacy in 1945 of José Luis Bustamante y Rivero, a lawyer from Arequipa with liberal leanings, who represented a coalition of middle- and upper-class elements. APRA, again a legal party, obtained a majority of seats in the lower house and half the seats in the Senate. Bustamante generally followed an independent course, and the Apristas withdrew their support. After Apristas staged an abortive insurrection in Callao, near Lima, the President outlawed the party.

The dictatorship of Manuel Odría. In October 1948 General Manuel Odría seized power, protesting the President's lack of firmness in dealing with the radicals, and extreme measures were taken to suppress the Apristas. Haya de la Torre found refuge in the Colombian embassy, where he stayed for five years before leaving Peru.

Odría led an authoritarian regime in which political stability allowed the revival of prosperity. The Korean conflict of the early 1950s benefited foreign trade because of heavy U.S. demand for Peruvian minerals, and a friendly policy toward foreign capital prompted large-scale investments.

Return to elected government. In the election of 1956, Manuel Prado, who was supported by Odría, won a second term, defeating Fernando Belaúnde Terry. A surprising feature of the election was the decline of APRA, some of whose members joined Belaúnde's National Front Party.

Prado countered the financial crisis inherited from Odría by appointing as minister of the treasury Pedro Beltrán,

whose policies contributed to a 4½ percent annual increase in the gross national product. The fishing industry based on the Peru Current expanded. Beltrán's measures did not, however, lessen the pressure from the landless Indians and the underpaid urban proletariat. (J.P.M./T.M.D.)

In 1962 none of the presidential candidates received the one-third vote necessary for election; the decision went to the congress, but the military seized the government. A new election called in 1963 by the junta was won by Belaúnde's party, now called Popular Action.

Reforms under Belaúnde

Belaúnde promised solutions to the country's economic and social problems. An agrarian act of 1964 provided for expropriation of unused or misused agricultural properties; by 1966 more than 500,000 acres had been distributed. Community development projects and irrigation schemes were instituted, and a network of roads was planned. Indians were encouraged to colonize land in the foothills east of the Andes. Education was promoted with the establishment of new universities and literacy campaign.

MILITARY RULE (1968–80)

In October 1968 the military forced the resignation of Belaúnde. The junta, headed by Juan Velasco Alvarado, imprisoned opposing politicians and suspended constitutional liberties, and the government expropriated the holdings of the International Petroleum Company, straining relations with the United States.

Economic nationalism. In 1969 the junta embarked on a massive program of economic nationalization and agrarian reform. Accordingly, in August 1969, the government seized the Paramonga chemical and paper plant, which belonged to one of the largest U.S. interests in Peru. It sought also to control essential industries and public services through outright ownership and by "Peruvianization"—insistence that a majority of the stock of a foreign company be held by nationals. The occurrence on May 31, 1970, of a major earthquake in northern Peru partly jeopardized the financial stability of the regime.

Agrarian and fishing reforms

The junta appealed to the peasants by expropriating landed estates, which were operated by government-directed collectives or by individuals or Indian communes. The junta signed a contract in July 1971 with a Yugoslav company for the construction of a canal in the Piura Valley to irrigate 330,000 additional acres, and two more projects were subsequently initiated. Fishing was encouraged, but the disappearance of the anchovies in 1972 because of El Niño dealt a serious blow to the economy. In 1973 the government moved to nationalize the fish meal industry, valued at $500,000,000. With the organization of Petroperú, a state-owned company, the petroleum industry expanded.

An education reform bill, promulgated in March 1972, recognized the equality of women, established rural schools, granted autonomy to the universities, and promoted the use of the Indian languages Quechua or Aymara in the schools in the Andes and east of the Sierra.

To prevent criticism of its tight dictatorship, the junta censored the press, closed some radio stations and newspapers, and acquired control of privately owned television stations. As part of an innovative trans-Pacific policy, Japanese investments and contacts were encouraged by the government. Friendship with the Soviet Union led to the exchange of ambassadors with Communist-bloc nations.

New junta. Economic factors fostered resentment among a number of groups toward the Velasco regime. The decline in fish meal exports and in copper prices brought an end to the economic boom, while loans obtained abroad for agrarian reform and huge copper and petroleum projects increased the foreign debt. In August 1975 a new junta took power headed by General Francisco Morales Bermúdez, former minister of finance and economy, and Peruvian policies were constantly altered as repeated changes in the Cabinet took place. Morales made a shift toward more moderate right-wing policies. The National Agrarian Confederation was dissolved in 1978; the state fishing enterprise was denationalized; mining projects were opened to private investors; and more foreign investment was encouraged.

RETURN TO CIVILIAN RULE

The Morales government committed itself to reestablishing constitutional rule, and a popularly elected Constituent Assembly was summoned in June 1978 to draft a new constitution. The Apristas formed the largest bloc of the assembly, and Haya de la Torre was elected president. The new constitution was signed on July 12, 1979.

Elections were held in May 1980, and Fernando Belaúnde Terry was returned to the presidency. His party, Popular Action, headed a majority coalition in the legislature. Belaúnde immediately returned to their previous owners newspapers that had been confiscated by the military junta. The legislature attempted to reorganize the economy and encourage private enterprise. However, the economy was hurt by an increase in imports due to Belaúnde's free-market policies, lower world prices for Peru's major export commodities, high international interest rates on the nation's burgeoning foreign debt, and a devastating El Niño in 1982–83. Aggravating the economic problems was the rise of terrorism, led by the neo-Maoist Sendero Luminoso (Shining Path) and the Tupac Amaru Revolutionary Movement. When the inflation rate rose a staggering 3,240 percent between July 1980 and June 1985, the economy almost collapsed.

Presidency of Alan García Pérez

In the 1985 elections APRA, capitalizing on the nation's plight, had its presidential candidate elected for the first time in its history. He was the young and charismatic Alan García Pérez. The new leader shocked the international community when he announced that Peru would pay no more than 10 percent of its export earnings toward a nearly $14,000,000,000 foreign debt. Adopting a populist stance domestically, García attempted to reactivate the economy, end human rights abuses in the war against the guerrillas, gain control over the drug traffickers, and rally the population. The International Monetary Fund added a blow to the country's condition when it declared Peru to be ineligible for future loans and credits until García adopted more orthodox economic and debt repayment measures. Facing a deteriorating economic situation, the president moved to nationalize the banks in 1987, an act that eroded his personal popularity. The end of García's term was marked by runaway inflation, a series of crippling general strikes, and even rejection by his own party.

With inflation, guerrilla war, and the drug trade major concerns, the 1990 presidential elections resulted in a runoff between Mario Vargas Llosa of the Democratic Front Movement, or Fredemo, and Alberto Fujimori of Change 90. Vargas Llosa, a novelist, advocated a drastic anti-inflation program that alarmed many of Peru's poor. Fujimori, who won the election, gained support by decrying Vargas Llosa's plan. However, less than two weeks after taking office, Fujimori instituted austerity measures as harsh as those he had earlier decried, including suddenly raising the price of gasoline by 3,000 percent. The program wiped out inflation but caused immediate hardships, notably among the poor.

In April 1992 the military assisted Fujimori in staging an *autogolpe* ("self-administered coup"), in which Congress was dissolved. Another legislature was soon elected, and a new constitution was promulgated the following year. Fujimori promoted neoliberal economic policies such as privatizing state-owned mines and utility companies. He also took credit for successful antiterrorism campaigns that included capturing Abimael Guzmán Reynoso, the leader of Sendero Luminoso, in 1992, and storming the Japanese ambassador's residence in Lima in 1996, where dozens of hostages had been held by Tupac Amaru members. Fujimori won a second term in 1995, but charges of fraud and unconstitutionality accompanied his election to a third term in 2000. His government crumbled later that year after the head of the secret police, Vladimiro Montesinos, was found to have bribed a congressman.

(J.P.M./T.M.D./Ed.)

For later developments in the history of Peru, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and the *Micropædia*, see the *Propædia*, sections 951, 964, 966, and 974, and the *Index*.

BIBLIOGRAPHY

Physical and human geography: An overview of Peru is provided by RICHARD F. NYROP (ed.), *Peru, A Country Study,* 3rd ed. (1981). An analysis of the geography is *Gran geografía del Perú: naturaleza y hombre,* 8 vol. (1986). Other useful studies include DAVID A. ROBINSON, *Peru in Four Dimensions* (1964, reprinted 1971); CARLOS PEÑAHERRERA DEL AGUILA, *Geografía general del Perú: síntesis,* vol. 1, *Aspectos físicos* (1969); and CLAUDE COLLIN DELAVAUD, *Les Régions côtieres du Pérou septentrional: occupation du sol, aménagement régional* (1968), available also in Spanish, *Las regiones costeñas del Peru septentrional* (1984). Peru's natural regions and vegetation are discussed in JAVIER PULGAR VIDAL, *Geografía del Perú: las ocho regiones naturales del Perú,* 8th ed. (1981).

Notable works on the topic of Indians and peasants include HILDEBRANDO CASTRO POZO, *Nuestra comunidad indígena,* 2nd ed. (1979); and MOISÉS SÄENZ, *Sobre el indio peruano y su incorporación al medio nacional* (1933), both early classics; THOMAS M. DAVIES, JR., *Indian Integration in Peru: A Half Century of Experience, 1900–1948* (1974); PAUL L. DOUGHTY, *Huaylas: An Andean District in Search of Progress* (1968), an anthropological study; WILFREDO KAPSOLI E., *Los movimientos campesinos en el Perú, 1879–1965,* 2nd ed. (1982), the best study of peasant uprisings; and PIERRE L. VAN DEN BERGHE and GEORGE P. PRIMOV, *Inequality in the Peruvian Andes: Class and Ethnicity in Cuzco* (1977), a study of Indian domination by the "urban mestizo elite." Good ethnographies include STEPHEN B. BRUSH, *Mountain, Field, and Family: The Economy and Human Ecology of an Andean Valley* (1977); and SUSAN C. BOURQUE and KAY BARBARA WARREN, *Women of the Andes: Patriarchy and Social Change in Two Peruvian Towns* (1981).

General economic studies include ROSEMARY THORP and GEOFFREY BERTRAM, *Peru, 1890–1977: Growth and Policy in an Open Economy* (1978); and E.V.K. FITZGERALD, *The Political Economy of Peru, 1956–78: Economic Development and the Restructuring of Capital* (1979). The economic geography of Peru is discussed in EMILIO ROMERO, *Geografía económica del Perú,* 5th ed. (1966). Labour studies include DAVID CHAPLIN, *The Peruvian Industrial Labor Force* (1967); and DENIS SULMONT S., *Historia del movimiento obrero en el Perú, de 1890 a 1977* (1977). Studies on agriculture and fishing include BALTAZAR CARAVEDO MOLINARI, *Estado, pesca y burguesía, 1939–1973: teoría y realidad* (1979); and ARTHUR J. COUTU and RICHARD A. KING, *The Agricultural Development of Peru* (1969).

Samplings and discussions of Peruvian literature can be found in LUIS ALBERTO SÁNCHEZ, *La literatura peruana: derrotero para una historia cultural del Perú,* 5th ed., 5 vol. (1981); and LYNN A. DARROCH (ed.), *Between Fire and Love: Contemporary Peruvian Writing* (1980). On the political role of the Roman Catholic Church and popular religiosity, see JEFFREY L. KLAIBER, *Religion and Revolution in Peru, 1824–1976* (1977). See also GUSTAVO GUTIÉRREZ, *A Theology of Liberation: History, Politics, and Salvation* (1973; originally published in Spanish, 1972), by a leading theoretician of the liberation theology movement.

(J.S.K./T.M.D.)

History: JORGE BASADRE, *Historia de la república del Perú, 1822–1933,* 6th rev. ed., 17 vol. (1968–70), is the basic point of departure for any study of Peru. HENRY F. DOBYNS and PAUL L. DOUGHTY, *Peru: A Cultural History* (1976), is a sensitive treatment of Peru's historical tradition, with particular emphasis on race relations and ecological adaptation. FREDRICK B. PIKE, *The Modern History of Peru* (1967, reprinted 1969), is competent and comprehensive, while his *United States and the Andean Republics: Peru, Bolivia, and Ecuador* (1977), is a brilliant synthesis of the incompatability of Hispanic and Anglo cultures. DAVID SCOTT PALMER, *Peru: The Authoritarian Tradition* (1980), treats Peruvian history within the context of Latin-American development.

The best accounts of Peru's pre-conquest peoples are EDWARD P. LANNING, *Peru Before the Incas* (1967); LUIS G. LUMBRERAS, *The Peoples and Cultures of Ancient Peru* (1974; originally published in Spanish, 1969); JOHN V. MURRA, *Formaciones económicas y políticas del mundo andino* (1975); and FRANKLIN PEASE G.Y., *Del Tawantinsuyu a la historia del Perú* (1978), which emphasizes the indigenous origins of modern Peruvian culture. Superb ethnohistorical accounts of the impact of conquest and of the early colonial period are NATHAN WACHTEL, *The Vision of the Vanquished: The Spanish Conquest of Peru Through Indian Eyes, 1530–1570* (1977; originally published in French, 1971); KAREN SPALDING, *Huarochirí: An Andean Society Under Inca and Spanish Rule* (1984); and STEVE J. STERN, *Peru's Indian Peoples and the Challenge of Spanish Conquest: Huamanga to 1640* (1982). See also JAMES LOCKHART, *Spanish Peru, 1532–1560: A Colonial Society* (1968); and NOBLE DAVID COOK, *Demographic Collapse: Indian Peru, 1520–1620* (1981), which treats the devastating loss of Indian life due to European contact.

Treatments of Spanish colonial institutions and economy include J.R. FISHER, *Government and Society in Colonial Peru: The Intendant System, 1784–1814* (1970), and *Silver Mines and Silver Miners in Colonial Peru, 1776–1824* (1977); ROBERT G. KEITH, *Conquest and Agrarian Change: The Emergence of the Hacienda System on the Peruvian Coast* (1976); and NICHOLAS P. CUSHNER, *Lords of the Land: Sugar, Wine, and Jesuit Estates of Coastal Peru, 1600–1767* (1980). LEON G. CAMPBELL, *The Military and Society in Colonial Peru, 1750–1810* (1978), studies the origins of Peruvian civil-military relations. TIMOTHY E. ANNA, *The Fall of the Royal Government in Peru* (1979), views independence as an accidental occurrence; while HERACLIO BONILLA *et al., La independencia en el Perú,* 2nd ed. (1981), argues that Peru had to be conquered. HERACLIO BONILLA, *Guano y burguesía en el Perú,* 2nd ed. (1984), analyzes the rise of the guano industry and its impact on capitalist development. FLORENCIA E. MALLON, *The Defense of Community in Peru's Central Highlands: Peasant Struggle and Capitalist Transition, 1860–1940* (1983), treats nascent Peruvian capitalism from a rural perspective. ROBERT N. BURR, *By Reason or Force: Chile and the Balancing of Power in South America, 1830–1905* (1965, reprinted 1974), remains the best diplomatic history of the War of the Pacific.

Excellent political analyses of the 20th century include FRANÇOIS BOURRICAUD, *Power and Society in Contemporary Peru* (1970; originally published in French, 1967); CARLOS A. ASTIZ, *Pressure Groups and Power Elites in Peruvian Politics* (1969); and JULIO COTLER, *Clases, estado y nación en el Perú,* 2nd ed. (1982). The best studies of the crucial period 1919–40 are STEVE STEIN, *Populism in Peru: The Emergence of the Masses and the Politics of Social Control* (1980); and BALTAZAR CARAVEDO MOLINARI, *Clases, lucha política y gobierno en el Perú, 1919–1933* (1977). APRA, one of the most important political movements in Latin America, is the subject of ROBERT J. ALEXANDER (ed. and trans.), *Aprismo: The Ideas and Doctrines of Víctor Raúl Haya de la Torre* (1973); PETER F. KLARÉN, *Modernization, Dislocation, and Aprismo: Origins of the Peruvian Aprista Party, 1870–1932* (1973); and FREDRICK B. PIKE, *The Politics of the Miraculous in Peru: Haya de la Torre and the Spiritualist Tradition* (1986). Economic analyses of Belaúnde's first term include PEDRO-PABLO KUCZYNSKI, *Peruvian Democracy Under Economic Stress: An Account of the Belaúnde Administration, 1963–1968* (1977); and RICHARD C. WEBB, *Government Policy and the Distribution of Income in Peru, 1963–1973* (1977). Guerrilla movements are discussed by members of the movements in HUGO BLANCO, *Land or Death: The Peasant Struggle in Peru* (1972); and HECTOR BÉJAR, *Peru 1965: Notes on a Guerrilla Experience* (1970; originally published in Spanish, 1969).

Civil–military relations are treated in VÍCTOR VILLANUEVA, *Ejército peruano: del caudillaje anárquico al militarismo reformista* (1973), and his seminal study of the Center for High Military Studies, *El CAEM y la revolución de la fuerza armada* (1972); and in BRIAN LOVEMAN and THOMAS M. DAVIES, JR. (eds.), *The Politics of Antipolitics: The Military in Latin America* (1978). The most important literature on the 1968–80 military government includes ABRAHAM F. LOWENTHAL (ed.), *The Peruvian Experiment: Continuity and Change Under Military Rule* (1975), an overly optimistic account; CYNTHIA MCCLINTOCK and ABRAHAM F. LOWENTHAL (eds.), *The Peruvian Experiment Reconsidered* (1983), a scathing postmortem; CYNTHIA MCCLINTOCK, *Peasant Cooperatives and Political Change in Peru* (1981); ALFRED STEPAN, *The State and Society: Peru in Comparative Perspective* (1978), a study of authoritarian rule; GEORGE D.E. PHILIP, *The Rise and Fall of the Peruvian Military Radicals 1968–1976* (1978), an analysis of the reforms undertaken by Velasco; and STEPHEN M. GORMAN (ed.), *Post-Revolutionary Peru: The Politics of Transformation* (1982), an analysis of the successes, failures, and legacy of the military government.

(T.M.D.)

Peter I the Great, of Russia

Peter I the Great (Russian Pyotr I Veliky, in full Pyotr Alekseyevich), tsar of Russia from 1682 and emperor from 1721, was one of his country's greatest statesmen, organizers, and reformers. He was born in Moscow on June 9 (May 30, old style), 1672, the son of Tsar Alexis by his second wife, Natalya Kirillovna Naryshkina. Unlike his half-brothers, sons of his father's first wife, Mariya Ilinichna Miloslavskaya, Peter proved a healthy child, lively and inquisitive. It is probably significant to his development that his mother's former guardian, Artamon Sergeyevich Matveyev, had raised her in an atmosphere open to progressive influences from the West.

Youth and accession. When Alexis died in 1676 Peter was only four years old. His elder half-brother, a sickly youth, then succeeded to the throne as Fyodor III; but, in fact, power fell into the hands of the Miloslavskys, relatives of Fyodor's mother, who deliberately pushed Peter and the Naryshkin circle aside. When Fyodor died childless in 1682, a fierce struggle for power ensued between the Miloslavskys and the Naryshkins: the former wanted to put Fyodor's brother, the delicate and feebleminded Ivan V, on the throne; the Naryshkins stood for the healthy and intelligent Peter. Representatives of the various orders of society, assembled in the Kremlin, declared themselves for Peter, who was then proclaimed tsar; but the Miloslavsky faction exploited a revolt of the Moscow *streltsy,* or musketeers of the sovereign's bodyguard, who killed some of Peter's adherents, including Matveyev. Ivan and Peter were then proclaimed joint tsars; and eventually, because of Ivan's precarious health and Peter's youth, Ivan's 25-year-old sister Sophia was made regent. Clever and influential, Sophia took control of the government; excluded from public affairs, Peter lived with his mother in the village of Preobrazhenskoye, near Moscow, often fearing for his safety. All this left an ineradicable impression on the young tsar and determined his negative attitude toward the *streltsy.*

One result of Sophia's overt exclusion of Peter from the government was that he did not receive the usual education of a Russian tsar; he grew up in a free atmosphere instead of being confined within the narrow bounds of a palace. While his first tutor, the former church clerk Nikita Zotov, could give little to satisfy Peter's curiosity, the boy enjoyed noisy outdoor games and took especial interest in military matters, his favourite toys being arms of one sort or another. He also occupied himself with carpentry, joinery, blacksmith's work, and printing.

Boyhood influences

By courtesy of the Rijksmuseum, Amsterdam



Peter I the Great, portrait by Aert de Gelder (1645–1727). In the Rijksmuseum, Amsterdam.

Near Preobrazhenskoye there was a *nemetskaya sloboda* ("German colony") where foreigners were allowed to reside. Acquaintance with its inhabitants aroused Peter's interest in the life of other nations, and an English sailboat, found derelict in a shed, whetted his passion for seafaring. Mathematics, fortification, and navigation were the sciences that appealed most strongly to Peter. A model fortress was built for his amusement, and he organized his first "play" troops, from which, in 1687, the Preobrazhensky and Semyonovsky Guards regiments were formed—to become the nucleus of a new Russian Army.

Early in 1689 Natalya Naryshkina arranged Peter's marriage to the beautiful Eudoxia (Yevdokiya Fyodorovna Lopukhina). This was obviously a political act, intended to demonstrate the fact that the 17-year-old Peter was now a grown man, with a right to rule in his own name. The marriage did not last long: Peter soon began to ignore his wife, and in 1698 he relegated her to a convent.

In August 1689 a new revolt of the *streltsy* took place. Sophia and her faction tried to use it to their own advantage for another coup d'état, but events this time turned decisively in Peter's favour. He removed Sophia from power and banished her to the Novodevichy convent; she was forced to become a nun after a *streltsy* rebellion in 1698. Though Ivan V remained nominally joint tsar with Peter, the administration was now largely given over to Peter's kinsmen, the Naryshkins, until Ivan's death in 1696. Peter, meanwhile continuing his military and nautical amusements, sailed the first seaworthy ships to be built in Russia. His games proved to be good training for the tasks ahead.

Peter as sole ruler

External events. At the beginning of Peter's reign, Russia was territorially a huge power, but with no access to the Black Sea, the Caspian, or to the Baltic, and to win such an outlet became the main goal of Peter's foreign policy.

The Azov campaigns (1695–96). The first steps taken in this direction were the campaigns of 1695 and 1696, with the object of capturing Azov from the Crimean Tatar vassals of Turkey. On the one hand, these Azov campaigns could be seen as fulfilling Russia's commitments, undertaken during Sophia's regency, to the anti-Turkish "Holy League" of 1684 (Austria, Poland, and Venice); on the other they were intended to secure the southern frontier against Tatar raids, as well as to approach the Black Sea. The first campaign ended in failure (1695), but this did not discourage Peter: he promptly built a fleet at Voronezh to sail down the Don River and in 1696 Azov was captured. To consolidate this success Taganrog was founded on the northern shore of the Don Estuary, and the building of a large navy was started.

The Grand Embassy (1697–98). Having already sent some young nobles abroad to study nautical matters, Peter, in 1697, went with the so-called Grand Embassy to western Europe. The embassy comprised about 250 people, with the "grand ambassadors" Franz Lefort, F.A. Golovin, and P.B. Voznitsyn at its head. Its chief purposes were to examine the international situation and to strengthen the anti-Turkish coalition, but it was also intended to gather information on the economic and cultural life of Europe. Travelling incognito under the name of Sgt. Pyotr Mikhaylov, Peter familiarized himself with conditions in the advanced countries of the West. For four months he studied shipbuilding, working as a ship's carpenter in the yard of the Dutch East India Company at Saardam; after that he went to Great Britain, where he continued his study of shipbuilding, working in the Royal Navy's dockyard at Deptford, and he also visited factories, arsenals, schools, and museums and even attended a session

Familiarization with the West

of Parliament. Meanwhile, the services of foreign experts were engaged for work in Russia.

On the diplomatic side of the Grand Embassy, Peter conducted negotiations with the Dutch and British governments for alliances against Turkey; but the Maritime Powers did not wish to involve themselves with him because they were preoccupied with the problems that were soon to come to a crisis, for them, in the War of the Spanish Succession.

The destruction of the streltsy (1698). From England, Peter went on to Austria; but while he was negotiating in Vienna for a continuance of the anti-Turkish alliance, he received news of a fresh revolt of the *streltsy* in Moscow. In the summer of 1698 he was back in Moscow, where he suppressed the revolt. Hundreds of the *streltsy* were executed, the rest of the rebels were exiled to distant towns, and the corps of the *streltsy* was disbanded.

The Northern War (1700–21). When it became clear that Austria, no less than the Maritime Powers, was preparing to fight for the Spanish Succession and to make peace with Turkey, Peter saw that Russia could not contemplate a war without allies against the Turks, and he abandoned his plans for pushing forward from the Sea of Azov to the Black Sea. By the Russo-Turkish Peace of Constantinople (Istanbul, 1700) he retained possession of Azov. He was now turning his attention to the Baltic instead, following the tradition of his predecessors.

The Swedes occupied Karelia, Ingria, Estonia, and Livonia and blocked Russia's way to the Baltic coast. To dislodge them, Peter took an active part in forming the great alliance, comprising Russia, Saxony, and Denmark–Norway, which started the Northern War in 1700. This war lasted for 21 years and was Peter's main military enterprise. In planning it and in sustaining it he displayed iron willpower, extraordinary energy, and outstanding gifts of statesmanship, generalship, and diplomacy. Mobilizing all the resources of Russia for the triumph of his cause, constantly keeping himself abreast of events, and actively concerning himself with all important undertakings, often at his personal risk, he could be seen sometimes in a sailor's jacket on a warship, sometimes in an officer's uniform on the battlefield, and sometimes in a labourer's apron and gloves with an axe in a shipyard.

Peter's main military enterprise

The defeat of the Russians at Narva (1700), very early in the war, did not deter Peter and, in fact, he later described it as a blessing: "Necessity drove away sloth and forced me to work night and day." He subsequently took part in the siege that led to the Russian capture of Narva (1704) and in the battles of Lesnaya (1708) and of Poltava (1709). At Poltava, where Charles XII of Sweden suffered a catastrophic defeat, the plan of operations was Peter's own: it was his idea to transform the battlefield by works of his military engineers—the redoubts erected in the path of the Swedish troops to break their combat order, to split them into little groups, and to halt their onslaught. Peter also took part in the naval battle of Gangut (Hanko, or Hangö) in 1714, the first major Russian victory at sea.

The treaties concluded by Russia in the course of the war were made under Peter's personal direction. He also travelled abroad again for diplomatic reasons—*e.g.*, to Pomerania in 1712 and to Denmark, northern Germany, Holland, and France in 1716–17.

In 1703, on the banks of the Neva River, where it flows into the Gulf of Finland, Peter began construction of the city of St. Petersburg (now Leningrad) and established it as the new capital of Russia in 1712. By the Treaty of Nystad (September 10 [August 30, O.S.], 1721) the eastern shores of the Baltic were at last ceded to Russia, Sweden was reduced to a secondary power, and the way was opened for Russian domination over Poland.

In celebration of his triumph, the Senate on November 2 (October 22, O.S.), 1721, changed Peter's title from tsar to that of emperor (imperator) of all the Russias.

The popular revolts (1705–08). The peasant serfs and the poorer urban workers had to bear the greatest hardships in wartime and moreover were intensively exploited in the course of Peter's great work for the modernization and development of Russia (see below *Internal reforms*). Their sufferings, combined with onerous taxation, provoked a number of revolts, the most important of which were that of Astrakhan (1705–06) and that led by Kondraty Afanasyevich Bulavin in the Don Basin (1707–08). These revolts were cruelly put down.

Other events during the Northern War

The Turkish War (1710–13). In the middle of the Northern War, when Peter might have pressed further the advantage won at Poltava, Turkey declared war on Russia. In the summer of 1711 Peter marched against the Turks through Bessarabia into Moldavia, but he was surrounded, with all his forces, on the Prut River. Obliged to sue for peace, he was fortunate to obtain very light terms from the inept Turkish negotiators, who allowed him to retire with no greater sacrifice than the retrocession of Azov. The Turkish government soon decided to renew hostilities; but the Peace of Adrianople (Edirne) was concluded in 1713, leaving Azov to the Turks. From that time on Peter's military effort was concentrated on winning his war against Sweden.

The Tsarevich Alexis and Catherine (to 1718). Peter had a son, the tsarevich Alexis, by his discarded wife Eudoxia. Alexis was his natural heir, but he grew up antipathetic to Peter and receptive to reactionary influences working against Peter's reforms. Peter, meanwhile, had formed a lasting liaison with a low-born woman, the future empress Catherine I, who bore him other children and whom he married in 1712. Pressed finally to mend his ways or to become a monk in renunciation of his hereditary rights (1716), Alexis took refuge in the dominions of the Holy Roman emperor Charles VI, but he was induced to return to Russia in 1718. Thereupon proceedings were brought against him on charges of high treason, and after torture he was condemned to death. He died in prison, presumably by violence, before the formal execution of the sentence.

The Persian campaign (1722–23). Even during the second half of the Northern War, Peter had sent exploratory missions to the East—to the Central Asian steppes in 1714, to the Caspian region in 1715, and to Khiva in 1717. The end of the war left him free to resume a more active policy on his southeastern frontier. In 1722, hearing that the Ottoman Turks would take advantage of Persia's weakness and invade the Caspian region, Peter himself invaded Persian territory. In 1723 Persia ceded the western and southern shores of the Caspian to Russia in return for military aid.

Death. The campaign along the parched shores of the Caspian obviously put a great strain on Peter's health, already undermined by enormous exertions and also by the excesses in which he occasionally indulged himself. In the autumn of 1724, seeing some soldiers in danger of drowning from a ship aground on a sandbank in the Gulf of Finland, he characteristically plunged himself into the icy water to help them. Catching a chill, he became seriously ill in the winter but even so continued to work; indeed, it was at this time that he drew up the instructions for the expedition of Vitus Bering to Kamchatka.

Peter died in St. Petersburg on February 8 (January 28, O.S.), 1725, leaving an empire that stretched from Arkhangelsk (Archangel) on the White Sea to Mazanderan on the Caspian and from the Baltic Sea to the Pacific Ocean. Though he had in 1722 issued a decree reserving to himself the right to nominate his successor, he did not in fact nominate anyone. His widow Catherine, whom he had crowned as empress in 1724, succeeded him to the temporary exclusion of his grandson, the future Peter II.

The empire at Peter's death

Internal reforms. At the beginning of Peter's reign, Russia was backward by comparison with the countries of western Europe. This backwardness inhibited foreign policy and even put Russia's national independence in danger. Peter's aim, therefore, was to overtake the developed countries of western Europe as soon as possible, in order both to promote the national economy and to ensure victory in his wars for access to the seas. Breaking the resistance of the boyars, or members of the ancient landed aristocracy, and of the clergy and severely punishing all other opposition to his projects, he initiated a series of reforms that affected, in the course of 25 years, every field of the national life—administration, industry, commerce, technology, and culture.

The towns. At the beginning of Peter's reign there was

already some degree of economic differentiation between the various regions of Russia; and in the towns artisans were establishing small businesses, small-scale production was expanding, and industrial plants and factories were growing up, with both hired workers and serfs employed. There was thus a nascent bourgeoisie, which benefitted considerably from Peter's plans for the development of the national industry and trade. The reform of the urban administration was particularly significant.

Administrative reforms

By a decree of 1699, townspeople (artisans and tradesmen) were released from subjection to the military governors of the provinces and were authorized to elect municipalities of their own, which would be subordinated to the Moscow municipality, or *ratusha*—the council of the great merchant community of the capital. This reform was carried further in 1720, with the establishment of a chief magistracy in St. Petersburg, to which the local town magistracies and the elected municipal officers of the towns (mayors, or *burmistry,* and councillors, or *ratmany*) were subordinated.

All townspeople, meanwhile, were divided between "regulars" and "commons" (inferiors). The regulars were subdivided between two guilds—the first comprising rich merchants and members of the liberal professions (doctors, actors, and artists); the second, artisans (classified according to their vocations) and small tradesmen. A merchant belonged to the first or to the second guild according to the amount of his capital; and those who were also manufacturers had special privileges, coming under the jurisdiction of the College of Manufactures and being exempt from the billeting of troops, from elective rotas of duty, and from military service. The commons were hired labourers, without the privileges of regulars.

Thanks to the reforms, the economic activity and the population of the towns increased. Anyone engaged in trade was legally permitted to settle in a town and to register himself in the appropriate category, and there was a right of "free commerce for people of every rank."

The provinces and the districts. In order to create a more flexible system of control by the central power, Russia was territorially divided in 1708 into eight *guberny*, or governments, each under a governor appointed by the tsar and vested with administrative, military, and judicial authority. In 1719 these *guberny* were dissolved into 50 *provintsy,* or provinces, which in turn were subdivided into districts. The census of 1722, however, was followed by the substitution of a poll tax for the previous hearth tax; and this provoked a wave of popular discontent, against which Peter decided to distribute the army regiments (released from active service by the Peace of Nystad) in garrisons throughout the country and to make their maintenance obligatory on the local populations. Thus came into being the "regimental districts," which did not coincide with the administrative. The regimental commanders, with their own sphere of jurisdiction and their own requirements, added another layer to the already complex system of local authority.

The central government. In the course of Peter's reign, medieval and obsolescent forms of government gave place to effective autocracy. In 1711 he abolished the *boyarskaya duma,* or boyar council, and established by decree the Senate as the supreme organ of state—to coordinate the action of the various central and local organs, to supervise the collection and expenditure of revenue, and to draft legislation in accordance with his edicts. Martial discipline was extended to civil institutions, and an officer of the guards was always on duty in the Senate. From 1722, moreover, there was a procurator general keeping watch over the daily work of the Senate and its chancellery and acting as "the eye of the sovereign."

When Peter came to power the central departments of state were the *prikazy,* or offices, of which there were about 80, functioning in a confused and fragmented way. To replace most of this outmoded system, Peter in 1718 instituted 9 "colleges" (*kollegy*), or boards, the number of which was by 1722 expanded to 13. Their activities were controlled, on the one hand, by the General Regulation and, on the other, by particular regulations for individual colleges, and indeed there were strict regulations for every branch of the state administration. Crimes against the state came under the jurisdiction of the Preobrazhensky Office, responsible immediately to the tsar.

Industry. A secondary purpose of Peter's Grand Embassy to western Europe in 1697 (see above *The Grand Embassy*) had been to obtain firsthand acquaintance with advanced industrial techniques, and the exigencies of his great war against Sweden, from 1700, made industrial development an urgent matter. In order to provide armaments and to build his navy (Russia had virtually no warships at all), metallurgical and manufacturing industries on a grand scale had to be created; and Peter devoted himself tirelessly to meeting these needs. Large capital investments were made, and numerous privileges were accorded to businessmen and industrialists. These privileges included the right to buy peasant serfs for labour in workshops, with the result that a class of "enlisted" serfs came into existence, living in specified areas and bound to the factories. The methods of other countries were further studied, and foreign experts were invited to Russia. The overall result was satisfactory: the army and the navy were supplied with their material needs; a great number of manufacturing establishments were founded (mainly with serf labour); the metallurgical industry was so far advanced that by the middle of the 18th century Russia led Europe in this field; and the foreign-trade turnover was increased sevenfold in the course of the reign.

Europeanization of industry

The armed forces. Peter established a regular army on completely modern lines for Russia in the place of the unreliable *streltsy* and the militia of the gentry. While he drew his officers from the nobility, he conscripted peasants and townspeople into the other ranks. Service was for life. The troops were equipped with flintlock firearms and bayonets of Russian make; uniforms were provided; and regular drilling was introduced. For the artillery, obsolete cannons were replaced with new mortars and guns designed by Russian specialists or even by Peter himself (he drew up projects of his own for multicannon warships, fortresses, and ordnance). The Army Regulations of 1716 were particularly important; they required officers to teach their men "how to act in battle," "to know the soldier's business from first principles and not to cling blindly to rules," and to show initiative in the face of the enemy. For the navy, Peter's reign saw the construction, within a few years, of 52 battleships and hundreds of galleys and other craft; thus a powerful Baltic fleet was brought into being. Several special schools prepared their pupils for military or naval service and finally enabled Peter to dispense with foreign experts.

Cultural and educational measures. From January 1, 1700, Peter introduced a new chronology, making the Russian calendar conform to European usage with regard to the year, which in Russia had hitherto been numbered "from the Creation of the World" and had begun on September 1 (he adhered however to the Julian Old Style as opposed to the Gregorian New Style for the days of the month). In 1710 the Old Church Slavonic alphabet was modernized into a secular script.

Peter was the first ruler of Russia to sponsor education on secular lines and to bring an element of state control into that field. Various secular schools were opened; and since too few pupils came from the nobility, the children of soldiers, officials, and churchmen were admitted to them. In many cases, compulsory service to the state was preceded by compulsory education for it. Russians were also permitted to go abroad for their education and indeed were often compelled to do so (at the state's expense). The translation of books from western European languages was actively promoted. The first Russian newspaper, *Vedomosti* ("Records"), appeared in 1703. The Russian Academy of Sciences was instituted in 1724.

Beside his useful measures, Peter often enforced superficial Europeanization rather brutally; for example, when he decreed that beards should be shorn off and Western dress worn. He personally cut the beards of his boyars and the skirts of their long coats (*kaftany*). The Raskolniki (Old Believers) and merchants who insisted on keeping their beards had to pay a special tax, but peasants and the Orthodox clergy were allowed to remain bearded.

Control over church and nobility

The church. In 1721, in order to subject the Orthodox Church of Russia to the state, Peter abolished the Patriarchate of Moscow. Thenceforward the patriarch's place as head of the church was taken by a spiritual college, namely the Holy Synod, consisting of representatives of the hierarchy obedient to the tsar's will. A secular official—the *ober-prokuror,* or chief procurator—was appointed by the tsar to supervise the Holy Synod's activities. The Holy Synod ferociously persecuted all dissenters and conducted a censorship of all publications.

Priests officiating in churches were obliged by Peter to deliver sermons and exhortations that were intended to make the peasantry "listen to reason" and to teach such prayers to children that everyone would grow up "in fear of God" and in awe of the tsar. The regular clergy were forbidden to allow men under 30 years old or serfs to take vows as monks.

The church was thus transformed into a pillar of the absolutist regime. Partly in the interests of the nobility, the extent of land owned by the church was restricted; Peter disposed of ecclesiastical and monastic property and revenues at his own discretion, for state purposes.

The nobility. Peter's internal policy served to protect the interest of Russia's ruling class—the landowners and the nascent bourgeoisie. The material position of the landed nobility was strengthened considerably under Peter. Almost 100,000 acres of land and 175,000 serfs were allotted to it in the first half of the reign alone. Moreover, a decree of 1714 that instituted succession by primogeniture and so prevented the breaking up of large properties also removed the old distinction between *pomestya* (lands granted by the tsar to the nobility in return for service) and *votchiny* (patrimonial or allodial lands) so that all such property became hereditary.

Moreover, the status of the nobility was modified by Peter's Table of Ranks (1722). This replaced the old system of promotion in the state services, which had been according to ancestry, by one of promotion according to services actually rendered. It classified all functionaries—military, naval, and civilian alike—in 14 categories, the 14th being the lowest and the 1st the highest; and admission to the 8th category conferred hereditary nobility. Factory owners and others who had risen to officer's rank could accede to the nobility, which thus received new blood. The predominance of the boyars ended.

Personality and achievement. Peter was of enormous height, more than six and one-half feet (two metres) tall; he was handsome and of unusual physical strength. Unlike all earlier Russian tsars, whose Byzantine splendours he repudiated, he was very simple in his manners; for example, he enjoyed conversation over a mug of beer with shipwrights and sailors from the foreign ships visiting St. Petersburg. Restless, energetic, and impulsive, he did not like splendid clothes that hindered his movements; often he appeared in worn-out shoes and an old hat, still more often in military or naval uniform. He was fond of merrymaking and knew how to conduct it, though his jokes were frequently crude; and he sometimes drank heavily and forced his guests to do so too. A just man who did not tolerate dishonesty, he was terrible in his anger and could be cruel when he encountered opposition: in such moments only his intimates could soothe him—best of all his beloved second wife, Catherine, whom people frequently asked to intercede with him for them. Sometimes Peter would beat his high officials with his stick, from which even Prince A.D. Menshikov, his closest friend, received many a stroke. One of Peter's great gifts of statesmanship was the ability to pick talented collaborators for the highest appointments, whether from the foremost families of the nobility or from far lower levels of society.

As a ruler, Peter often used the methods of a despotic landlord—the whip and arbitrary rule. He always acted as an autocrat, convinced of the wonder-working power of compulsion by the state. Yet with his insatiable capacity for work he saw himself as the state's servant, and whenever he put himself in a subordinate position he would perform his duties with the same conscientiousness that he demanded of others. He began his own army service in the lowest rank and required others likewise to master their profession from its elements upward and to expect promotion only for services of real value.

Peter's influence on Russia

Peter's personality left its imprint on the whole history of Russia. A man of original and shrewd intellect, exuberant, courageous, industrious, and iron-willed, he could soberly appraise complex and changeable situations so as to uphold consistently the general interests of Russia and his own particular designs. He did not completely bridge the gulf between Russia and the Western countries, but he achieved considerable progress in development of the national economy and trade, education, science and culture, and foreign policy. Russia became a great power, without whose concurrence no important European problem could thenceforth be settled. His internal reforms achieved progress to an extent that no earlier innovator could have envisaged.

BIBLIOGRAPHY. *Pis'ma i bumagi Imperatora Petra Velikogo,* 11 vol. (1887–1964), contains Peter's correspondence as well as valuable documents on Russian history up to 1711. Biographies include M.M. BOGOSLOVSKIĬ, *Petr I,* 5 vol. (1940–48, reissued 1969), a detailed study up to 1700; IAN GREY, *Peter the Great, Emperor of All Russia* (1960); M.S. ANDERSON, *Peter the Great* (1978); ALEX DE JONGE, *Fire and Water: A Life of Peter the Great* (1979); ROBERT K. MASSIE, *Peter the Great: His Life and World* (1980, reprinted 1991); and HENRI TROYAT, *Peter the Great* (1987; originally published in French, 1979), a popularized account.

Peter's reign and the reforms he instituted are analyzed in SERGEĬ M. SOLOV'EV, *Publichnyia chteniia o Petrie Velikom* (1872, reissued 1984), by a famous Russian historian; B.H. SUMNER, *Peter the Great and the Emergence of Russia* (1950, reissued 1972); REINHARD WITTRAM, *Peter I, Czar und Kaiser,* 2 vol. (1964), and *Peter der Grosse: der Eintritt Russlands in die Neuzeit* (1954); IVAN I. GOLIKOV, *Dieianiia Petra Velikago,* 2nd ed., 15 vol. (1837–43), on his reforms; JAMES CRACRAFT, *The Church Reform of Peter the Great* (1971); ALEXANDER V. MULLER (ed. and trans.), *The Spiritual Regulation of Peter the Great,* trans. from Russian (1972); and EVGENII V. ANISIMOV, *The Reforms of Peter the Great: Progress Through Coercion in Russia* (1993; originally published in Russian, 1989). J.G. GARRARD (ed.), *The Eighteenth Century in Russia* (1973), provides a collection of essays on different aspects of the Westernization of Russia. Peter's military campaigns and his role as the founder of the new Russian army are explored in the works of a prominent Soviet historian, EVGENIĬ V. TARLE, *Russkiĭ flot i vneshniaia politika Petra I* (1949), also available in a German translation, *Russisch-englische Beziehungen unter Peter I* (1954), and *Severnaia voĭna i shvedskoe nashestvie na Rossiiu* (1958). Foreign relations are described by LEONID A. NIKIFOROV, *Russko-angliĭskie otnosheniia pri Petre I* (1950); and B.H. SUMNER, *Peter the Great and the Ottoman Empire* (1949, reissued 1965).

XENIA GASIOROWSKA, *The Image of Peter the Great in Russian Fiction* (1979), is a study of some 60 historical novels written since Pushkin's time. NICHOLAS V. RIASANOVSKY, *The Image of Peter the Great in Russian History and Thought* (1985), examines Peter's influence and how he has been perceived in Russia from 1700 to 1983.

Works that put Peter the Great and his reign into historical perspective include VASILI KLYUCHEVSKY, *The Rise of the Romanovs,* trans. and ed. by LILIANA ARCHIBALD (1970; originally published in Russian, 1912); E.M. ALMEDINGEN, *The Romanovs: Three Centuries of an Ill-Fated Dynasty* (1966); JOHN D. BERGAMINI, *The Tragic Dynasty: A History of the Romanovs* (1969), based on English- and French-language sources; IAN GREY, *The Romanovs: The Rise and Fall of a Dynasty* (1970); and W. BRUCE LINCOLN, *The Romanovs: Autocrats of All the Russias* (1981).

(L.A.N./Ed.)

Philadelphia

Philadelphia has been described as the elegant but rather jaded great lady of American cities. A more realistic look at Philadelphia, however, shows it to be a very modern and vigorous city, arising in gracious counterpoint to the deep serenity of an older city that has provided gentle but firm intellectual, economic, and humanitarian direction to the nation at whose birth it played midwife.

Philadelphia, which is the fifth largest city in the United States and the largest in Pennsylvania, displays many characteristics of a small town. Its many trees, parks, and other open spaces and its quiet pace of life reflect in various ways the genteel Quaker heritage bestowed on the city by its founder, William Penn. Nearly everywhere are dignified reminders of the colonial and Revolutionary city and of Benjamin Franklin, a Philadelphian by adoption, who left his imprint on innumerable ongoing institutions, both cultural and commercial, in the city.

Beneath this facade, however, Philadelphia represents an urban cluster of national and international stature. The Port of Philadelphia and Camden is one of the largest freshwater ports in the world and one of the busiest shipping centres along the nation's East Coast, specializing in international cargo. The enormous industrial production of the city and the surrounding metropolitan area represents a continuation of Philadelphia's early leadership in the Industrial Revolution and in American commerce and finance generally. Lying in the midst of the vast urban community stretching down the Atlantic Seaboard, Philadelphia is an integral part of the turbulent fabric of contemporary social and economic life as well as a tranquil oasis joining together the spirit of America, past and present.

This article is divided into the following sections:

Physical and human geography 532
- The landscape 532
  - The city site
  - The city layout
- The people 533
- The economy 533
  - Industry
  - Research
  - Finance
  - Transportation
- Administration and social conditions 534
  - Government
  - Public services
  - Education
- Cultural life 534

History 535
- Foundation and early settlement 535
- The growth of the city 535
  - Cultural dominion
  - Political evolution
  - The 20th century

Bibliography 536

Physical and human geography

THE LANDSCAPE

The city site. Philadelphia spreads over about 136 square miles (352 square kilometres) from the Delaware on the east across the Schuylkill and beyond. The coextensive boundaries of Philadelphia city and county remain essentially as defined by the Consolidation Act of 1854.

The city layout. The gridiron of streets in central Philadelphia, the first U.S. city to be so arranged, follows the original plans of William Penn. Midway between the two rivers, Penn Square, occupied by City Hall, is the centre of the plan. Four shady, fountained squares—Logan, Franklin, Washington, and Rittenhouse—dot the quadrants. Westward from Penn Square along John F. Kennedy Boulevard is Penn Center, a prime symbol of rejuvenation in Philadelphia since the 1950s. The multilevel complex comprises high-rise offices and hotels, with interior courts and malls and underground walkways lined with shops and restaurants. A statue of Penn stands atop City Hall, and for decades Philadelphia developers agreed to construct buildings no higher than the top of his hat. However, in the late 1980s taller buildings began to appear, such as One Liberty Place, which rises some 940 feet (287 metres) in 61 stories.

Benjamin Franklin Parkway provides a splendid vista as it cuts diagonally northwestward from Penn Square through the grid, encircling Logan Square and leading into Fairmount Park. The nation's largest landscaped park within city limits and the centre of the Centennial Exposition of 1876, Fairmount is one of the most frequent foregrounds for photographs of Philadelphia's skyline, adding to the city's reputation for shaded, sculpted elegance. Once a section of outlying estates of the wealthy, the park contains many fine old mansions maintained by the city as museums. Through it winds the narrow valley of Wissahickon Creek, whose rugged beauty has inspired generations of poets and painters.

The oldest sections of Philadelphia—Southwark, Society Hill, and the Independence Hall area—lie to the east, along and inland from the Delaware. Southwark is the oldest, having been settled by Swedes in 1643. Those of its ancient and dilapidated houses that have escaped bulldozing for riverfront expressways resemble the edifices of Society Hill before its restoration in the 1950s. This latter area contains some of the city's finest old houses, taverns, and churches, though some high-rise apartment buildings strike a dissonant note in the hearts of many Philadelphians.

Historic preservations and restorations

Independence National Historical Park, established in 1956 and designated a World Heritage site in 1979, contains Independence Hall, with the nearby Liberty Bell Pavilion, and many other buildings used by the Revolutionary and early federal governments. Nearby, Elfreth's Alley, dating from 1690, contains 33 houses that make up the oldest continously inhabited street in the country. Included in the complex are Carpenters' Hall, site of the meeting of the First Continental Congress, and Philosophical Hall, home of the American Philosophical Society.

Throughout these areas and elsewhere, domestic architecture is characterized primarily by two- and three-story redbrick structures fronting directly on the sidewalks and containing impressive examples of colonial design both outside and inside. The Philadelphia Historical Commission has certified about 7,500 buildings and structures, ranging from 17th-century houses to a bridge constructed in 1950. Numerous buildings throughout the city have been restored, but since the late 1990s many others have been demolished to make way for new buildings and parking lots. Even in its downtown section, Philadelphia continues to be a city of shops rather than of huge merchandising outlets. It has nothing approaching New York City's Fifth Avenue as a street of large stores. Chestnut Street is an attractive street of shops but has only one major department store. Pine and other streets are noted for such specialties as antique shops. Major features of the skyline are the massive contours of the city's many banking and financial institutions.

The long stretches of Philadelphia north, south, and west

of the intersection of Broad and Chestnut streets, a "city hub" of sorts just below Penn Square, contain numerous distinctive sections, often identified for generations with various ethnic groups that have filled the city during its long history. Among the more notable is the Germantown section of North Philadelphia, settled in Penn's time by Germans and the home in the 18th century of wealthy Philadelphians fleeing the periodic yellow-fever epidemics of the riverside city; the area is now ethnically mixed, with several middle-class African American neighbourhoods. North Philadelphia has large African American and Puerto Rican communities. South Philadelphia contains notable Italian, Irish, and other sections settled by European immigrants mainly in the 19th century.

THE PEOPLE

Though Philadelphia has had most of the characteristics of an ethnic and racial melting pot nearly from its start, it lacks the steaming hurly-burly visible everywhere in its behemoth neighbour, New York City, about 90 miles (150 kilometres) northeast. The Philadelphians by and large are not a street people, and their orientation has been so traditionally toward the home that the city became known as a "city of homes." Philadelphians are also great joiners, giving the city more social and other clubs than any other in America. Many of these are based in ethnic neighbourhoods, others are citywide, and still others serve the affluent "Main Liners" who reside in the plush suburbs that grew up in past centuries along the main line of the former Pennsylvania Railroad.

Much of this tradition dates from Philadelphia's early years, from the beliefs and attitudes not only of the Quakers but also of such German pietist sects as the Mennonites, who stressed personal religious experience rather than institutionalized formulations. A basic sobriety underlying these tenets led to many stringent laws that remain in both the city and state. On the other hand, these groups strove for tolerance in all matters. Pennsylvania was among the few colonies admitting Roman Catholics and Jews, and the Quakers long were leaders in seeking social justice and the alleviation of inhumanity in ethnic and other human relations. Philadelphia still has several endowed charities of Quaker origin to provide the poor with food—as well as one of the oldest and strongest municipal commissions on human relations.

Quaker tradition and ethnic mix

During the 1800s the Protestant sects were joined by Roman Catholics, initially Irish and German and later Italian, Polish, and Slavic. Eastern European Jews also immigrated, eventually comprising a significant portion of the city's population. Cultural conflicts and competition for housing and jobs created tensions between the oldstock groups and the new arrivals. A series of anti-Catholic nativist riots rocked Philadelphia in the Jacksonian period, with an especially violent incident in 1844. Although immigrants came to Philadelphia in the hundreds of thousands in the late 19th and early 20th centuries, the city had a relatively low proportion of foreign-born when compared with other major cities at that time. Some Irish, Italian, Polish, and Jewish newcomers did find opportunities for considerable economic advancement in Philadelphia; for many, however, hopes never materialized and the urban experience proved rough and oftentimes shattering. The slow progress of these ethnic groups toward full acceptance is represented by the fact that Philadelphia did not elect its first Irish mayor until 1963 and its first Italian mayor until 1971.

Philadelphia has been the focal point of one of the historically most important African American communities in the nation. During the early national and antebellum periods, many black leaders came from or centred their activities in Philadelphia, and the free black community there was intensely involved in many efforts to abolish slavery, assist fugitive slaves, and advance the general social and economic well-being of blacks. Between 1829 and 1860, however, the Philadelphia black community was victimized by several racist mob actions. At the beginning of the 20th century Philadelphia was the site of the largest African American community of any northern city. The migration of Southern-born blacks to the city continued to augment the city's population throughout the first half of the 20th century. Discrimination in housing resulted in the creation of overcrowded black districts. By the late 20th century about two-fifths of all Philadelphians and in excess of half of the public school population were black. The suburbs, by contrast, were almost entirely white.

Problems of substandard and insufficient housing, public health, crime, drug abuse, and other ills, by no means confined to the ghettos, somewhat dimmed the city's image in the late 20th century. Efforts at public housing date back to 1937. Other municipal and federal bodies have continued efforts at urban renewal.

THE ECONOMY

Industry. Many of Philadelphia's business enterprises were established in the 1700s, but the city's economy has changed greatly. Such old and once-dominant industries as textile mills met stiff competition from the South. Manufacturing was given a boost by World War II but began to decline later; thousands of jobs were lost in the 1960s. Nonmanufacturing and services, however, continue to increase employment and now dominate the economy. High-technology industries such as communications, computer software, and Internet commerce became important in the late 20th century. Printing, publishing, and food processing are also major industries. Manufactures include chemicals, industrial machinery, fabricated metal products, electronics, transportation equipment, scientific instruments, apparel, paper products, and rubber and plastic products. The Philadelphia Naval Shipyard, which once employed more than 10,000 people, was closed in 1996; two years later a smaller commercial shipbuilding concern began production at the site.

An economic overview

From the mid-20th century Philadelphia diminished in comparison with its outlying metropolitan ring, as large numbers of people and businesses migrated from the city to the suburbs. The Valley Forge area has become a nest of business and industrial locations. Much of this development is under the sponsorship of the nonprofit Philadelphia Industrial Development Corporation.

Research. George Washington approved the first U.S. patent in 1790, to a Philadelphian named Samuel Hopkins for a better way of making potash. In today's economy research and invention are keys to progress. In West Philadelphia is located the University City Science Center, a nonprofit project of 30 academic and scientific institutions with a multimillion-dollar commitment to research. General Electric, E.I. du Pont de Nemours, Unisys, Monsanto, Rohm and Haas, and Exelon are other large concerns with extensive research programs within the metropolitan area. The region is one of the country's largest health education and research centres; several major biotechnology and pharmaceutical companies have research facilities there. There is also significant research on petroleum refining. The computer age was born out of experiments at the University of Pennsylvania, where the first electronic computer was operated in 1946. A unit of the Boeing Company at Ridley Park is one of the world's most advanced plants for helicopter research and design.

Finance. From its early days until the 1850s, Philadelphia was the financial capital of the United States, but it then lost this position to New York City. It was the birthplace of American banking, and the first building and loan association was founded there. The city continues to be a major financial centre, with its institutions remaining the oldest and among the most prosperous in the nation. Philadelphia also supplies the stock and exchange services demanded by modern business. The Philadelphia Commercial Exchange was set up in 1868 as an outgrowth of the older Corn Exchange Association of 1863 to regulate the then-flourishing grain and flour trade. The Philadelphia Bourse was organized in 1891, and the Maritime Exchange was founded in 1875. Philadelphia has the oldest stock exchange (founded in 1790) in the United States. The Philadelphia–Baltimore–Washington Stock Exchange was joined by Pittsburgh in 1969 and in 1976 was renamed the Philadelphia Stock Exchange. A U.S. Mint is also located in the city.

Transportation. A complex system of public and private trolley and bus lines was consolidated in 1963 by the state legislature's creation of the Southeastern Pennsylvania Transportation Authority (SEPTA) to plan, develop, and coordinate a regional transportation system and to fund projects by the sale of bonds. Express buses provide rapid service between the inner and outer city.

There are several toll- and tax-supported bridges over the Delaware, two of which are solely for pedestrian use. The Schuylkill is bridged at a number of points and has a subway tunnel. The Delaware River Port Authority administers the Port of Philadelphia and Camden; the Benjamin Franklin, Walt Whitman, Betsy Ross, and Commodore Barry bridges over the Delaware; a rapid-transit line between Philadelphia and Lindenwold, N.J.; and a ferry on the Delaware. Philadelphia International Airport is less than seven miles (11 kilometres) from the city centre.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Philadelphia's government centres on a mayor and a city council. The mayor, who is the chief executive for the city, is directly elected to a four-year term and may seek reelection to one additional term. The mayor has the power to initate and veto legislation, review and plan the day-to-day operations of the government, and determine fiscal policy. The city council consists of 17 members, seven of whom are elected at-large and 10 according to electoral district; members of the council serve four-year terms and may seek reelection

Public services. Though the Philadelphia metropolitan area has a per capita income that is the highest of any such area in the state, a large percentage of Philadelphians live below the poverty level. City, state, and federal agencies administer a full range of ameliorative social-service programs. A department of public health operates a variety of services through a number of health districts. With its Quaker tradition, Philadelphia is known for its humanitarian concern and has a great variety of privately supported child-care, hospital, and other social services.

Education. Philadelphia's public school system is an independent governmental unit operating under a board of education appointed by the mayor on the recommendation of a citizens' nominating panel. The city's institutions of higher learning are headed by the privately endowed University of Pennsylvania, which was founded in 1740, opened the nation's first medical school (1765), and was the first designated "university" in the United States (1779). Temple University (1884) has acquired the reputation of "the people's college" and is known for work in the arts and social sciences. Roman Catholic universities include LaSalle and St. Joseph's. Drexel University was founded in 1891. Such liberal arts colleges as Haverford (1833), Swarthmore (1864), and Bryn Mawr (1885), as well as Villanova University (1842), are located in the suburbs. Philadelphia is also a centre for medical education, with several general schools and specialized schools in pharmacy, podiatry, optometry, and osteopathy. The Curtis Institute of Music (1924) is a leading music school, while the Pennsylvania Academy of Fine Arts (1805) is the oldest art school in the United States.

Institutions of higher learning

CULTURAL LIFE

Much of the essence of Philadelphia lies in the features described previously—its small-town atmosphere, its parks and tree-shaded downtown squares and streets, and its innumerable memorials to the American past, as well as its teeming riverside and factories and its diverse business institutions. There are other factors as well that contribute in their way to an understanding of Philadelphia's culture, considered in its broadest implications to comprise the lifestyles of its people.

In addition to Fairmount Park, Philadelphia has Pennypack Park in the northeast, a semiwilderness setting with bridle paths, bird-watching trails, and an abundance of deer and other wildlife. More than 100 other parks are located throughout the city.

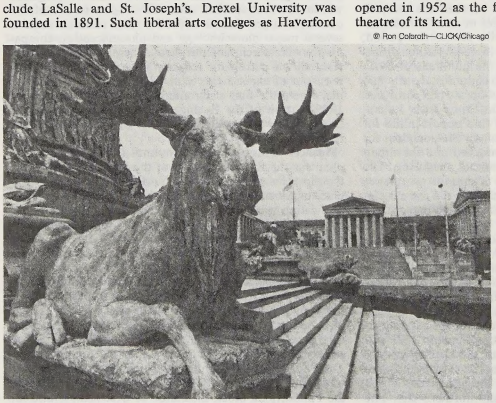
Recreational facilities

Philadelphians have always been a sports-loving group, whether passive or participatory, though professional teams have always been the object, like the city itself, of a good-natured deprecation that is tolerated in residents but not in strangers. City baseball fans enjoyed a long-awaited triumph in 1980 when the Philadelphia Phillies won the World Series. Other professional sports teams include the Eagles (football), 76ers (basketball), and Flyers (hockey). Fox hunting in the surrounding countryside is of old Quaker origins, while the Schuylkill is a major rowing site for collegiate and other individuals and crews, and Germantown harbours remnants of a once-lively citywide enthusiasm for cricket.

In colonial days Philadelphia was known as the "Athens of America," and it retains a high place in the artistic achievement of the nation. The Academy of Music, opened in 1857, is the home of the Philadelphia Orchestra, which is among the finest orchestras of the world and was among the first to broadcast and make recordings. At Fairmount Park is the Mann Music Center, which presents classical and popular music.

Philadelphia was the nation's theatrical centre until well after the Revolution, its stages having hosted the greatest players of Europe and America. The Walnut Street Theatre, opened in 1809, is the oldest playhouse in active use in the English-speaking world. The Playhouse in the Park opened in 1952 as the first city-owned and city-operated theatre of its kind.

© Ron Cobrath—CLICK/Chicago



The Philadelphia Museum of Art.

Museums and galleries

Philadelphia was a pioneer in museums of all kinds. Charles Willson Peale's museum was housed in Independence Hall in the 1800s, but the art museums are now led by the Philadelphia Museum of Art. One of the world's great museums, it houses priceless collections of art from the Western Middle Ages onward and from Asia, as well as an extensive modern collection. Others include the Rodin Museum, featuring the largest collection of sculptures by Auguste Rodin outside of Paris. The Historical Society of Pennsylvania has some 700 portraits by Benjamin West, James Copley, Gilbert Stuart, the four Peales, and other early American painters. The output of visual artists in Philadelphia and vicinity has been by and large conservative, though Thomas Eakins gained fame for work beyond his American contemporaries, and Mary Cassatt was among the few women in the Impressionist school of the late 19th century.

Other museums include the Museum of the Philadelphia Civic Center, originally organized in 1894 as the Commercial Museum to promote trade exhibitions. The Academy of Natural Sciences is the oldest of its kind in the United States. The Franklin Institute Science Museum has extensive exhibits, whereas the Museum of Archaeology and Anthropology is a major feature of the University of Pennsylvania. Among the newer museums in Philadelphia are the Afro-American Historical and Cultural Museum and the Philadelphia Mummers Museum, both established in 1976, and the Independence Seaport Museum at Penn's Landing. Penn's Landing is a waterfront park area where concerts and festivals are held; several historic ships are also located there. Other museums include the Atwater Kent Museum, devoted to city history, and Swedish, Polish, and Jewish ethnic museums. The Barnes Foundation in Merion has an arboretum and a collection of Postimpressionist art. The Philadelphia Zoo, founded in 1859, is the oldest zoo in the country. The city is also the location of Edgar Allen Poe National Historic Site, which preserves the author's 1840s home, and Independence National Historical Park, which in addition to the locations mentioned above includes the former site of Benjamin Franklin's home. Nearby Valley Forge National Historical Park preserves the site where George Washington and the Continental Army spent the winter of 1777–78 during the American Revolution. There also are many small museums housed in restored buildings throughout the city.

History

FOUNDATION AND EARLY SETTLEMENT

Penn's plan

William Penn acquired the province of Pennsylvania in 1681 from King Charles II of England as a place where his fellow Quakers could enjoy freedom of worship and a chance to govern themselves and develop their own way of life. The King made the grant, signed on March 4, 1681, and proclaimed it a few weeks later, on April 2, partly to settle a debt owed to Penn's father, Adm. Sir William Penn, upon his death and also to complete the settlement of the Middle Atlantic region with Englishmen. Penn sent his cousin William Markham to take charge of affairs of government and also to lay out the city Penn named Philadelphia, city of "brotherly love," the name symbolizing his idealistic concepts. From England, Penn wrote in 1681 asking that "the Rivers and Creeks be sounded on my side of the Delaware River...in order to settle a great Towne, and be sure to make your choice where it is most navigable, high, dry, and healthy." He wanted every house to be placed in the middle of its own plot to provide ground about it "that it may be a greene Country Towne, which will never be burnt, and always be wholesome."

Penn arrived in 1682 but had little chance to enjoy his city. He was forced to travel to England in 1684 and was unable to return until 1699. By then, Philadelphia was a flourishing town with many shops and trading houses, as well as several hundred dwellings and about 10,000 people clustered close to the riverfront. Penn's governor declared the city already was the equal of New York "in trade and riches." Penn's policies throughout the colony of religious toleration and the right of the people to take part in the government, in addition to growing prosperity, soon began to attract thousands of English, German, and Scots-Irish settlers, and most came by way of Philadelphia.

Philadelphia by the 1770s had grown to at least 30,000 persons in the central city, and it was the third most important business centre in the British Empire, overshadowed only by Liverpool and London. This position was due in large measure to the city's site at the confluence of the Delaware and Schuylkill rivers, which provided the city access to inland farm regions and to the coal and ore resources that supplied the early iron industry.

A visitor in 1756 wrote that "Everybody in Philadelphia deals more or less in trade," a tribute not merely to Philadelphia's location but also to the shrewd business talents of the Quaker merchants. A large and profitable system of triangular trade involved foodstuffs and wood products, such as lumber and barrel staves, that went from Philadelphia to the West Indies and there were exchanged for sugar, rum, and other West Indian products. These were carried to English ports, where they in turn were exchanged for English manufactures to be brought back to Philadelphia. The prosperous farm country of interior Pennsylvania supplied the Philadelphia merchant with goods for the West Indies, and a profitable coastal trade existed with other colonies and directly with England. By the 1750s Philadelphians had invested heavily in the flourishing charcoal-iron industry. Anthracite coal became an important mineral resource of Pennsylvania 100 years later, and the Philadelphia capital played a leading part in this industry as well as in the mining of bituminous coal farther to the west. Philadelphia continued its leadership in foreign commerce until about 1810, when New York City, with an even more advantageous location, overtook it. Philadelphia surrendered its position as financial capital of the nation in the 1850s.

Philadelphia's leadership in commerce and culture

Shipyards had flourished along the Delaware since colonial days. Most of what came to the city was raw material for manufacture, and Philadelphia became a major centre of the early Industrial Revolution in the United States. In 1785 Oliver Evans invented the first gristmill operated entirely by mechanical power. The city was a pioneer in textile manufacturing and took the raw iron from inland furnaces and made it into tools and implements, such as saws, huge iron castings for cotton-mill machinery, and the first American-built steam locomotives. By 1860 the value of Philadelphia's manufactures ran into several hundred million dollars, about 30 percent of the national total. Textiles, ships, iron products, leather, refined sugar, and boots and shoes were leaders, giving important aid to the Union in the Civil War.

THE GROWTH OF THE CITY

Cultural dominion. Prosperity was translated into personal and community wealth, and with these social and economic advantages Philadelphia assumed early leadership in the arts, in science, and in culture. Benjamin Franklin, who as a young man had migrated from Boston to Philadelphia, became an American leader in scientific and intellectual affairs. Philadelphia had the nation's first free library, its first hospital, and its first learned society, the American Philosophical Society—all founded by Franklin. Along with Franklin, there were men such as Benjamin Rush, the great physician, and David Rittenhouse, an astronomer, mathematician, inventor, and early Philadelphia aristocrat, and many others. The city excelled in printing and publishing: by 1776 there were 23 printers and newspapers with circulations of from 500 to 3,000 copies. Fine private and public buildings were erected. One of them was Andrew Hamilton's Independence Hall, originally—and still—better known by Philadelphians as "the statehouse." Led by Benjamin West, Charles Willson Peale, and Gilbert Stuart, oil painting flourished in the colonial and Federal periods.

The city's strategic location near the midpoint of colonial settlement and its important status as a vital political, economic, and cultural centre—as well as concern that Pennsylvania might not favour the Revolution—brought to Philadelphia the delegates who formed the First Continental Congress in 1774 and a year later the Second Continental Congress, which proclaimed the Declaration of

Independence and governed throughout the Revolution. The Constitutional Convention met in Philadelphia in 1787 and framed the federal Constitution, and the city served as the capital of the United States from 1790 to 1800. The nation celebrated its centennial in Philadelphia in 1876.

Political evolution. The city–county consolidation in 1854 was a result of the inability of a colonial-type government by committees to adapt to the needs of a growing city for new public services—*e.g.*, better streets, police, transportation, sanitation, and schools. Ironically, Philadelphia had, in fact, led most cities in providing urban improvements, but they were too few and too slow, and the breakdown of law and order in the 1840s forced changes. All of Pennsylvania fell into the hands of Philadelphia-based Republican political machines that, like their counterparts (usually Democratic) in other cities, were becoming increasingly sophisticated in methods of manipulating the political processes, especially through the newer immigrant groups, and profiting from the economic life of the state and city. They were instrumental in electing governors and U.S. senators, and, in the Depression years of the 1930s, the hard-hit city refused aid from the Democratic administration in Washington, D.C. Powerful bankers and industrial and business leaders, living for the most part outside the city, favoured this form of government because it kept taxes low, imposed little or no regulation on business, and maintained an aura of social calm through benign neglect or quiet but forceful repression.

Political machines and civic malaise

The 20th century. In about 1900 Philadelphia had been described as "corrupt but content," a status quo that Philadelphians were indeed content with until 1939, when a group known as the Young Turks and influenced by the nationwide New Deal of the Democratic Party began to agitate for charter reform and a city planning commission. Women and blacks were brought into the city's political life for the first time.

The reform movement that began in 1939 was joined in 1948 by top business and financial leaders, who organized the Greater Philadelphia Movement, and together they took up the battle for an entirely new city charter, which went into effect in 1951. This document effectively removed the city council from its administrative role and increased the staff and powers of the mayor. A strong civil-service commission improved professional employment. The council, seven of whose members were elected by all the voters and 10 by districts, was to review and approve taxes and budgets and enact ordinances. Philadelphia thus had the first modern big-city charter in the United States; its approval by a two-to-one popular vote made it an expression of the desire of Philadelphians for better, more efficient, and honest city government.

The first mayors under the new charter were Joseph S. Clark and Richardson Dilworth. Planning was made a virtue, and a $150,000,000 plan was launched at once for improvements in sewerage and sanitation, playgrounds, lighting and streets, police and fire protection, and other basic services long neglected. The Philadelphia Housing Authority and the Philadelphia Redevelopment Authority, with federal and private aid, approached Philadelphia's housing in terms of improving neighbourhoods rather than total clearance followed by new high-rise apartments. Affected areas included the Independence Hall area, neighbouring Society Hill, and the historic waterfront. The city entered a period of physical and political rebuilding, although ethnic conflicts and a steady decline in the central city's population loomed ahead.

During the late 1960s Philadelphia, like other major American cities, was shaken by race riots. This led, in 1971, to a backlash in the election as mayor of Frank Rizzo, a tough former police commissioner oriented toward "law and order." In 1979, however, Philadelphians turned toward more moderate rule by rejecting the attempt of Rizzo to alter the city charter and thereby win permission to seek a third term. In 1983 the city elected its first black mayor, W. Wilson Goode. The city's population began to decrease in the 1970s and continued to do so through the early years of the 21st century, as increasing numbers of residents moved into the surrounding suburbs.

In 1985 Philadelphians were shocked by a deadly siege on the headquarters of MOVE, a radical, anarchist African American group organized in the 1970s that had engaged in frequent conflict with police and neighbours. Police bombed the group's building, killing 11 people inside and burning down some 60 surrounding houses in Philadelphia's most destructive residential fire ever.

Under the administrations of Edward Rendell (1992–2000) and his successor, John F. Street, the city promoted itself as a focus of tourism and culture. The Center City area was revitalized with new hotels, restaurants, and a convention centre. The Avenue of the Arts was created along Broad Street to unite the area's cultural institutions (including the Pennsylvania Academy of Fine Arts and the Academy of Music) as an entertainment destination. Independence National Historical Park was restored and redesigned in the 1990s and the first years of the 21st century.

BIBLIOGRAPHY

General: ROBERT H. WILSON, *Official Handbook for Visitors: Philadelphia*, 300th anniversary ed. (1982), a city guide; *Philadelphia Magazine,* a periodical that often features a sharply critical and satirical approach.

Historical works: JOHN L. COTTER, DANIEL G. ROBERTS, and MICHAEL PARRINGTON, *The Buried Past: An Archaeological History of Philadelphia* (1992); CARL BRIDENBAUGH and JESSICA BRIDENBAUGH, *Rebels and Gentlemen: Philadelphia in the Age of Franklin* (1942, reprinted 1978), a good picture of colonial and Revolutionary times and people; THOMAS M. DOERFLINGER, *A Vigorous Spirit of Enterprise: Merchants and Economic Development in Revolutionary Philadelphia* (1986), examining aspects of political, social, and economic history; BROOKE HINDLE, *The Pursuit of Science in Revolutionary America, 1735–1789* (1956, reprinted 1974), an excellent source for early science in Philadelphia; AMERICAN PHILOSOPHICAL SOCIETY, *Historic Philadelphia: From the Founding Until the Early Nineteenth Century* (1953), a study in depth; HORACE MATHER LIPPINCOTT, *Early Philadelphia: Its People, Life, and Progress* (1917); *Pennsylvania Magazine of History and Biography* (quarterly), the prestigious journal of the Historical Society of Pennsylvania (well indexed); RUSSELL F. WEIGLEY, NICHOLAS B. WAINWRIGHT, and EDWIN WOLF II (eds.), *Philadelphia: A 300 Year History* (1982), a collection of scholarly essays; SAM BASS WARNER, JR., *The Private City: Philadelphia in Three Periods of Its Growth,* 2nd ed. (1987), a provocative interpretation of the impact of urbanization; and WILLIAM W. CUTLER III and HOWARD GILLETTE, JR. (eds.), *The Divided Metropolis: Social and Spatial Dimensions of Philadelphia, 1800–1975* (1980), an insightful anthology of essays. The following provide valuable insights into the texture of working-class life and the nature of interethnic and interracial relations in the city: BILLY G. SMITH, *The "Lower Sort": Philadelphia's Laboring People, 1750–1800* (1990); BRUCE LAURIE, *Working People of Philadelphia, 1800–1850* (1980); CYNTHIA L. SHELTON, *The Mills of Manayunk: Industrialization and Social Conflict in the Philadelphia Region, 1787–1837* (1986); ALLEN F. DAVIS and MARK H. HALLER (eds.), *The Peoples of Philadelphia: A History of Ethnic Groups and Lower-Class Life, 1790–1940* (1973); MICHAEL FELDBERG, *The Philadelphia Riots of 1844: A Study of Ethnic Conflict* (1975); GARY B. NASH, *Forging Freedom: The Formation of Philadelphia's Black Community, 1720–1840* (1988); and ROGER LANE, *Roots of Violence in Black Philadelphia, 1860–1900* (1986).

Specialized studies: NATHANIEL BURT, *The Perennial Philadelphians: The Anatomy of an American Aristocracy* (1963, reprinted 1975), an excellent study of Philadelphia society; E. DIGBY BALTZELL, *Puritan Boston and Quaker Philadelphia: Two Protestant Ethics and the Spirit of Class Authority and Leadership* (1979), an important comparative history; MAXWELL STRUTHERS BURT, *Philadelphia, Holy Experiment* (1945), a good general description with great feeling for the city; GEORGE B. TATUM, *Penn's Great Town: 250 Years of Philadelphia Architecture Illustrated in Prints and Drawings* (1961), a valuable review of the cultural heritage; FREDRIC M. MILLER, MORRIS J. VOGEL, and ALLEN F. DAVIS, *Still Philadelphia: A Photographic History, 1890–1940* (1983); W.E.B. DU BOIS, *The Philadelphia Negro: A Social Study* (1899, reprinted 1973), an excellent monograph on the urban experience of blacks and a landmark in American sociological studies; VINCENT P. FRANKLIN, *The Education of Black Philadelphia* (1979), a solid study of the social and educational history of black Philadelphians during the first half of the 20th century; and CAROLYN ADAMS *et al., Philadelphia: Neighborhoods, Division, and Conflict in a Postindustrial City* (1991).

(S.K.S./J.B.B.T./M.S.Ma./Ed.)

Philippines

The Republic of the Philippines (Pilipino: Republika ñg Pilipinas) is an archipelago consisting of some 7,100 islands and islets lying about 500 miles off the southeast coast of Asia. The total land area of the Philippines is 115,800 square miles (300,000 square kilometres). It is bounded by the Philippine Sea to the east, the Celebes Sea to the south, and the South China Sea to the west and north. The Philippines takes its name from Philip II, who was king of Spain during the Spanish colonization of the islands in the 16th century. Manila is the biggest city and the national capital. It is located on Luzon, the largest island, which has a land area of 40,420 square miles.

The archipelago spreads out in the form of a triangle, with the islands south of Palawan, the Sulu Archipelago, and the island of Mindanao forming (west–east) its southern base and the Batan Islands, in the north, its apex. The islands stretch for about 1,150 miles (1,850 kilometres) from north to south and for about 700 miles from east to west at their widest extent. Only about two-fifths of the islands and islets have names, and only some 350 have areas of one square mile or more. The large islands fall into three groups: (1) the Luzon group in the north and west, consisting of Luzon, Mindoro, and Palawan; (2) the Visayan group in the centre, consisting of Bohol, Cebu, Leyte, Masbate, Negros, Panay, and Samar; and (3) Mindanao in the south.

Because it was under Spanish rule for 333 years and under U.S. tutelage for a further 48 years, the Philippines has many cultural affinities with the West. It is, for example, the fourth most populous country in which English is an official language and the only predominantly Roman Catholic country in Southeast Asia. Its peoples, however, are Asian in consciousness and in aspiration. In many ways Filipino society is composed of paradoxes, perhaps the most apparent being the great extremes of wealth and poverty in the nation. The Philippines is a country of rich resources, but it is in the process of developing its full potentialities. It is primarily agricultural, although a high degree of domestic and foreign investment has spurred the rapid development of its industrial potential. Educationally, it is among the most advanced of Asian countries, having a high literacy rate.

This article is divided into the following sections:

Physical and human geography 537
- The land 537
  - Relief and drainage
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 540
  - Ethnic composition
  - Linguistic groups
  - Religion
  - Demography
- The economy 541
  - Resources
  - Agriculture
  - Industry
  - Finance and trade
  - Administration of the economy
  - Transportation
- Administration and social conditions 542
  - Government
  - Justice
  - Armed forces
  - Education
  - Health and welfare
- Cultural life 544
  - Folk traditions
  - Museums
  - The media

History 544
- Pre-Spanish history 544
- The Spanish period 545
- The 19th century 545
- The Philippine revolution 546
- The period of U.S. influence 546
- World War II 547
- The early republic 548
- The Marcos regime and after 548
  - Martial law
  - The downfall of Marcos and return of democratic government

Bibliography 549

Physical and human geography

THE LAND

Relief and drainage. The irregular configuration of the archipelago, the coastline of more than 21,500 miles, the great extent of mountainous country, the narrow and interrupted coastal plains, the generally northward trend of the river systems, and the lakes are the outstanding physical features of the Philippines. The islands are composed of volcanic, coral, and all principal rock formations. The mountain ranges for the most part run in the same general direction as the islands themselves, approximately north to south. The Cordillera Central, the central mountain chain of Luzon, running north to the Luzon Strait from the northern boundary of the central plain, is the most prominent range. It consists of two and in places three parallel ranges, each of which averages 5,900 feet (1,800 metres) in height. The Sierra Madre, extending along the Pacific coast from northern to central Luzon, is the longest mountain range in the country. This range and the Cordillera Central merge in north-central Luzon to form the Caraballo Mountains. To the north of the latter, and between the two ranges, is the fertile Cagayan Valley. The narrow Ilocos, or Malayan, range, lying close along the west coast of northern Luzon, rises in places to more than 5,000 feet and is seldom less than 3,500 feet in height; it is largely volcanic. In the southwestern part of northern Luzon are the rugged Zambales Mountains, consisting of more or less isolated old volcanic stocks (rock formed under great heat and pressure deep beneath the Earth's surface).

Mountains of Luzon

The central plain of Luzon, about 150 by 50 miles, is only about 100 feet above sea level except near its centre. The greater part of southern Luzon is occupied by isolated volcanoes and irregular masses of hills and mountains. The highest peak is Mayon Volcano (8,077 feet) near Legazpi in Albay province.

Through the island of Palawan, about 25 miles wide and more than 250 miles long, there extends a range with an average height of from 4,000 to 5,000 feet. Each of the Visayan Islands except Samar and Bohol is traversed longitudinally by a single range with occasional spurs. Several peaks in Panay reach an altitude of 6,000 feet or more. Canlaon Volcano, on Negros, is 8,087 feet high, and other peaks reach altitudes of more than 6,000 feet.

Relief of Mindanao

There are several important ranges on Mindanao, the Diuata (Diwata) Mountains along the eastern coast being the most prominent. West of this lies the fertile valley of the Agusan River, 20 to 30 miles wide. Farther west and southwest is the Cotabato Valley of the Mindanao River

(Río Grande de Mindanao), and between its lower course and the southern coast lies a range trending northwest–southeast. On the southeastern border of the basin of the Mindanao River is Mount Apo, at 9,692 feet (2,954 metres) the highest peak in the Philippines. Around Lake Lanao are a number of volcanic peaks, and a low cordillera extends through the Zamboanga Peninsula in the far west.

The plains lying amid the mountains—for example, the central plain of Luzon and the central plain of Panay—have the greatest density of population in the islands, except Cebu, where the people live mostly on the coastal plain because the interior is high and rugged. The most important rivers are the Cagayan, Agno, Pampanga, Pasig, and Bicol on Luzon and the Mindanao and Agusan on Mindanao. The Pasig, which flows through the city of Manila, is commercially important. The largest lake in the archipelago is Laguna de Bay, located south of Manila.

Volcanoes are a conspicuous feature of the landscape, but there is relatively little volcanic activity. There are altogether about 50 volcanoes, of which more than 10 are known to be active. All gradations of volcanoes can be seen, from the almost perfect cone of Mayon, which has been compared to Mount Fuji in Japan, to old, worn-down volcanic stocks, the present forms of which give little indication of their origin. The several distinct volcanic areas are in south-central and southern Luzon and

MAP INDEX

Cities and towns

Abuyog ..... 10 45 N 125 01 E
Angeles ..... 15 09 N 120 35 E
Antipolo ..... 14 35 N 121 10 E
Aparri ..... 18 22 N 121 39 E
Bacolod ..... 10 40 N 122 56 E
Baganga ..... 07 35 N 126 34 E
Bago ..... 10 32 N 122 50 E
Baguio ..... 16 25 N 120 36 E
Bais ..... 09 35 N 123 07 E
Balabac ..... 07 59 N 117 04 E
Baler ..... 15 46 N 121 34 E
Balimbing ..... 05 05 N 119 58 E
Baliuag ..... 14 57 N 120 54 E
Bangued ..... 17 36 N 120 37 E
Bantayan ..... 11 10 N 123 43 E
Batangas ..... 13 45 N 121 03 E
Baybay ..... 10 41 N 124 48 E
Bayombong ..... 16 29 N 121 09 E
Bislig ..... 08 13 N 119 58 E
Bogo ..... 11 03 N 124 00 E
Bonggaw ..... 05 02 N 119 46 E
Bontoc ..... 17 05 N 120 58 E
Borongan ..... 11 37 N 125 26 E
Bulan ..... 12 40 N 123 52 E
Burauen ..... 10 58 N 124 53 E
Butuan ..... 08 54 N 125 35 E
Cabanatuan ..... 15 29 N 120 58 E
Cadiz ..... 10 57 N 123 18 E
Cagayan de Oro ..... 08 29 N 124 39 E
Calapan ..... 13 25 N 121 10 E
Calbayog ..... 12 04 N 124 35 E
Caloocan ..... 14 39 N 120 58 E
Candon ..... 17 12 N 120 27 E
Canlaon ..... 10 22 N 123 12 E
Capiz, see Roxas
Catarman ..... 12 30 N 124 38 E
Catbalogan ..... 11 46 N 124 53 E
Cauayan ..... 16 56 N 121 46 E
Cavite ..... 14 29 N 120 55 E
Cebu ..... 10 18 N 123 54 E
Claveria ..... 12 54 N 123 15 E
Coron ..... 12 00 N 120 12 E
Cotabato ..... 07 13 N 124 15 E
Daet ..... 14 05 N 122 55 E
Dagupan ..... 16 03 N 120 20 E
Danao ..... 10 32 N 124 02 E
Dansalan, see Marawi
Dapitan ..... 08 39 N 123 25 E
Davao ..... 07 04 N 125 36 E
Dimasalang ..... 12 12 N 123 51 E
Dinagat ..... 09 59 N 125 35 E
Dipolog ..... 08 35 N 123 20 E
Dolores ..... 12 02 N 125 29 E
Dumaguete ..... 09 18 N 123 18 E
General Luna ..... 09 47 N 126 09 E
General Santos ..... 06 07 N 125 10 E
Gingoog ..... 08 50 N 125 07 E
Guiuan ..... 11 02 N 125 43 E
Ilagan ..... 17 10 N 121 54 E
Iligan ..... 08 14 N 124 14 E
Iloilo City ..... 10 42 N 122 33 E
Ipil ..... 07 47 N 122 35 E
Iriga ..... 13 25 N 123 25 E
Irosin ..... 12 42 N 124 02 E
Isabela ..... 06 42 N 121 58 E
Isulan ..... 06 37 N 124 40 E
Janiuay ..... 10 58 N 122 30 E
Jaro ..... 11 11 N 124 47 E
Jolo ..... 06 03 N 121 00 E
Kagayan de Sulu, see Mahalu
Kalibo ..... 11 43 N 122 22 E
Koronadal ..... 06 30 N 124 51 E
La Carlota ..... 10 25 N 122 55 E
Lagawe ..... 16 49 N 121 06 E
Laoag ..... 18 12 N 120 36 E
Laoang ..... 12 34 N 125 00 E
Lapu-Lapu (Opon) ..... 10 19 N 123 57 E
Lazi ..... 09 08 N 123 38 E
Legazpi ..... 13 08 N 123 44 E
Lianga ..... 08 38 N 126 06 E
Liloy ..... 08 08 N 122 40 E
Lingayen ..... 16 01 N 120 14 E
Lipa ..... 13 57 N 121 10 E
Los Baños ..... 14 13 N 121 11 E
Lubang ..... 13 52 N 120 07 E
Lucban ..... 14 06 N 121 33 E
Lucena ..... 13 56 N 121 37 E
Lupon ..... 06 54 N 126 00 E
Luuk (Tungkil) ... 06 04 N 121 47 E
Maasin ..... 10 08 N 124 50 E
Magugpo, see Tagum
Mahalu (Kagayan de Sulu) ..... 07 01 N 118 30 E
Malaybalay ..... 08 09 N 125 05 E
Malolos ..... 14 51 N 120 49 E
Mambajao ..... 09 15 N 124 43 E
Mandaue ..... 10 20 N 123 56 E
Mangaldan ..... 16 04 N 120 24 E
Manila ..... 14 35 N 121 00 E
Marawi (Dansalan) ..... 08 01 N 124 18 E
Margosatubig ... 07 34 N 123 10 E
Masbate ..... 12 22 N 123 36 E
Mati ..... 06 57 N 126 13 E
Muddas, see Siasi
Naga (Nueva Caceres) ..... 13 37 N 123 11 E
Narra ..... 09 16 N 118 25 E
Naval ..... 11 34 N 124 23 E
Nueva Caceres, see Naga
Odiongan ..... 12 24 N 121 59 E
Olongapo ..... 14 50 N 120 16 E
Opon, see Lapu-Lapu
Ormoc ..... 11 00 N 124 37 E
Oroquieta ..... 08 29 N 123 48 E
Oton ..... 10 42 N 122 29 E
Ozamiz ..... 08 08 N 123 50 E
Pagadian ..... 07 49 N 123 25 E
Palayan ..... 15 33 N 121 06 E
Pandan ..... 14 03 N 124 10 E
Pasacao ..... 13 31 N 123 03 E
Polillo ..... 14 43 N 121 56 E
Polomolok ..... 06 13 N 125 04 E
Pototan ..... 10 55 N 122 40 E
Puerto Princesa .. 09 46 N 118 44 E
Quezon City ..... 14 38 N 121 00 E
Romblon ..... 12 35 N 122 15 E
Roxas (Capiz) ... 11 35 N 122 45 E
Sagay ..... 10 57 N 123 25 E
San Antonio ..... 14 57 N 120 05 E
San Carlos ..... 10 30 N 123 25 E
San Carlos ..... 15 55 N 120 20 E
San Fernando ... 15 01 N 120 41 E
San Fernando ... 16 37 N 120 19 E
San Jacinto ..... 12 34 N 123 44 E
San Jose ..... 10 45 N 121 56 E
San José ..... 12 21 N 121 04 E
San Jose ..... 15 48 N 121 00 E
San Pablo ..... 14 04 N 121 19 E
Santa Cruz ..... 13 29 N 122 02 E
Siasi (Muddas) .. 05 33 N 120 49 E
Silay ..... 10 48 N 122 58 E
Sitangkai ..... 04 40 N 119 24 E
Socorro ..... 09 37 N 125 58 E
Sorsogon ..... 12 58 N 124 00 E
Surigao ..... 09 45 N 125 30 E
Tabango ..... 11 19 N 124 22 E
Tabuk ..... 17 24 N 121 25 E
Tacloban ..... 11 15 N 125 00 E
Tacurong ..... 06 42 N 124 42 E
Tagaytay ..... 14 07 N 120 58 E
Tagbilaran ..... 09 39 N 123 51 E
Tagig ..... 14 32 N 121 04 E
Tagoloan ..... 08 32 N 124 45 E
Tagum (Magugpo) ... 07 28 N 125 48 E
Talibon ..... 10 07 N 124 18 E
Talisay ..... 10 15 N 123 51 E
Tandag ..... 09 04 N 126 12 E
Tangub ..... 08 03 N 123 44 E
Tanjay ..... 09 31 N 123 09 E
Tarlac ..... 15 29 N 120 35 E
Tiwi ..... 13 27 N 123 41 E
Toledo ..... 10 23 N 123 38 E
Trece Martires ... 14 16 N 120 50 E
Tuguegarao ..... 17 37 N 121 44 E
Tungkil, see Luuk
Valencia ..... 07 57 N 125 03 E
Vigan ..... 17 34 N 120 23 E
Virac ..... 13 35 N 124 15 E
Zamboanga ..... 06 54 N 122 04 E

Physical features and points of interest

Agno, *river* ..... 16 02 N 120 08 E
Agusan, *river* ..... 09 00 N 125 31 E
Apo, Mount, *volcano* ..... 06 59 N 125 16 E
Babuyan Channel ..... 18 44 N 121 40 E
Babuyan Islands .. 19 15 N 121 40 E
Balabac Island .. 07 57 N 117 01 E
Balabac Strait ... 07 40 N 117 00 E
Balintang Channel ..... 19 49 N 121 40 E
Bashi Channel ... 21 25 N 121 30 E
Basilan Island ... 06 34 N 122 03 E
Basilan National Park ..... 06 36 N 122 01 E
Batangas Bay ... 13 43 N 121 00 E
Batan Islands ... 20 30 N 121 50 E
Bay, Laguna de, *lake* ..... 14 23 N 121 15 E
Bicol, *river* ..... 13 44 N 123 07 E
Bohol, *island* ..... 09 50 N 124 10 E
Bohol (Mindanao) Sea ..... 09 10 N 124 25 E
Bojeador, Cape .. 18 30 N 120 34 E
Bugsuk Island ... 08 15 N 117 18 E
Busuanga Island .. 12 05 N 120 05 E
Cagayan, *river* ... 18 22 N 121 37 E
Cagayan Islands .. 09 40 N 121 16 E
Cagayan Sulu Island ..... 07 01 N 118 30 E
Cagayan Valley .. 17 30 N 121 45 E
Calamian Group, *islands* ..... 12 00 N 120 00 E
Camiguin Island .. 09 11 N 124 42 E
Canlaon Volcano .. 10 25 N 123 08 E
Caraballo Mountain ..... 16 08 N 121 05 E
Catanduanes Island ..... 13 45 N 124 15 E
Cebu, *island* ..... 10 20 N 123 45 E
Celebes Sea .... 05 00 N 122 00 E
Cordillera Central .. 17 20 N 120 57 E
Corregidor Island .. 14 23 N 120 35 E
Cotabato Valley .. 07 00 N 124 30 E
Culion Island ... 11 50 N 119 55 E
Davao Gulf ..... 06 40 N 125 55 E
Dinagat Island ... 10 12 N 125 35 E
Dingalan Bay ... 15 18 N 121 25 E
Diuata Mountains .. 09 00 N 125 50 E
Dumaran Island .. 10 33 N 119 51 E
Escarpada Point .. 18 31 N 122 13 E
Homonhon Island .. 10 44 N 125 43 E
Illana Bay ..... 07 25 N 123 45 E
Jolo Island ..... 05 58 N 121 06 E
Lake Naujan National Park ... 13 20 N 121 18 E
Lamon Bay ..... 14 25 N 122 00 E
Lanao, Lake, see Sultan Alonto, Lake
Leyte, *island* ..... 10 50 N 124 50 E
Leyte Gulf ..... 10 50 N 125 25 E
Limasawa Island .. 09 56 N 125 05 E
Lingayen Gulf ... 16 15 N 120 14 E
Lubang Island ... 13 46 N 120 11 E
Luzon, *island* ... 16 00 N 121 00 E
Luzon Strait ... 21 00 N 122 00 E
Mactan Island ... 10 18 N 123 58 E
Mainit, Lake .... 09 26 N 125 32 E
Mainit Hot Spring National Park ... 07 52 N 125 56 E
Manila Bay ..... 14 30 N 120 45 E
Marinduque, *island* ..... 13 24 N 121 58 E
Masbate, *island* .. 12 15 N 123 30 E
Mayon Volcano .. 13 15 N 123 41 E
Mindanao, *island* .. 08 00 N 125 00 E
Mindanao, *river* .. 07 07 N 124 24 E
Mindanao Sea, see Bohol Sea
Mindoro, *island* ... 12 50 N 121 05 E
Mindoro Strait ... 12 20 N 120 40 E
Moro Gulf ..... 06 51 N 123 00 E
Mount Isarog National Park ... 13 39 N 123 23 E
Naujan, Lake .... 13 10 N 121 21 E
Negros, *island* ... 10 00 N 123 00 E
Palawan, *island* .. 09 30 N 118 30 E
Palawan Passage ..... 10 00 N 118 00 E
Pampanga, *river* .. 14 47 N 120 39 E
Panay, *island* .... 11 15 N 122 30 E
Panay Gulf ..... 10 15 N 122 15 E
Philippine Sea .... 15 00 N 126 00 E
Polillo Island .... 14 50 N 121 57 E
Pulog, Mount .... 16 36 N 120 54 E
Romblon Island .. 12 33 N 122 17 E
Samar, *island* .... 12 00 N 125 00 E
San Agustin, Cape ..... 06 16 N 126 11 E
San Bernardino Strait ..... 12 32 N 124 10 E
San Ildefonso Peninsula ..... 16 10 N 122 05 E
Sarangani Islands ..... 05 25 N 125 26 E
Siargao Island ... 09 53 N 126 02 E
Sibuyan Island ... 12 25 N 122 34 E
Sibuyan Sea .... 12 50 N 122 40 E
Sicapoo, Mount .. 18 01 N 120 56 E
Sierra Madre, *mountains* ..... 16 20 N 122 00 E
Siquijor Island ... 09 11 N 123 34 E
South China Sea .. 15 00 N 117 00 E
Sultan Alonto (Lanao), Lake ... 07 53 N 124 15 E
Sulu Archipelago .. 06 00 N 121 00 E
Sulu Sea ..... 09 00 N 120 00 E
Taal, Lake ..... 13 55 N 121 00 E
Tablas Island .... 12 24 N 122 02 E
Tawitawi Island .. 05 10 N 120 00 E
Tinaca Point .... 05 33 N 125 20 E
Visayan Islands (Visayas) ..... 11 00 N 123 30 E
Visayan Sea .... 11 35 N 123 51 E
Y'Ami Island .... 21 07 N 121 57 E
Zambales Mountains ..... 15 45 N 120 05 E
Zamboanga Peninsula ..... 07 32 N 122 16 E


Scale 1:8,130,000
1 inch equals approx. 128 miles
0 25 50 75 100 125 mi
0 50 100 150 200 km
Cities over 600,000
Cities 200,000 to 600,000
Cities under 200,000
National capitals
Canals
Dams
National parks
Spot elevations in metres (1 m = 3.28 ft)
Lambert Conformal Conic Projection
LUZON
MINDANO
PALAWAN
MINDORO
SAMAR
LEYTE
CEBU
NEGROS
PANAY
BABUYAN ISLANDS
BATAN ISLANDS
POLILLO ISLANDS
CATANDUANES ISLAND
MASBATE
ROMBLON ISLAND
TABLAS ISLAND
SIBUYAN ISLAND
CALAMIAN GROUP
CUYO ISLAND
BUSUK ISLAND
BALABAC ISLAND
CAGAYAN SULU ISLAND
JOLO ISLAND
TAWITAWI ISLAND
SULU ARCHIPELAGO
CAMIGUIN ISLAND
SIARGAO ISLAND
DINAGAT ISLAND
HOMONHON ISLAND
SARANGANI ISLANDS
DUMARAN ISLAND
CAGAYAN ISLANDS
LIMASAWA ISLAND
ZAMBOANGA PENSULA
SOUTH CHINA SEA
PHILIPPINE SEA
SULU SEA
CELEBES SEA
MALAYSIA
Luzon Strait
Balintang Channel
Bashi Channel
Babuyan Channel
Balabac Strait
Surigao Strait
Moro Gulf
Davao Gulf
Manila Bay
Lingayen Gulf
Lagonoy Gulf
Ragay Gulf
Leyte Gulf
Panay Gulf
Illana Bay
Dapitan Bay
Dingalan Bay
Lamon Bay
Tayabas Bay
Ormoc Bay
Carigara Bay
Sibuyan Sea
Visayan Sea
Mindoro Strait
Verde Island Passage
Sibutu Passage
Basilan Strait
Sulu Sea
Cape Bojeador
Escarpada Point
Aparri
Mount Alzapan 2,801
Laoag
Banguod
Vigan
Tabuk
Tuguegarao
Candon
Ilagan
Bontoc
Lagawe
Cauayan
San Fernando
Baguio
Mount Pulog 2,934
Bayombong
Dagupan
Mangaldan
Urdaneta
CARABALLO MTS
San Jose
San Ildefonso Peninsula
San Carlos
Baler
Tarlac
Palauan
Cabanatuan
Angeles
San Fernando
Olongapo
Balanga
Malolos
Caloocan
Quezon City
Antipolo
Manila
Pasig
Taguig
Cavite
Los Baños
Trece Martires
Tagaytay
Lucban
Batangas
Calapan
Santa Cruz
Mount Malindang 2,402
Lucena
Daet
Virac
Pandan
Siruma
Iriga
Tiwi
Masbate
Legazpi
Sorsogon
Claveria
Pilar
Surigao
Catarman
Borongan
Romblon
Odiongan
San Jose
Tablas
Calbayog
Catbalogan
Dimasalang
Roxas
Kalibo
Antique
Tacloban
Ormoc
Bantayan
Iloilo City
Passi
Dumaguete
Bais
Tulay
Bacolod
Silay
Escalante
Cadiz
Sagay
Roxas City
Guiuan
Borongan
Abuyog
Baybay
Maasin
Sogod
Toledo
Danao
Mandaue
Cebu City
Talisay
Lapu-Lapu
Consolacion
Cabadbaran
Siargao
Surigao
General Luna
Socorro
Dapitan
Dipolog
Oroquieta
Iligan
Ozamiz
Tangub
Ipil
Pagadian
Malaybalay
Valencia
Bukidnon
Cagayan de Oro
Iligan
Marawi
Lanao
Tagoloan
Gingoog
Butuan
Tandag
Bislig
Mount Hibut Apo Range Natural Park
Bagenga
Tagum
Davao
Digos
Mount Apo 2,954
Kidapawan
Cotabato
Tacurong
Isulan
Koronadal
Polomolok
General Santos
Cape San Agustin
Tinaca Point
Zamboanga
Isabela
Basilan National Park
Basilan Island
Jolo
Luuk
Siasi
Bongao
Mapun
Sitangkai
Puerto Princesa
Narra
Busuk Island
Balabac
Balabac Island
Cagayan Sulu Island
Mahaut
Tawitawi Island
Jolo Island
Malaysia
Sulu Archipelago
Celebes Sea
© Encyclopædia Britannica Inc.

on the islands of Negros, Mindanao, Jolo, and elsewhere. Tremors and earthquakes are common.

Climate. The climate of the Philippines is tropical and is strongly affected by monsoon (rain-bearing) winds, which blow from the southwest from approximately May to October and from the northeast from November to February. Thus, temperatures remain relatively constant from north to south during the year, and seasons consist of periods of wet and dry. Throughout the country, however, there are considerable variations in the frequency and amount of precipitation. The western shores facing the South China Sea have marked dry and wet seasons. The dry season generally begins in December and ends in May, the first three months being cool and the second three hot; the rest of the year consists of the wet season. The dry season shortens progressively to the east until it ceases to occur. During the wet season, rainfall is heavy in all parts of the archipelago except for an area extending southward through the centre of the Visayan group to central Mindanao and then southwestward through the Sulu Archipelago; rain is heaviest along the eastern shores facing the Pacific Ocean.

The two seasons

Typhoons

From June to December typhoons often strike the archipelago. Most of these storms come from the southeast, their frequency generally increasing from south to north; in some years the number of typhoons reaches 25 or more. Typhoons are heaviest in Samar, Leyte, eastern Quezon province, and the Batan Islands, and when accompanied by floods or high winds they may cause great loss of life and property. Mindanao is generally free from typhoons.

November through February constitutes the most agreeable season; the air is cool and invigorating at night, and the days are pleasant and sunny. During the hot part of the dry season in most places, and especially in the cities of Cebu, Davao, and Manila, the temperature sometimes rises as high as 100° F (38° C). Overall temperatures decline with altitude, however, and cities and towns located at higher elevations—such as Baguio in northern Luzon, Majayjay and Lucban south of Manila, and Malaybalay in central Mindanao—experience a pleasant climate throughout the year; at times the temperature in these places dips as low as 43° F (6° C).

Plant and animal life. Much of the land area of the islands is covered with forest, and many of the mountains are densely forested. In northern Luzon the principal mountain tree is pine. In many places where the forests have been burned away their place has been taken by tall, coarse cogon grass (genus *Imperata*) or similar grasses. There are about 10,000 known species of flowering plants and ferns. Much of the vegetation resembles that found in Malaysia, and the vegetation of the coastal areas, including the mangrove swamps, is practically identical with that of similar areas throughout the Malay Archipelago. Himalayan elements occur in the mountains of northern Luzon, while a few Australian types are found at various altitudes. A majority of the plant species are indigenous, including the dominant forest species. Lauan (Philippine mahogany) often composes a high percentage of the forest. There are many other valuable hardwoods as well as many softwoods. About 800 species of orchids, some very rare, are found.

There are about 220 species of mammals, among which are carabaos (water buffalo), goats, horses, hogs, cats, dogs, monkeys, squirrels, lemurs, mice, pangolins (scaly anteaters), chevrotains (mouse deer), mongooses, civet cats, and red and brown deer. The tamarau, a very small species of carabao, is found only on Mindoro. There are at least 56 species of bats, many peculiar to the Philippines. Fossil remains show that there were once elephants on the islands.

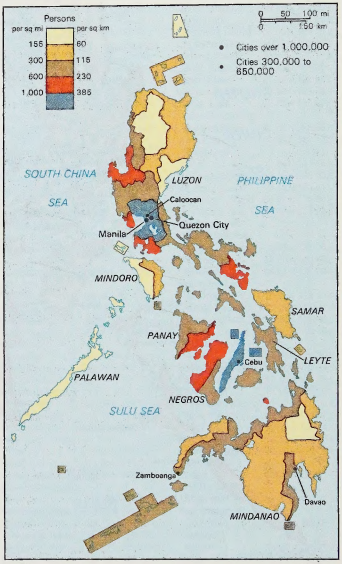
There are some 500 species of birds, many of which are migrants. Birdlife includes jungle fowl, pigeons, peacocks, pheasants, doves, parrots, hornbills, kingfishers, sunbirds, tailorbirds, weaverbirds, herons, and quails. Many species of birds are unique to the island of Palawan. The rare and endangered monkey-eating eagle is restricted mainly to isolated areas on Mindanao and in the Sierra Madre on Luzon.

Settlement patterns. Before the arrival of the Spanish in the 16th century, the people lived in small independent villages called barangays, each ruled by a petty king called a *datu*. Rural settlement today is centred upon poblaciones, or small towns, originally founded by the Spanish. Roads or trails were built from the poblacion in four to six directions, like the spokes of a wheel, and along the roadsides barrios, or villages, grew up, subdivided into smaller units called sitios. More people now live in the barrios than in the poblaciones. A large barrio may be composed of as many as 1,000 households and have a population of 4,000 or more. The cultivation of rice or corn (maize) and fishing provide the basic subsistence.

The barangay

Manila, Cebu, Jaro, Vigan, and Nueva Caceres (now called Naga) were granted city charters by the Spanish. More chartered cities were founded under U.S. administration and since independence in 1946. Quezon City was

Chartered cities



Population density of the Philippines.

chartered in 1939 and became the capital of the Philippines in 1948. In 1975 Manila, Pasay City, Caloocan, and Quezon City and 13 adjoining municipalities were placed under one administrative unit known as Metropolitan, or Metro, Manila. The city proper of Manila was again made the national capital in 1976; Metropolitan Manila is designated the National Capital Region. In addition to the cities there are a great many municipalities and a number of municipal districts.

THE PEOPLE

Ethnic composition. The people of the Philippines are called Filipinos. Their ancestors, who were of Malay stock, came from the southeastern Asian mainland as well as from what is now Indonesia. From the 10th century, contacts with China resulted in a group of mixed Filipino-Chinese descent, who account for a minority of the population. A small percentage of Chinese nationals also live in the country. Spanish-Filipinos and Filipino-Americans may be distinguished by their fairer complexion,

The Filipinos

taller stature, and aquiline nose structure. The relatively small numbers of emigrants from the Indian subcontinent added to the population's racial mixture. There are small numbers of resident U.S. nationals (excluding military personnel) and Spaniards. The aboriginal inhabitants of the islands were the Negritos, or Pygmies, also called Aetas or Balugas; they now constitute only a small percentage of the total population.

Linguistic groups. Estimates of the total number of native languages and dialects spoken in the Philippines differ, but scholarly studies suggest that there are some 70 of them. These languages are all closely related, belonging to the Austronesian (Malayo-Polynesian) family of languages. Traditionally, eight major linguistic groups are identified. These are (1) the Tagalog group, concentrated in Manila, central and south central Luzon, and the islands of Mindoro and Marinduque; (2) Cebuano, in Cebu, Bohol, eastern Negros, western Leyte, and parts of Mindanao; (3) Ilocano, dominant in many parts of northern Luzon; (4) Hiligaynon (Ilongo), spoken in parts of Panay and western Negros; (5) Bicol, spoken in southern Luzon and on the island of Catanduanes; (6) Waray-Waray, spoken in Samar and Leyte; (7) Pampango, spoken in parts of central Luzon; and (8) Pangasinan, spoken in central Luzon. Other notable languages are Magindanao and Maranao, spoken in parts of Mindanao, and Aklanon, spoken in Panay.

The Pilipino language

The national, and most widely spoken, language of the Philippines is Pilipino (also called Filipino), based on Tagalog. Pilipino and English are the two official languages and mediums of instruction. Tagalog is the richest of all Philippine languages and has the most extensive written literature.

Religion. The great majority of Filipinos are Roman Catholic; adherents of the Philippine Independent Church (the Aglipayans), Muslims, and Protestants are the largest religious minorities, including a Protestant sect called Iglesia ni Kristo. There are also some Buddhists and animists. Roman Catholicism has been strengthened by an increase in the number of Filipinos in the church hierarchy, the building of seminaries, and the increased involvement of the church in the political and social life of the country.

Demography. The population density of the Philippines is high—some 500 people per square mile (193 per square kilometre) on the average—although the distribution of the population is uneven. Population density in urban areas is extremely high—that of the National Capital Region is about 30,000 per square mile—while density in such outlying areas as the Batan Islands and northeastern Mindanao is considerably lower than the national average. Because it is largely a Roman Catholic country and because most of its population is rural, the country has a high birth rate. Efforts since the mid-20th century at reducing the overall growth rate have had some success, but reductions in the birth rate have been partially offset by reductions in the death rate.

Demographic trends

Especially since World War II, population has tended to move from rural areas to towns and cities. At the beginning of the 20th century more than 80 percent of the population was rural, but by the late 1980s the number had dropped to less than 60 percent. There is also a considerable amount of Filipino emigration, particularly of manual labourers and professionals. Many emigrants have gone to the United States, Okinawa, Guam, and Canada; in addition, a large number of skilled and semiskilled workers have taken temporary overseas assignments, mainly in the Middle East.

THE ECONOMY

Resources. The Philippines is rich in mineral resources. There are major deposits of gold in northern and southern Luzon; iron ore in northern Mindanao and on nearby islands and in central Luzon; copper in west-central Luzon; lead and zinc in western Mindanao; and high-grade chromium ore (chromite) in west-central and southern Luzon, northern Mindanao, and central Palawan. Deposits of silver, nickel, mercury, molybdenum, cadmium, and manganese occur in several other places. Nonmetallic minerals include limestone for cement, found on Cebu, Luzon, and Romblon; salt and asbestos on Luzon; marble on Romblon and Panay; asphalt on Leyte; mineral waters on Luzon; gypsum on Luzon; sulfur on Luzon, Leyte, and Mindanao; guano and phosphate rock on Cebu and Bohol; coal and silica on Cebu and Palawan; and petroleum off the northwest shore of Palawan.

At one time about half of the Philippines' total land area was covered with forests. Of this, a large part abounded with trees of commercial value, especially lauan, narra (species of *Pterocarpus,* used in cabinetmaking), ipil (*Intsia bijuga*), molave (*Vitex littoralis*), and *kamagong* (*Diospyros discolor*). Heavy logging and inadequate reforestation measures, however, have reduced considerably the amount of forested land.

No fewer than 2,000 varieties of fish are in the seas surrounding the islands and in lakes, rivers, estuaries, and fish ponds. The most important commercial fishes are milkfish (a herringlike fish), anchovy, herring, sardine, mackerel, grouper, snapper, cavalla, mullet, barracuda, mudfish, and caesio. Fish are raised in ponds in some provinces of Luzon and Panay. There are pearl beds in the Sulu Archipelago to the south, and mother-of-pearl used for making buttons is exported to China, Hong Kong, and elsewhere. Tortoiseshells and shells used for windows are exported to Singapore, Malaysia, and other countries.

Hydroelectric projects

Hydroelectric power in the past has supplied only a small proportion of the country's electrical output, with thermal plants, most of which burn imported oil, supplying the major proportion. But the completion of several dam projects on Luzon and the expansion of another project on Mindanao have increased the percentage of power generated by hydroelectric installations; irrigation and flood control are additional benefits of some of the projects. Dependence on foreign oil has also been reduced by the construction of geothermal and conventional coal-fired thermal plants.

Agriculture. The Philippines is chiefly an agricultural country. Its soil is rich and fertile, and crops grow abundantly throughout the year. The principal farm products are rice, corn (maize), coconut, sugarcane, abaca (Manila hemp), tobacco, maguey (used for making such products as rope), and pineapple. Many tropical fruits are also raised, the most important being banana, mango, lanseh, or lanzon (*Lansium domesticum*), citrus, and papaya. A wide variety of vegetables are raised for domestic consumption.

Rice production

Rice, the principal staple crop, is grown everywhere, but especially in central and north-central Luzon, south-central Mindanao; western Negros and eastern and cen-

Shostal Associates



Mayon Volcano overlooking surrounding rice fields in Albay province, southern Luzon.

tral Panay. Much of the total farmland is used for rice growing. Since the early 1970s, rice production in the Philippines has improved considerably, and in some years the country has had surplus rice available for export. Factors contributing to this increase in output include the development and use of higher-yielding strains of rice, the construction of feeder roads and irrigation canals, and the use of chemical fertilizers and insecticides. Use of scientific farming techniques in the Philippines, however, has had its drawbacks. The newer strains of rice have required the application of chemicals that are expensive and that generally must be imported.

Coconut production

The Philippines is one of the world's largest producers of coconuts and a major exporter of coconut products. The area devoted to coconut production is second only to that used for rice and corn. The Philippines also produces more than one-third of the world's copra.

Sugar is one of the country's top exports and earns a substantial amount of foreign exchange. Sugarcane is grown extensively in central and north-central Luzon, western Negros, and on Panay. Much of the sugar crop is exported to the United States. Abaca, or Manila hemp, the source of a useful plant fibre, is also an important export. It is grown extensively in eastern Mindanao, southeastern Luzon, and on Leyte and Samar.

High-quality timber and veneer products traditionally have been leading exports, although their importance has declined. Other forest products include dyewoods, rattan, tanbarks, gutta-percha, beeswax, and rubber. Fishing is also one of the most important of Filipino industries. Canned tuna is the principal fish exported, and fish provides an important proportion of the protein in the Filipino diet. Commercial fishing is carried on primarily off Palawan, Negros, Mindanao, and Panay.

Industry. The principal minerals mined include gold, silver, iron, copper, manganese, zinc, lead, coal, and chromite; some petroleum is also extracted. Many of these minerals—notably copper—are produced for export and thus are vulnerable to fluctuations in world demand. Much industrial growth has taken place in the Philippines, particularly in the 1950s and the 1970s, but the manufacturing sector is still in an early stage of development. Many factories are licensees of foreign companies or act as subcontractors for foreign firms, turning out finished products for export from imported semifinished goods. Major manufactured goods include processed foods and beverages, petroleum products, textiles and wearing apparel, and chemicals.

Finance and trade. In addition to the Central Bank of the Philippines, there are commercial banks, thrift banks, and rural banks. Banking operations are also conducted by such other institutions as the Philippine National Bank and the Development Bank, of the Philippines. The principal exports are electronic equipment, coconuts and coconut products, minerals (copper, gold, iron ore), sugar, and garments. The principal imports are fuel, machinery, transport equipment, and electrical equipment. Major trading partners are the United States, Japan, The Netherlands, Germany, and Hong Kong.

Exports and imports

Administration of the economy. The economy of the Philippines is based on free enterprise; individuals and nongovernmental entities are free to participate in its development and management, sometimes with the aid of government credit.

The public sector. The government has assisted the private sector of the economy by exempting certain new industries from taxation for a certain period. Only nominal taxes are imposed on selected industries, and loans on favourable terms are available to others. Government banks, such as the Central Bank of the Philippines, the Philippine National Bank, and the Development Bank, were established to encourage business, agriculture, and industry. Much of the original capital investment of many private rural banks was provided by the government. Private development banks have likewise received government assistance. The National Economic and Development Authority (NEDA) and the Board of Investments were created to help both public and private sectors in planning further economic development.

The government plans national economic development through the NEDA, the National Planning Commission, and other agencies. In so doing, it has sought to increase economic independence; while there have been advantages in the preferential trade relations between the United States and the Philippines, this trade has resulted in the development of industries that would primarily benefit the American rather than the Filipino economy. Although the United States and Japan have continued to be the Philippines' top trading partners, a number of new markets have been explored. The government has also been engaged in business in its own right. It owns such enterprises as the National Development Corporation, the Philippine Ports Authority, and the Philippine National Railways.

Government businesses

Labour relations. The trade-union movement is well-established in Manila and in most other towns and cities. Farmers and tenants are also organized, as are teachers and government employees. The right of all workers to organize unions has been recognized in the constitution promulgated in 1987. Management, for its part, has organized company unions. Relations between trade unions and the employers' union have been generally untroubled. The Bureau of Labor Relations has been effective in settling disputes between labour and management.

Taxation. The government derives its revenue from three major sources: taxation, earnings and other credits, and extraordinary income, including the transfer from special funds (that is to say, funds derived primarily from unexpended balances in the budget that are deposited as savings accounts). Revenue is collected principally through the Bureau of Internal Revenue and the Bureau of Customs. Excise taxes constitute the single largest portion of government income; import duties, corporate and individual income taxes, and sales taxes provide a substantial portion of the remainder. Local governments have only limited taxing powers.

Transportation. Thousands of miles of roads—gravel, asphalt, and concrete—link the towns on the archipelago's many islands, with expressways running north and south from Manila. Elsewhere on Luzon, most of the road from Manila to Laoag, in the far north of the island, and the greater part of the road to Sorsogon, in the extreme south, are hard-surfaced, as are roads branching out from Manila to the city of Baguio and to the heavily populated provinces to the north and south. Thousands of miles of roads of various types have also been constructed on Mindanao, Mindoro, and Palawan and in the Visayas. A major achievement in road construction in the country is the Pan-Philippine Highway (also called the Maharlika Highway), a system of hard-surface roads, bridges, and ferries that connects the islands of Luzon, Samar, Leyte, and Mindanao.

The Pan-Philippine Highway

Two principal railroads operate on Luzon and Panay. The Luzon main line extends from San Fernando in the north to the city of Legazpi in the south. The railroad in Panay connects Iloilo City with Roxas.

The most important port is Manila. Manila North Harbor handles domestic trade, while Manila South Harbor handles shipping from abroad. Other international ports are Cebu (next in importance to Manila), Iloilo City, Davao, Zamboanga, Cagayan de Oro, and Legazpi.

The international airport at Manila, like those at Hong Kong and Singapore, is a focal point for air routes. A separate terminal is used by domestic airlines. The Mactan airport, on Mactan Island near Cebu, also handles international flights. There are a number of other airports for domestic flights, all of which have service to and from Manila.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. *Constitutional framework.* The Philippines has been governed under three constitutions, the first of which was promulgated in 1935, during the period of U.S. administration. It was closely modeled on the U.S. Constitution and included provisions for a bicameral legislative branch, an executive branch headed by a president, and an independent judiciary. During the period of martial law (1972–81) under President Ferdinand E. Marcos, the old constitution was abolished and

replaced by a new constitution (adopted in January 1973) that changed the Philippine government from a U.S.-style presidential system to a parliamentary form; the president became head of state, and executive power was vested in a prime minister and Cabinet. President Marcos, however, also served (until 1981) as prime minister and ruled by decree. Subsequent amendments and modifications of this constitution replaced the former bicameral legislature with a unicameral body and gave the president even more powers, including the ability to dissolve the legislature and (from 1981) to appoint a prime minister from among members of the legislature.

Constitution of 1987

After the downfall of Marcos in 1986, a new constitution, similar to the 1935 document, was drafted and was ratified in a popular referendum held in February 1987. Its key provision was a return to a bicameral legislature, called the Congress of the Philippines, consisting of a 250-member House of Representatives and a 24-member Senate. House members are elected from districts, although a number of them are appointed; they can serve no more than three consecutive three-year terms. Senators, elected at large, can serve a maximum of two six-year terms. The first legislative election under the new constitution was held in May 1987. The president, the head of state, can be elected to only a single six-year term, and the vice president to two consecutive six-year terms. The president appoints the Cabinet, which consists of the heads of the various ministries responsible for running the day-to-day business of the government. Most presidential appointments are subject to the approval of a Commission of Appointments, which consists of equal numbers of senators and representatives.

Local government. The country is divided administratively into 73 provinces, which are grouped into 12 regions; the National Capital Region has special status. Each province is headed by an elected governor. Local political subdivisions that also have elected officials include cities and municipalities; during the Marcos regime the ancient barangay was reinstated as the smallest unit of government.

Political parties. Partisan political activity was vigorous before 1972, the principal rivals being the Nacionalista and Liberal parties. Martial law restrictions all but eliminated partisan politics. Marcos' organization, the New Society Movement (Kilusan Bagong Lipunan; KBL), was created from elements of the Nacionalista Party and other supporters and became predominant. Organized political opposition was revived for legislative elections held in 1978; and, since the downfall of Marcos, partisan politics has returned to its pre-1972 level, with a large number of political parties emerging. Certain outlawed political organizations also operate within the country, the two principal ones being the Moro National Liberation Front (MNLF), a group seeking autonomy for Muslim Filipinos (Moros), and the National Democratic Front (NDF), a Communist-led insurgency movement.

Justice. Justices and judges enjoy fixed tenure and reasonably high compensation. The judiciary has consequently been less criticized than other branches of the government; and since 1900 the Supreme Court has retained the trust of the people, although trust in the entire judiciary diminished considerably during the Marcos regime. Judicial independence was reaffirmed by the constitution of 1987.

The extrajudicial *lupon*

In order to reduce the load of the lower courts, local committees of citizens called Pacification Committees (Lupon Tagapayapa) have been organized to effect extrajudicial settlement of minor cases between barangay residents. In each *lupon* there is a Conciliation Body (Pangkat Tagapagkasundo), the main function of which is to bring opposing parties together and effect amicable settlement of differences. The committee cannot impose punishment, but otherwise its decision is binding.

Armed forces. The Department of National Defense is divided into four services: the army, the navy, the air force, and the constabulary (the national police force), each with a commanding officer whose rank is at least that of brigadier general or, in the navy, commodore. Several army battalions served in the Korean and Vietnam wars, though in the latter the men performed civic instead of combat duties. The armed forces are responsible for containing the antigovernment military actions of the NDF and the MNLF, for controlling smuggling, and for maintaining order. The armed forces also engage in nonmilitary activities, such as providing disaster relief, constructing civilian housing, and participating in literacy campaigns. The commander in chief of the armed forces (the president of the Philippines) is a civilian.

U.S. military bases

Under a series of agreements reached in 1947, the United States maintains several bases in the Philippines and provides the Philippines with military equipment and training. Since the treaty dealing with the bases was revised in 1978, Philippine sovereignty over them has been recognized; all U.S. bases in the country have been placed under the overall command of a Filipino, and the Philippine flag flies over them. Two of the largest U.S. military bases—those at Clark Field and Subic Bay—were closed in the early 1990s.

The constabulary is divided into commands in the 12 administrative regions and the National Capital Region. Each province, including the cities within its territorial jurisdiction, is assigned a constabulary unit under a provincial commander. The police throughout the country are integrated under the supervision of the National Police Commission. There are also more than 500 private armies organized by landowners and local politicians. President Ramos sought to disband the private military forces in 1993.

Education. The largest governmental ministry administers education as well as such cultural institutions as the Institute of National Language (Surian ng Wilkang Pambansa), the National Library, the National Museum of the Philippines, and the National Historical Institute.

Elementary education lasts for six years, secondary education for four, and college courses for four to nine years. Vocational schools offer specialized training for one to two years. Relatively few students finish high school. In view of this, greater emphasis has been placed on nonformal educational opportunities for adults and out-of-school youths. A Bureau of Continuing Education supervises programs in adult literacy, vocational and technical training, health, and family planning.

There are some four dozen state-run universities and colleges, about one-third of them in Metropolitan Manila, as well as a number of private institutions. The University of Santo Tomas was founded as early as 1611. There are also many technical institutions and community colleges in the provinces.

Pilipino is the medium of instruction in all elementary school subjects except science, mathematics, and the English language, which are taught in English. A chronic shortage of supplies and facilities has been partially remedied by a textbook program begun in the mid-1970s and the large-scale manufacture of prefabricated classrooms.

Health and welfare. Health and welfare are the responsibility of the departments of health and social services. The national government maintains a general hospital, a mental hospital, an orthopedic hospital, and a hospital for war veterans and their families. There are also government and private hospitals, as well as rural health units. Incorporated into the Department of Social Services is the Philippine National Red Cross, and a number of private social welfare agencies cooperate with the department.

The rate of mortality has been considerably reduced, particularly among infants and mothers, and there has been a steady increase in average life expectancy. The improvement in health is credited to better prenatal care and the services of more trained midwives, doctors, and nurses, improved housing and sanitation, the provision of health services to government employees and Social Security members, the increasing number of medical and nursing school graduates, and the requirement that a medical graduate render rural service. Nonetheless, the demand for health care continues to outstrip available resources; a large number of trained medical professionals emigrate, particularly to the United States, and many of the poorest people still rely on the services of practitioners of traditional medicine and unlicensed midwives.

Housing

There is a serious housing shortage everywhere, although it is especially acute in Manila. In many places people live in their own dwellings, but the houses are often substandard and lack elementary facilities for health and sanitation. To help meet this problem the government has relocated thousands of squatters from Manila to resettlement areas in nearby provinces. The Marcos regime instituted housing projects in which model communities were built that provided the residents with hygienic dwellings, a number of amenities, and facilities for raising livestock and for pursuing cottage industries and other means of making a living. This program met with some success.

CULTURAL LIFE

Philippine society is an incongruous blend of diversity and homogeneity. Geographically the country is part of the East, but in culture it is strongly Western. Forces of assimilation have constantly worked to overcome differences caused by the relative physical isolation of various groups of people throughout the archipelago who had come over time from disparate ethnic and cultural backgrounds. Nearly four centuries of Western rule, however, left the most indelible imprint on the country, serving as a conduit for the introduction of Western culture and as the catalyst for the emergence of a sense of Philippine political and cultural unity. The Christian churches built by the Spaniards and the mosques in the Muslim areas provided a spiritual anchor, while the educational system established by the United States and expanded by the Filipinos became a strong factor for socioeconomic progress. Nonetheless, traditionally strong family ties and other Asian moorings have remained. The revival of the barangay as the smallest unit of government has contributed to the revival of ancient traditions; and Asian and African history and literature—neglected in the past—have received more attention. Thus, the Philippines has been strengthening its Asian ties without abandoning its Western cultural acquisitions.

Early Spanish chroniclers testified that the Filipinos carved the images of their anitos (gods and goddesses) and ancestors in wood. They also played a variety of musical instruments, including flutes, nose flutes, and guitars. They performed appropriate songs and dances to celebrate courtship, marriage, the harvest, and other occasions. Many of these songs and dances, accompanied by the same musical instruments, have been preserved or reinterpreted and are presented by dance and song groups.

Major dance groups

Some of these groups—among them the Bayanihan, Filipinescas, Barangay, and Hariraya groups, as well as the Karilagan Ensemble and the groups associated with the guilds of the Manila and Fort Santiago theatres—have also performed abroad. Many Filipino musicians have risen to prominence, such as the composer and conductor Antonio J. Molina, the composer Felipe P. de Leon, known for his nationalistic themes, and the opera singer Jovita Fuentes. The Cultural Center of the Philippines, the Folk Arts Theatre, and the restored Manila Metropolitan Theatre (all in Manila) provide homes for the performing arts, featuring local and foreign opera and ballet. To encourage the development of arts the government gives awards of recognition and has established an Arts Center in Los Baños south of Manila.

Filipino painters have included Juan Luna; Fernando Amorsolo, who is known for his traditional rural scenes; the muralists Carlos V. Francisco and Vicente Manansala; and the modernists Victorio Edades and Arturo Rogerio Luz. Among the sculptors, Guillermo Tolentino and Napoleon Abueva are prominent. Tribesmen from mountainous regions in northern Luzon and craftsmen living northwest of Manila and in Paete on the eastern shore of Laguna de Bay are known for wood carving. Romblon and other nearby islands are noted for their marble sculptures. Among Filipino architects, the most notable include Juan F. Nakpil, Otilio Arellano, Fernando Ocampo, Leandro Locsin, Juan Arellano, Carlos Arguelles, and Tomas Mapua.

Folk traditions. Filipinos have a rich folklore tradition. Myths and legends deal with such subjects as the origin of the world, the first man and woman on Earth, why the sky is high, why the sea is salty, and why there are different races. Other tales are associated with the Spanish conquest. Muslim Filipinos have an epic called *Darangen,* and the Ilocanos of northern Luzon have another entitled *Biag ni Lamang.* Most folktales and popular cultural traits are still to be found in remote barrios and *sitios.* Dean S. Fansler, Maximo Ramos, and Armando and Carolina Malay have compiled collections of Philippine folktales.

Philippine epics

Museums. The Bureau of Science and the National Museum, both in Manila, have gathered artifacts connected with many of the folktales and popular beliefs. Several provinces—including Cagayan, Bulacan, Aklan, and Iloilo—have built their own museums. A few universities, such as the University of Santo Tomas, Silliman University in eastern Negros, and Mindanao State University in Lanao del Sur, likewise have added museums to their campuses. Nongovernmental organizations such as the Conservation Society and some historical groups have also sought to preserve the folklore heritage. The Cultural Centre of the Philippines opened in 1970 to encourage interest in the nation's heritage.

The media. A highly independent press developed in the Philippines when it was under U.S. administration, but many newspapers ceased publication during the period of martial law. Limited press freedom was granted in the early 1980s, with full freedoms returning after the change of government in 1986. Newspapers are published in English, Pilipino, and many of the country's vernacular dialects. The major English-language dailies—all published in Manila—include the *Bulletin Today, Times Journal,* and *Manila Times;* the *New Philippines Daily Express* has English and Pilipino editions. Among the major weeklies is *Veritas Weekly,* published by the Roman Catholic Church. The operators of radio and television stations belong to a national organization called the Association of Broadcasters in the Philippines.

For statistical data on the land and people of the Philippines, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR. (Eu.M.A./G.C.B.)

History

In ancient times, the inhabitants of the Philippines were a diverse agglomeration of peoples who arrived in various waves of immigration from the Asian mainland and who maintained little contact with each other. The Philippines is the only nation in Southeast Asia that became subject to Western colonialism before it had had the opportunity to develop either a centralized government ruling over a large territory or an advanced elite culture. Chinese traders were known to have been resident from about AD 1000, and some cultural influences from South Asia, such as a Sanskrit-based writing system, were carried to the islands by the Indonesian empires of Śrivijaya and Majapahit; but in comparison with other parts of the region, the influence of both China and India on the Philippines was of little importance. The peoples of the Philippine archipelago, unlike most of the other peoples of Southeast Asia, never adopted Hinduism or Buddhism.

PRE-SPANISH HISTORY

According to what can be inferred from somewhat later accounts, the Filipinos of the 15th century must have been primarily shifting cultivators, hunters, and fishers. Sedentary cultivation was the exception. Only in the mountains of northern Luzon, where elaborate rice terraces were built in the early centuries AD, were livelihood and social organization linked to a fixed territory. The lowland peoples lived in extended kinship groups known as barangays under the leadership of a *datu,* or chieftain. The barangay, which ordinarily numbered no more than a few hundred individuals, was usually the largest stable economic and political unit.

Barangay society

Within the barangay, the status system, though not rigid, appears to have consisted of three broad classes: the *datu* and his family and the nobility, freeholders, and "dependents." This third category consisted of three levels—sharecroppers, debt peons, and war captives—with the last two levels being termed slaves by Spanish observers. The

slave status was inherited but, through manumission and interclass marriage, seldom extended over more than two generations. The fluidity of the social system was, in part, the consequence of a bilateral kinship system in which lineage was reckoned equally through the male and female line. Marriage was apparently stable, though divorce was socially acceptable under certain circumstances.

Animism was the religion of the early Filipino, a mixture of monotheism and polytheism in which the latter dominated. The propitiation of evil spirits required numerous rituals, but there was no religious hierarchy. In religion, as in social structure and economic activity, there was considerable variation between—and even within—islands.

This pattern, however, began to change in the 15th century, when Islām was introduced to Mindanao and the Sulu islands through Brunei. Along with changes in religious beliefs and practices came new political and social institutions. By the mid-16th century two sultanates had been established, bringing under their sway a number of barangays. A powerful *datu* as far north as Manila embraced Islām. It was in the midst of this wave of Islāmic proselytism that the Spaniards arrived. Had they come a century later or had their motives been strictly commercial, Filipinos today might be a predominantly Muslim people.

THE SPANISH PERIOD

Spanish colonialist motives were not, however, strictly commercial. The Spanish at first viewed the Philippines as a stepping-stone to the riches of the Spice Islands, but, even after the Portuguese and Dutch had foreclosed that possibility, the Spaniards still maintained their presence in the archipelago.

The Portuguese navigator and explorer Ferdinand Magellan headed the first Spanish foray to the Philippines when he made landfall on Cebu in March 1521; a short time later he met an untimely death on the nearby island of Mactan. After three further expeditions had ended in disaster, Philip II, "the most Catholic of kings," for whom the islands are named, sent out Miguel López de Legazpi, who established the first permanent Spanish settlement, in Cebu, in 1565. The Spanish city of Manila was founded in 1571, and by the end of the 16th century most of the coastal and lowland areas from Luzon to northern Mindanao were under Spanish control. Friars marched with soldiers and soon accomplished the nominal conversion to Roman Catholicism of all the natives under Spanish administration. But the Muslims of Mindanao and Sulu, whom the Spanish called Moros, were never completely subdued by Spain; it took a 20th-century U.S. army to accomplish that task.

The encomienda

Spanish rule for the first 100 years was exercised in most areas through a type of tax farming imported from the Americas known as the encomienda. But abusive treatment of the native tribute payers and neglect of religious instruction by encomenderos, as well as frequent withholding of revenues from the crown, led to the abandonment of the system by the end of the 17th century. The governor general, himself appointed by the king, began to appoint his own civil and military governors to rule directly.

Central government in Manila retained a medieval cast until the 19th century, and the governor general was so powerful that he was often likened to an independent monarch. He dominated the Audiencia, or high court, was captain general of the armed forces, and enjoyed the privilege of engaging in commerce for private profit.

Manila dominated the islands not only as the political capital. The galleon trade with Acapulco, Mex., assured Manila's commercial primacy as well. This exchange of Chinese silks for Mexican silver kept in Manila those Spaniards seeking quick profit, and it also attracted a large Chinese community. The Chinese, despite periodic massacres at the hands of suspicious Spaniards, persisted and soon established a dominance of commerce that they have never relinquished.

Manila was also the ecclesiastical capital of the Philippines. The governor general was civil head of the church in the islands, but the archbishop vied with him for political supremacy. In the late 17th and 18th centuries, buttressed with the legal status of lieutenant governor, the archbishop frequently won. Augmenting their political power, religious orders, Roman Catholic hospitals and schools, and bishops acquired great wealth, mostly in land. Royal grants and devises formed the core of their holdings, but many arbitrary extensions were made beyond the boundaries of the original grants.

Power of the church

The power of the church derived not simply from wealth and official status. The priests and friars had a command of native languages rare among lay Spaniards, and in the provinces they outnumbered civil officials. Thus, they were an invaluable source of information to the colonial government. The cultural goal of the Spanish clergy was nothing less than the full Christianization and Hispanization of the Filipino. In the first decades of missionary work, native animism was vigorously suppressed; old practices were not tolerated. But as the Christian laity grew in number and the zeal of the clergy waned, it became more and more difficult to prevent the preservation of ancient beliefs and customs under Roman Catholic garb. Thus, even in the area of religion, pre-Spanish Filipino culture was not entirely destroyed.

Economic and political institutions were also altered under Spanish impact but perhaps less thoroughly than in the religious realm. The priests tried to move all the people into pueblos, or villages, surrounding the great stone churches. But the dispersed demographic patterns of the old barangays largely persisted. Nevertheless, the *datu*'s once hereditary position became subject to Spanish appointment.

Agricultural technology changed very slowly until the late 18th century, as shifting cultivation gradually gave way to more intensive sedentary farming, partly under the guidance of the friars. The socioeconomic consequences of the Spanish policies that accompanied this shift reinforced class differences. The *datus* and other representatives of the old noble class took advantage of the introduction of the Western concept of absolute ownership of land to claim, as their own, fields cultivated by their various retainers, even though traditional land rights had been limited to usufruct. These heirs of pre-Spanish nobility were known as the *principalia* and played an important role in the friar-dominated local government.

THE 19TH CENTURY

By the late 18th century, political and economic changes in Europe were finally beginning to affect Spain and, thus, the Philippines. Important as a stimulus to trade was the gradual elimination of the monopoly enjoyed by the galleon to Acapulco. The last galleon arrived in Manila in 1815, and by the mid-1830s Manila was open to foreign merchants almost without restriction. The demand for Philippine sugar and abaca, or hemp, grew apace, and the volume of exports to Europe expanded even further after the completion of the Suez Canal in 1869.

The growth of commercial agriculture resulted in the appearance of a new class. Alongside the landholdings of the church and the rice estates of the pre-Spanish nobility there arose haciendas of coffee, hemp, and sugar, often the property of enterprising Chinese-Filipino mestizos. Some of the families that gained prominence in the 19th century still play an important part in Philippine economics and politics.

Not until 1863 was there public education in the Philippines, and even then the church controlled the curriculum. Less than 20 percent of those who went to school could read and write Spanish, and far fewer could speak it properly. The limited higher education in the colony was entirely under clerical direction, but by the 1880s many sons of the wealthy were sent to Europe to study. There, nationalism and a passion for reform blossomed in the liberal atmosphere. Out of this talented group of overseas Filipino students arose the so-called Propaganda Movement. Magazines, poetry, and pamphleteering flourished. José Rizal, this movement's most brilliant figure, produced two political novels—*Noli me tangere* (1886; *The Social Cancer*) and *El filibusterismo* (1891; *The Reign of Greed*)—which had a wide impact in the Philippines.

In 1892 Rizal returned home and formed the Liga Filipina, a modest reform-minded society, loyal to Spain, that breathed no word of independence. But Rizal was quickly arrested by the overly fearful Spaniards, exiled to a remote island in the south, and finally executed in 1896. Meanwhile, within the Philippines there had developed a firm commitment to independence among a somewhat less privileged class.

The Katipunan

Shocked by the arrest of Rizal in 1892, these activists quickly formed the Katipunan under the leadership of Andres Bonifacio, a self-educated warehouseman. The Katipunan was dedicated to the expulsion of the Spaniards from the islands, and preparations were made for armed revolt. Filipino rebels had been numerous in the history of Spanish rule, but now for the first time they were inspired by nationalist ambitions and possessed the education that made success a real possibility.

THE PHILIPPINE REVOLUTION

In August 1896, Spanish friars uncovered evidence of the Katipunan's plans, and its leaders were forced into premature action. Revolts broke out in several provinces around Manila. After months of fighting, severe Spanish retaliation forced the revolutionary armies to retreat to the hills. In December 1897 a truce was concluded with the Spaniards. Emilio Aguinaldo, a municipal mayor and commander of the rebel forces, was paid a large sum and was allowed to go to Hong Kong with other leaders; the Spaniards promised reforms as well. But reforms were very slow in coming, and small bands of rebels, distrustful of Spanish promises, kept their arms; clashes grew more frequent.

Meanwhile, war had broken out between Spain and the United States. After the U.S. naval victory at Manila Bay in May 1898, Aguinaldo and his entourage returned to the Philippines with the help of Admiral George Dewey. Confident of U.S. support, he reorganized his forces and soon liberated several towns south of Manila. Independence was declared on June 12, which is now celebrated as Independence Day. In September a constitutional congress met in Malolos, north of Manila, drawing up a fundamental law derived from European and Latin-American precedents. A government was formed on the basis of that constitution in January 1899, with Aguinaldo as president.

Meanwhile, U.S. troops had landed and, with important Filipino help, forced the capitulation in August 1898 of the Spanish commander in Manila. The Americans, however, would not let Filipino forces enter the city. It was soon apparent to Aguinaldo and his advisers that earlier expressions of sympathy for Filipino independence by Dewey and U.S. consular officials had little significance. They felt betrayed.

U.S. commissioners to the peace negotiations in Paris had been instructed to demand the cession of the Philippines to the United States; such cession was confirmed with the signing of the peace treaty on Dec. 10, 1898. Ratification followed in the U.S. Senate in February 1899, but with only one vote more than the required two-thirds. Arguments of "manifest destiny" could not overwhelm a determined anti-imperialist minority.

War against the United States

By the time the treaty was ratified, hostilities had already broken out between American and Filipino forces. Since Filipino leaders did not recognize U.S. sovereignty over the islands and U.S. commanders gave no weight to Filipino claims of independence, the conflict was inevitable. It took two years of counterinsurgency warfare and some wise conciliatory moves in the political arena to break the back of the nationalist resistance. Aguinaldo was captured in March 1901 and shortly thereafter appealed to his countrymen to accept U.S. rule.

The Filipino revolutionary movement had had two goals, national and social. The first goal, independence, though realized briefly, was frustrated by the American decision to continue administering the islands. The goal of fundamental social change, manifest in the nationalization of friar lands by the Malolos Republic, was ultimately frustrated by the power and resilience of entrenched institutions. Share tenants who had rallied to Aguinaldo's cause, partly for economic reasons, merely exchanged one landlord for another. In any case, the proclamation of a republic in 1898 had marked the Filipinos as the first Asian people to try to throw off European colonialism.

THE PERIOD OF U.S. INFLUENCE

The juxtaposition of U.S. democracy and imperial rule over a subject people was sufficiently jarring to most Americans that, from the beginning, the training of Filipinos for self-government and ultimate independence—the Malolos Republic was conveniently ignored—was an essential rationalization for U.S. hegemony in the islands. Differences between Democratic and Republican policies focused on the speed with which self-government should be extended and the date on which independence should be granted.

In 1899 President William McKinley sent Cornell University president Jacob G. Schurman at the head of a five-man fact-finding commission to the Philippines. Schurman reported back that Filipinos wanted ultimate independence, but this had no immediate impact on policy. McKinley sent out the Second Philippine Commission in 1900, under William Howard Taft; by July 1901 it had established civil government.

In 1907 the Philippine Commission, which had been acting as both legislature and governor general's Cabinet, became the upper house of a bicameral body. The new 80-member Philippine Assembly was directly elected by a somewhat restricted electorate from single-member districts, making it the first elective legislative body in Southeast Asia. When Governor General Francis B. Harrison appointed a Filipino majority to the Commission in 1913, the American voice in the legislative process was further reduced.

Harrison was the only governor general appointed by a Democratic president in the first 35 years of U.S. rule. He had been sent by Woodrow Wilson with specific instructions to prepare the Philippines for ultimate independence, a goal that he enthusiastically supported. During Harrison's term, a Democratic Congress in Washington hastened to fulfill long-standing campaign promises to the same end. The Jones Act, passed in 1916, would have fixed a definite date for the granting of independence if the Senate had had its way, but the House prevented such a move. In its final form the act merely stated that it was the "purpose of the people of the United States" to recognize Philippine independence "as soon as a stable government can be established therein." Its greater importance was as a milestone in the development of Philippine autonomy. Under Jones Act provisions, the Commission was abolished and was replaced by a 24-member Senate, almost wholly elected. The electorate was expanded to include all literate males.

Some substantial restrictions on Philippine autonomy remained, however. Defense and foreign affairs remained exclusive U.S. prerogatives. American direction of Philippine domestic affairs was exercised primarily through the governor general and the executive branch of insular government. There was little more than one decade of thoroughly U.S. administration in the islands, however—too short a time in which to establish lasting patterns. Whereas Americans formed 51 percent of the civil service in 1903, they were only 29 percent in 1913 and 6 percent in 1923. By 1916 Filipino dominance in both the legislative and judicial branches of government also served to restrict the U.S. executive and administrative roles.

U.S. educational policy

By 1925 the only American left in the governor general's Cabinet was the secretary of public instruction, who was also the lieutenant governor general. This is one indication of the high priority given education in U.S. policy. In the first few years of U.S. rule, hundreds of schoolteachers came from the United States. But Filipino teachers were trained so rapidly that by 1927 they constituted nearly all of the 26,200 in public schools. The school population expanded fivefold in a generation; education consumed one-half of governmental expenditures at all levels, and educational opportunity in the Philippines was greater than in any other colony in Asia.

As a consequence of this pedagogical explosion, literacy doubled to nearly 50 percent in the 1930s, and educated

Filipinos acquired a common language and a linguistic key to Western civilization. By 1939, 27 percent of the population could speak English, a larger percentage than for any of the native dialects. Perhaps more important was the new avenue of upward social mobility that education offered. In educational policy was found the only successful U.S. effort to establish a socio-cultural basis for political democracy.

American attempts to create equality of economic opportunity were more modest and less successful. In a predominantly agricultural country the pattern of landownership is crucial. The trend toward greater concentration, which began in the 19th century, continued during the American period, despite some legal barriers. Vast American-owned plantations were forestalled, but legal restrictions had little effect on those politically well-connected Filipinos who were intent on amassing fortunes. The percentage of farmers under share tenancy doubled between 1900 and 1935; and the frustration of the tenants erupted in three small rebellions in central Luzon during the 1920s and 1930s.

Nor was U.S. trade policy conducive to the diffusion of economic power. From 1909 the Payne–Aldrich Tariff Act allowed free entry of Philippine products into the U.S. market; at the same time U.S. products, mostly manufactured, were exempted from tariff in the Philippines. The free flow of U.S. imports was a powerful deterrent to industrial growth. Export agriculture, especially sugar, prospered in the protected U.S. market. Owners of mills and large plantations profited most, thus reinforcing the political dominance of the landed elite.

American preparation of the Philippines for democratic self-government suffered from an inherent contradiction, perhaps not recognized at the time. Transferring governmental responsibility to those capable of undertaking it was not consistent with building a social and economic base for political democracy. Self-government meant, of necessity, assumption of power by those Filipinos who were already in positions of leadership in society. But these men came for the most part from the landed elite; preservation of their political and economic position was incompatible with equalization of opportunity. Even the expansion of an educated middle class did not necessarily result in a transformation of the pattern of power. Most middle-class aspirants for political leadership adjusted to the values and the practices of the existing power elite.

Formation of political parties

Filipino leaders quickly and skillfully utilized the opportunities for self-government that the Americans opened to them. The Filipino political genius was best reflected in an extralegal institution—the political party. The first party, the Federal Party, was U.S.-backed and stressed cooperation with the overlords, even to the point of statehood for the Philippines. But, when openly nationalist appeals were allowed in the 1907 election, the Nacionalista Party, advocating independence, won overwhelmingly. The Federalists survived with a new name, Progressives, and a new platform, ultimate independence after social reform. But neither the Progressives nor their successors in the 1920s, the Democrats, ever gained more than one-third of the seats in the legislature. The Nacionalista Party under the leadership of Sergio Osmeña and Manuel Quezon dominated Philippine politics from 1907 until independence.

More significant than the competition between the Nacionalistas and their opposition was the continuing rivalry between Osmeña and Quezon. In fact, understanding this personality conflict provides more insight into the realities of prewar Philippine politics than any examination of policy or ideology.

In 1933 the U.S. Congress passed the Hare–Hawes–Cutting Act, which set a date for Philippine independence. The act was a fulfillment of the vague pledge in the Jones Act; it was also responsive to the demands of a series of "independence missions" sent to Washington by the Philippine legislature. But this unprecedented transfer of sovereignty was decided upon in the dark days of the Great Depression—and with the help of some incongruous allies. The Depression had caused American farm interests to look desperately for relief, and those who suffered real or imaginary hurt from the competition of Philippine products sought to exclude those products. They had already failed in a direct attempt to amend the tariff on Philippine imports but found that the respectable cloak of the advocacy of independence increased the effectiveness of their efforts. Tied to independence was the end of free entry into American markets of Philippine sugar, coconut oil, rope, and other less important items. That these economic interests were able to accomplish what they did is partly explainable by the fact that their political clout was great compared to that of the small group of American traders and investors in the Philippines.

The Philippine legislature rejected the Hare–Hawes–Cutting Act, apparently as a result of the Osmeña–Quezon feud, much to the displeasure of American officialdom. But, when Quezon came to Washington the following year to work for a new bill, the same alliance of forces in the U.S. Congress obliged by producing the almost identical Tydings–McDuffie Act. It was endorsed by Quezon and accepted with alacrity by the Manila legislature. It provided for a 10-year commonwealth during which the U.S. would retain jurisdiction over defense and foreign affairs. Filipinos were to draft their own constitution, subject to the approval of the U.S. president.

A constitutional convention was quickly elected, a constitution framed (which bore a strong resemblance to its U.S. model), and the document approved by plebiscite and by President Franklin D. Roosevelt. The last governor general, Frank Murphy, became the first high commissioner, with more of a diplomatic than a governing role. The commonwealth was inaugurated on Nov. 15, 1935. The Nacionalista Party patched up its internal quarrels and nominated Quezon for president and Osmeña for vice president. They were elected overwhelmingly.

The Commonwealth

The commonwealth period was intended to be devoted to preparation for economic and political independence and perfection of democratic institutions. But even before the tragic events of 1941 the transition did not run smooth.

WORLD WAR II

Japanese aggression in China prompted much attention to military preparedness. Nearly one-fourth of the national budget was devoted to defense. General Douglas MacArthur, retiring as army chief of staff in Washington, was called by President Quezon to direct plans and preparations. Meanwhile, agrarian unrest festered, and leftist political activity grew. Quezon pushed significant reform legislation through the National Assembly, but implementation was feeble, despite the rapid accumulation of power in his hands.

The Japanese attack of Dec. 8, 1941, came at a time when the U.S. military buildup had hardly begun. Before Christmas, Manila was declared an "open city," while Quezon and Osmeña were evacuated to MacArthur's headquarters on Corregidor. Despite a desire, at one point, to return to Manila in order to surrender, Quezon was persuaded to leave the Philippines in March on a U.S. submarine; he was never to return. Osmeña also went. Filipino and American forces, under General Jonathan M. Wainwright, surrendered in May 1942. An Executive Commission, made up of more than 30 members of the old Filipino political elite, had been cooperating with Japanese military authorities in Manila since January.

The "independent Philippine Republic" under the Japanese

The Executive Commission lasted until September 1943, when it was superseded by an "independent Philippine Republic." The president, chosen by the Japanese, was José Laurel, former associate justice of the Commonwealth Supreme Court and the only Filipino to hold an honorary degree from Tokyo Imperial University. More than half of the Commonwealth Senate and more than one-third of the House served at one time in the Japanese-sponsored regime. Yet collaboration with Japan was neither as willing nor as widespread as elsewhere in Southeast Asia.

Even before the fall of Bataan Peninsula in April 1942, guerrilla units were formed throughout the Philippines. Most were led by middle-class officers and were enthusiastically pro-United States; in central Luzon, however, a major force was the Hukbalahap, which, under Communist leadership, capitalized on earlier agrarian unrest. Though in a number of instances collaborators secretly assisted guerrillas, many guerrillas in the hills were bitter

against those who appeared to benefit from the occupation. The differences between the two groups became an important factor in early postwar politics.

Soon after the U.S. landings on Leyte in October 1944, commanded by MacArthur, civil government was returned to the Commonwealth, at least in name. Sergio Osmeña, who had become president in exile on the death of Quezon in August, had few resources to deal with the problems at hand, however. Osmeña's role was complicated by the fact that MacArthur chose to lionize a leading collaborator who had also been in contact with U.S. military intelligence, Manuel A. Roxas. As president of the Senate, Roxas became, in effect, MacArthur's candidate for president. Roxas was nominated in January 1946 in a separate convention of the "liberal wing" of the Nacionalista Party, as it was first called. Thus was born the Philippines' second major political party, the Liberals.

Osmeña, though he had the advantages of incumbency, was old and tired and did not fully use the political tools he possessed. In April, Roxas was elected by a narrow margin. The following month he was inaugurated as the last chief executive of the Commonwealth, and on July 4, when the Republic of the Philippines was proclaimed, he became its first president.

THE EARLY REPUBLIC

Roxas, as expected, extended amnesty to all major collaborators with Japan. In the campaign for the election of 1949 there was an attempt to raise the collaboration issue against José Laurel, the Nacionalista presidential candidate, but it was not effective. In the fluidity of Philippine politics, "guerrillas" and "collaborators" were by that time to be found on both sides of all political fences.

Situation at independence

The Philippines had gained independence in the "ashes of victory." Intense fighting, especially around Manila in the last days of the Japanese retreat, had nearly destroyed the capital. The economy generally was in disarray. Rehabilitation aid was obviously needed, and President Roxas was willing to accept some onerous conditions placed implicitly and explicitly by the U.S. Congress. The Bell Act in the United States extended free trade with the Philippines for eight years, to be followed by 20 years of gradually increasing tariffs. The United States demanded and received a 99-year lease on a number of Philippine military and naval bases in which U.S. authorities had virtual territorial rights. And finally, as a specific requirement for release of U.S. war-damage payments, the Philippines had to amend its constitution to give U.S. citizens equal rights with Filipinos in the exploitation of natural resources—the so-called Parity Amendment.

The changing character of Philippine–U.S. relations has been a major theme in postwar Philippine history. The trend has been toward weakening of the link, achieved partly by diversification of Philippine external ties and partly by more articulate anti-American feeling. Economic nationalism, though first directed against the local Chinese community's dominance of retail trade, by the 1950s was focused on the special status of American business firms.

At independence, the military ties with the United States were as strong as the economic ones. Filipino troops fought against Communist forces in Korea, and noncombatant engineers augmented U.S. forces in the Vietnam War. Crucial to U.S. military action in Vietnam were bases in the Philippines. The Military Bases Agreement has been the greatest single cause of friction in relations between the United States and the Philippines. Since 1965, however, a series of agreements between the two countries have reduced the size and number of the U.S. bases and shortened base leases. In 1979 formal jurisdiction over the base areas passed to the Philippine government; and the constitution of 1987 formalized the process by which the bases agreement could be extended beyond the expiration in 1991 of base leases.

The nature and effectiveness of Filipino political institutions since independence has been a special concern of the former colonial power that helped establish them. For Filipinos, those institutions have determined the ability or inability to maintain domestic social order. Clumsy repression of dissent and the fraudulent election of the nation's second president, Elpidio Quirino, in 1949 set the stage for an intensification of the Communist-led Hukbalahap (Huk) Rebellion, which had begun in 1946. The rebellion also reflected a growing sense of social injustice among tenant farmers, especially in central Luzon. Suppression of the rebellion five years later, however, was attributable to American military aid as well as to the opening of the political process to greater mass participation, particularly during the campaign of Ramon Magsaysay, a uniquely charismatic figure in Filipino politics who was elected president in 1953. Magsaysay's attempts at social and economic reform failed largely because of the conservative outlook of the legislature and the bureaucracy. When Magsaysay died in a plane crash in 1957, leadership of the country fell to his vice president, Carlos P. Garcia. During Garcia's presidential term, and that of his reform-minded successor, Diosdado Macapagal (1961–65), unrest was usually channeled through the electoral process and peaceful protest.

Huk-balahap Rebellion

THE MARCOS REGIME AND AFTER

In November 1965, Ferdinand E. Marcos was elected to the presidency. His administration faced grave economic problems that were exacerbated by corruption, tax evasion, and smuggling.

In 1969 Marcos became the first elected president of the Philippines to win reelection. His campaign platform included the renegotiation of major treaties with the United States and trade with Communist countries. These promises reflected a change in the self-concept of the nation during the 1960s. The idea of the Philippines as an Asian outpost of Christianity was increasingly rejected in favour of the development of Asian cultural identity. Artists, musicians, and writers began to look to pre-Spanish themes for inspiration. More important was the trend toward seeking cultural identity through the national language, Pilipino. English remained the language of business, of most government documents, and of the greater part of university education. Demands that the government meet the social and economic needs of its citizenry continued.

A short-lived sign that the Filipino political system was again attempting to respond constructively to those needs was the choosing in 1970 of a widely representative Constitutional Convention in one of the most honest and peaceful elections in Philippine history. Large student demonstrations urged the convention to undertake a fundamental restructuring of political power.

Marcos, who was approaching the end of his constitutionally delimited eight years in office, had narrower goals: he pressed for the adoption of the parliamentary style of government, which would allow him to remain in power. He feared that the new constitution would not come into force before he lost the advantages of incumbency. At the same time, foreign investors, predominantly American, felt increased pressure from economic nationalists in the legislature.

Martial law. In September 1972 Marcos declared martial law, claiming that it was the last defense against the rising disorder caused by increasingly violent student demonstrations. One of his first actions was to arrest opposition politicians in Congress and the Constitutional Convention. Under martial law the regime was able to reduce violent urban crime, collect unregistered firearms, and suppress Communist insurgency in some areas. At the same time, a series of important new concessions were given to foreign investors, including a prohibition on strikes by organized labour, and a land-reform program was launched. In January 1973 Marcos proclaimed the ratification of a new constitution based on the parliamentary system, with himself as both president and prime minister. He did not, however, convene the interim legislature that was called for in the constitution.

Initial public reaction to martial law was mostly favourable except in Muslim areas of the south, where a separatist rebellion, led by the Moro National Liberation Front, broke out in 1973. Despite half-hearted attempts to negotiate a cease-fire, the rebellion continued to claim thousands of military and civilian casualties. Communist insurgency spread to new areas.

Disillusionment with martial law and with the consolidation of political and economic control by Marcos, his family, and close associates grew during the 1970s. The economy grew, but workers' real income dropped, and few farmers benefited from land reform. The precipitous drop in sugar prices in the early 1980s, coupled with lower prices and less demand for coconuts and coconut products—the most important export commodity—added to the country's economic woes; the government was forced to borrow large sums from the international banking community. Also troubling to the regime were more frequent reports of widespread corruption.

Elections for an interim National Assembly were finally held in 1978. The opposition—of which the primary group was led by the jailed former senator Benigno S. Aquino, Jr.—produced such a popular campaign that the official results, which gave the opposition virtually no seats, were believed to have been altered. In 1980 Benigno Aquino went into exile in the United States, and the following year, after suspending martial law, Marcos won a virtually uncontested election for a new six-year term.

The downfall of Marcos and return of democratic government. The assassination of Benigno Aquino as he returned to Manila in August 1983 was thought to have been the work of the military; it became the focal point of a renewed and strengthened opposition to Marcos' rule. By late 1985 Marcos, under mounting pressure both inside and outside the Philippines, called a snap presidential election for February 1986. Corazon C. Aquino, Benigno's widow, became the candidate of a coalition of opposition parties. Marcos was declared the official winner, but a public outcry over the election results precipitated a revolt that by the end of the month had driven Marcos from power. Aquino then assumed the presidency.

Election of 1986

Aquino's popularity and international support were instrumental in establishing the new government. Shortly after taking office, she abolished the constitution of 1973 and began ruling by decree. A new constitution was drafted and was ratified in February 1987 in a general referendum; legislative elections in May 1987 and the convening of a new bicameral congress in July 1987 marked the return of the form of government that had existed before the imposition of martial law in 1972.

Euphoria over the ouster of Marcos, however, was short-lived. The new government faced an enormous external debt, a depleted economy, and a growing threat from Moro and Communist insurgents. The Aquino administration also had to weather internal dissension, repeated coup attempts, and natural disasters such as a major earthquake and the 1991 eruption of Mount Pinatubo. The resumption of partisan politics, moreover, ended the coalition that had brought Aquino to power. Pro-Aquino candidates had won a sweeping victory in the 1987 legislative elections, but there was less support for her among those elected to provincial and local offices in early 1988. By the early 1990s, the criticisms charging her administration with weak leadership, corruption, and human rights abuses had begun to stick.

The presidential election of May 1992, in which Aquino was not a candidate, was a seven-way race in which the winner, Fidel Ramos, received less than 24 percent of the vote. Ramos, a former army chief of staff and defense minister under Aquino, was unpopular in some quarters because he had headed the agency that enforced martial law under Marcos before turning against Marcos to support Aquino in 1986. Some election observers had wryly noted that the winner might come to envy the losers, and indeed Ramos inherited the onus of having to deal with insurgencies from the right and the left, a severe energy crisis that produced daily electricity outages, an infrastructure in decay, a large foreign debt, and a population half of whom lived in deep poverty. (D.O.D.W./G.C.B.)

Ramos carried out social and economic reforms, and successful peace negotiations with Communist insurgents and Muslim separatists in the mid-1990s improved political stability and allowed the government to focus on economic development. The economy grew faster during this period than in the past due largely to increased foreign investment, political stability, and curbs on political corruption. As a result, the country weathered the Asian economic crisis of the late 1990s better than most other countries in the region.

In 1998 Joseph Estrada was elected president but was impeached and driven from office in 2000 on various charges including corruption and bribery. He was succeeded by his vice president, Gloria Macapagal Arroyo, a popular reform politician and the daughter of former president Diosdado Macapagal. (Ed.)

For later developments in the history of the Philippines, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 969 and 976, and the *Index.*

BIBLIOGRAPHY

Physical and human geography: General works include FUND FOR ASSISTANCE TO PRIVATE EDUCATION, *The Philippine Atlas,* 2 vol. (1975); DAVID JOEL STEINBERG, *The Philippines, a Singular and a Plural Place,* 4th ed. (2000); and RONALD E. DOLAN, (ed.), *Philippines, A Country Study,* 4th ed. (1993). FREDERICK L. WERNSTEDT *et al., Philippine Studies: Geography, Archeology, Psychology, and Literature* (1974); and JOHN V. HART (ed.), *Philippine Studies: History, Sociology, Mass Media, and Bibliography* (1978), and *Philippine Studies: Political Science, Economics, and Linguistics* (1981), provide useful summaries and bibliographies. The *Philippine Yearbook,* published by the government, contains current information.

DOMINGO C. SALITA, *Geography and Natural Resources of the Philippines* (1974); and DOMINGO C. SALITA and DOMINADOR Z. ROSELL, *Economic Geography of the Philippines* (1980), contain information on the economic resources of the Philippines. Two studies by the World Bank, *Philippines: Industrial Development Strategy and Policies* (1980), and *Philippines: A Framework for Economic Recovery* (1987), analyze economic conditions of the 1980s. Agricultural development policies are examined in RENE E. OFRENEO, *Capitalism in Philippine Agriculture* (1980); and *Vital Documents on Agrarian Reform in the New Society* (1979). Also useful are *Philippine Development Report* (irregular), published by the National Economic and Development Authority; and the *Philippine Statistical Yearbook.*

Social and cultural aspects are explored in MARY RACELIS HOLLNSTEINER (ed.), *Society, Culture, and the Filipino* (1979); LEONARD DAVIS, *The Philippines: People, Poverty, and Politics* (1987); ANDREW B. GONZALEZ, *Language and Nationalism: The Philippine Experience Thus Far* (1980); IRENE L. ORTIGAS and FELIX B. REGALADO, *Society and Culture in Rural Philippines,* rev. ed. (1978); ELENA YU and WILLIAM T. LIU, *Fertility and Kinship in the Philippines* (1980); ERIC S. CASINO, *The Filipino Nation* (1982); RENATO CONSTANTINO, *Neocolonial Identity and Counter-Consciousness: Essays on Cultural Decolonization* (1978); and GABRIEL CASAL *et al., The People and Art of the Philippines* (1981). (G.C.B.)

History: EUFRONIO M. ALIP, *Philippine History,* 10th rev. ed. (1974), and *Political and Cultural History of the Philippines,* new rev. ed., 2 vol. (1967), offer detailed coverage from the early period to the 20th century. RENATO CONSTANTINO, *The Philippines: A Past Revisited* (1975; also published as *A History of the Philippines: From the Spanish Colonization to the Second World War*), and *The Philippines: The Continuing Past* (1978); and HELEN R. TUBANGUI *et al., The Filipino Nation: A Concise History of the Philippines* (1982), are general histories. A good discussion of the early Spanish period is offered in JOHN L. PHELAN, *The Hispanization of the Philippines: Spanish Aims and Filipino Responses, 1565–1700* (1959, reprinted 1967). See also PEDRO S. DE ACHÚTEGUI and MIGUEL A. BERNAD, *Religious Revolution in the Philippines,* 2 vol. (1960–66), and *Documents Relative to the Religious Revolution in the Philippines* (1971); and ELIODORO G. ROBLES, *The Philippines in the Nineteenth Century* (1969). The Philippines' involvement in World War II is discussed in DAVID JOEL STEINBERG, *Philippine Collaboration in World War II* (1967). Studies of the Marcos and post-Marcos eras include PRIMITIVO MIJARES, *The Conjugal Dictatorship of Ferdinand and Imelda Marcos* (1976); MONINA ALLAREY MERCADO (ed.), *People Power: The Philippine Revolution of 1986* (1986); GARY HAWES, *The Philippine State and the Marcos Regime: The Politics of Export* (1987); and JOHN BRESNAN (ed.), *Crisis in the Philippines: The Marcos Era and Beyond* (1986), a comprehensive collection of essays by Filipino and American writers. Relations with the United States are treated in PETER W. STANLEY, *A Nation in the Making: The Philippines and the United States, 1899–1921* (1974); EMILIO AGUINALDO and VICENTE ALBANO PACIS, *A Second Look at America* (1957); and PETER W. STANLEY (ed.), *Reappraising an Empire: New Perspectives on Philippine-American History* (1984). (D.O.D.W./G.C.B.)

Philosophical Anthropology

The word anthropology was first used in the philosophical faculties of German universities at the end of the 16th century to refer to the systematic study of man as a physical and moral being. Philosophical anthropology is thus, literally, the systematic study of man conducted within philosophy or by the reflective methods characteristic of philosophy; it might in particular be thought of as being concerned with questions of the status of man in the universe, of the purpose or meaning of human life, and, indeed, with the issues of whether there is any such meaning and of whether man can be made an object of systematic study. What actually falls under the term philosophical anthropology, however, varies with conceptions of the nature and scope of philosophy. The fact that such disciplines as physics, chemistry, and biology—which are now classified as natural sciences—were until the 19th century all branches of natural philosophy serves as a reminder that conceptions of philosophy have changed.

Twentieth-century readings of philosophical anthropology are much narrower than those of previous centuries. Four possible meanings are now accepted: (1) the account of man that is contained in any comprehensive philosophy; (2) a particular philosophical orientation known as humanism (see HUMANISM), in which the study of man provides the foundation for all else—a position that has been prominent since the Renaissance; (3) a distinctive, 20th-century form of humanism that on occasion has claimed the label of "philosophical anthropology" for itself and that has taken the human condition, the personal being-in-the-world, as its starting point; and (4) any study of man that is regarded as unscientific. Philosophical anthropology has been used in the last sense by 20th-century antihumanists for whom it has become a term of abuse; antihumanists have insisted that if anthropology is to be possible at all it is possible only on the condition that it rejects the concept of the individual human subject. Humanism, in their eyes, yields only a prescientific, and hence a philosophical (or ideological), nonscientific anthropology.

By tracing the development of the philosophy of man, it will thus be possible to deal, in turn, with the four meanings of philosophical anthropology. First, however, it is necessary to discuss the concept of human nature, which is central to any anthropology and to philosophical debates about the sense in which and the extent to which man can be made an object of systematic, scientific study.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 10/52, and the *Index*.

This article is divided into the following sections:

The concept of human nature 550
Historical background of the philosophy of man 551
  Ancient Greece
  Medieval period
  Renaissance
The 16th and 17th centuries: the rise of scientific thought 553
  Rationalism versus Skepticism
  Development of anthropological studies
The 18th-century Enlightenment 555
  The natural history of man
  Man the rational subject
The 19th century 557
The 20th century: emergence of philosophical anthropology 558
  Frege and empiricist anthropology
  Husserl and philosophical anthropology
  Saussure, Freud, and antihumanism
Bibliography 561

THE CONCEPT OF HUMAN NATURE

The concept of human nature is a common part of everyday thought. The ordinary person feels that he comes to know human nature through the character and conduct of the people he meets. Behind what they do he recognizes qualities that often do not surprise him: he forms expectations as to the sort of qualities possessed by other human beings and about the ways they differ from, for example, dogs or horses. People are proud, sensitive, eager for recognition or admiration, often ambitious, hopeful or despondent, and selfish or capable of self-sacrifice. They take satisfaction in their achievements, have within them something called a conscience, and are loyal or disloyal. Experience in dealing with and observing people gives rise to a conception of a predictable range of conduct; conduct falling outside the range that is considered not to be worthy of a human is frequently regarded as inhuman or bestial whereas that which is exceptional—in that it lives up to standards which most people recognize but few achieve—is regarded as superhuman or saintly.

The common conception of human nature thus implicitly locates man on a scale of perfection, placing him somewhere above most animals but below saints, prophets, or angels. This idea was embodied in the theme, Hellenic in origin, of the Great Chain of Being—a hierarchical order ascending from the most simple and inert to the most complex and active: mineral, vegetable, animal, man, and finally divine beings superior to man. In the Middle Ages these divine beings constituted the various orders of angels, with God as the single, supremely perfect and omnipotent, ever-active being. There was a tendency in this theory to take for granted the commonality among all human beings, something by virtue of which they could be classified as fully human, which differentiates them from all other animals, and which gives them their place in the order of things. Yet, as with many notions that are habitually employed, the request for a precise definition of "human nature" proves highly problematic.

Great Chain of Being

The Greeks—most notably Plato and Aristotle—introduced the notion of form, nature, or essence as an explanatory, metaphysical concept. Variations on this concept were central to Western thought until the 17th century. Observation of the natural world raised the question of why creatures reproduced after their kind and could not be interbred at will and of why, for example, acorns grew into oaks and not into roses. To explain such phenomena it was postulated that the seeds, whether plant or animal, must each already contain within them the form, nature, or essence of the species from which they were derived and into which they would subsequently develop. This pattern of explanation is preserved in the modern biological concept of a genetic code that is embodied in the DNA molecular structure of each cell. There are important differences, however, between the modern concept of a genetic code and the older, Greek-derived concept of form or essence.

First, biologists are now able to locate, isolate, experimentally analyze, and manipulate DNA molecules in what has become known as genetic engineering. Being the structures responsible for physical development, DNA molecules represent the terms by which man can be biologically characterized. Forms or essences, on the other hand, were not observable; if they were granted any independent existence, it was as immaterial entities. The form, nature, or

essence of man or of any other kind of being was posited as a principle present in the thing, determining its kind by producing in it an innate tendency to strive to develop into a perfect example of itself—to fulfill its nature and to realize its full potential as a thing of a given kind. This gave rise to a teleological, or purposive, view of the natural world in which developments were explained by reference to the goal toward which each natural thing, by its nature, strives; *i.e.,* by reference to the ideal form it seeks to realize. By contrast, the genetic structure present in each cell is now invoked to explain the subsequent development of an organism in a "mechanistic" and nonpurposive way, in which development is shown to be dependent upon and determined by preexisting structures and conditions.

Second, genetic mutability forms an essential part of modern evolutionary biology. Not only are there genetic differences between individuals of a given species to account for differences between them in features, such as coloration, but random genetic mutation in the presence of changing environmental conditions may result in alterations to the genetic constitution of the species as a whole. Thus, in evolutionary biological theory species are not stable; natural kinds do not have the fixed, immutable forms or essences characteristic of biology before the advent of evolutionary theory.

Within either framework, if human nature is understood simply as man's special form of that which is biologically inherited in all species, there remains the delicate problem of discovering, in any given case, exactly what role environment plays in determining the actual characteristics of mature members of the species. Even in the case of purely physiological characteristics this may be far from straightforward: for example, the extent to which diet, exercise, and conditions of work determine such things as susceptibility to heart disease and cancer remains the subject of intensive scientific investigation. In the case of behavioral and psychological characteristics, such as intelligence, the problems are multiplied to the point where they are no longer problems that can be answered by purely empirical investigation. There is room for much conceptual debate about what is meant by intelligence and over what tests, if any, can be supposed to yield a direct measure of this capacity, and thus provide evidence that an individual's level of intelligence is determined at birth (by nature) rather than by subsequent exposure to the environment (nurture) that conditions the development of all his capacities.

Intelligence

This debate—whether the variation in intelligence levels is a product of the conditions into which people all having the same initial potential are born, or whether it is a reflection of variations in the capacities with which they are born—is very closely related to the question of whether there is such a thing as human nature common to all human beings, or whether there are intrinsic differences among those whom we recognize as belonging to the biological species *Homo sapiens.* This is because, as the name *Homo sapiens* suggests, man is traditionally thought to be distinguished from and privileged above other animals by virtue of his possession of reason, or intellect. When the intellect is positively valued as that which is distinctively human and which confers superiority on man, the thought that different races of people differ by nature in their intellectual capacities has been used as a justification for a variety of racist attitudes and policies. Those of another race, of supposedly lesser intellectual development, are classified as less than fully human and therefore as needing to be accorded less than full human rights. Similarly, the thought that women are by nature intellectually inferior to men has been used as a justification for their domination by men, for refusing them education, and even for according them the legal status of property owned by men. On the other hand, if differences in adult intellectual capacity are regarded as a product of the circumstances in which potentially similar people are brought up, the attitude is to consider all as equally human but some as having been more privileged when growing up than others.

More radically, the evidence for variations in intelligence levels may be questioned by challenging the objectivity of the standards relative to which these levels are assessed. It may be argued that conceptions of what constitutes a rational or intelligent response to a situation or to a problem are themselves culturally conditioned, a product of the way in which the members of the group devising the tests and making the judgments have themselves been taught to think. Such an argument has the effect of undermining claims by any one human group to intellectual superiority over others, whether these others be their contemporaries or their own forebears. Hence, they may also be used to discredit any idea of a progressive development of human intellectual capacities.

These debates about intelligence and rationality provide an example of the complexity of the impact of evolutionary biology on conceptions of human nature, for the dominant traditions in Western thought about human nature have tended to concentrate attention more on what distinguishes man from other animals than on the strictly biological constitution that he largely shares with them. Possession of reason or intellect is far from being the only candidate considered for such a distinguishing characteristic. Man has been characterized as essentially a tool user, or fabricator (Homo faber), as essentially social, as essentially a language user, and so on. These represent differing views concerning the fundamental feature that gives rise to all the other qualities regarded as distinctively human and which serve to mark man off from other animals. These characteristics all centre on mental, intellectual, psychological—*i.e.,* nonphysiological—characteristics and thus leave scope for debate about the relation between mind and body. So long as this question remains open, and so long as mental or intellectual constitution remains the central consideration in discussions of human nature, the question of changes in—and of the possible evolution of—human nature will remain relatively independent of those devoted to physiological change and hence of strictly biological evolution.

Human nature

Until the 15th century the standard assumption was that man had a fixed nature, one that determined both his place in the universe and his destiny. The Renaissance humanists, however, proclaimed that what distinguishes man from all other creatures is that he has no nature. This was a way of asserting that man's actions are not bound by laws of nature in the way that those of other creatures are. Man is capable of taking responsibility for his own actions because he has the freedom to exercise his will. This view received two subsequent interpretations.

First, the human character is indefinitely plastic; each individual is given determinate form by the environment in which he is born, brought up, and lives. In this case, changes or developments in human beings will be regarded as the product of social or cultural changes, changes that themselves are often more rapid than biological evolution. It is thus to disciplines such as history, politics, and sociology, rather than to biology, that one should look for an understanding of these processes. But if disciplines such as these must constitute the primary study of man, then the question of the extent to which this can be a strictly scientific study arises. The methods of history are not, and cannot be, those of the natural sciences. And the legitimacy of the claims of the so-called social or human sciences to genuine scientific status has frequently been called into question and remains a focus for debate.

Second, each individual is autonomous and must "make" himself. Assertion of the autonomy of man involves rejection of the possibility of discovering laws of human behaviour or of the course of history, for freedom is precisely not being bound by law, by nature. In this case, the study of man can never be parallel to the natural sciences with their theoretical structures based on the discovery of laws of nature. (M.E.T.)

HISTORICAL BACKGROUND OF THE PHILOSOPHY OF MAN

In the tradition of Western thought up to the 20th century, the study of man has been regarded as a part of philosophy. Two sayings that have been adopted as mottoes by those who see themselves as engaged in philosophical anthropology date from the 5th century BC. These are: "Man is the measure of all things" (Protagoras) and "Know thyself" (a saying from the Delphic oracle, echoed by Heracleitus and Socrates, among others). Both reflect

the specific orientation of philosophical anthropology as humanism, which takes man as its starting point and treats man and the study of man as the centre, or origin, on which all other disciplines ultimately depend.

Man, the world, and God have constituted three important foci of Western thought from the beginnings of its recorded history; the relative significance of these three themes, however, has varied from one epoch to another. Western thought has laid greater stress on the existence of the individual human being than have the great speculative systems of the East; in Brahmanism, for example, personal identity dissolves in the All. But even so it was not until the Renaissance that man became the primary focus of philosophical attention and that the study of human nature began to displace theology and metaphysics as "first philosophy"—the branch of philosophy that is regarded as forming the foundation for all subsequent philosophy and that provides the framework for all scientific investigation.

Ancient Greece. From late antiquity onward differing views of man were worked out within a framework that was laid down and given initial development by Plato and later by Aristotle. Plato and Aristotle concurred in according to metaphysics the status of first philosophy. Their differing views of man were a consequence of their differing metaphysical views.

Plato

Plato's metaphysics was dualistic: the everyday physical world of changeable things, which man comes to know by the use of his senses, is not the primary reality but is a world of appearances, or phenomenal manifestations, of an underlying timeless and unchanging reality, an immaterial realm of Forms that is knowable only by use of the intellect. This is the view expressed in the *Republic* in his celebrated metaphor of the cave, where the changeable physical world is likened to shadows cast on the wall of a cave by graven images. To know the real world the occupants of the cave must first turn around and face the graven images in the light that casts the shadows (*i.e.*, use their judgment instead of mere fantasy) and, second, must leave the cave to study the originals of the graven images in the light of day (stop treating their senses as the primary source of knowledge and start using their intellects). Similarly, human bodily existence is merely an appearance of the true reality of human being. The identity of a human being does not derive from the body but from the character of his or her soul, which is an immaterial (and therefore nonsexual) entity, capable of being reincarnated in different human bodies. There is thus a divorce between the rational/spiritual and the material aspects of human existence, one in which the material is devalued.

Aristotle

Aristotle, however, rejected Plato's dualism. He insisted that the physical, changeable world made up of concrete individual substances (people, horses, plants, stones, etc.) is the primary reality. Each individual substance may be considered to be a composite of matter and form, but these components are not separable, for the forms of changeable things have no independent existence. They exist only when materially instantiated. This general metaphysical view, then, undercut Plato's body–soul dualism. Aristotle dismissed the question of whether soul and body are one and the same as being as meaningless as the question of whether a piece of wax and the shape given to it by a seal are one. The soul is the form of the body, giving life and structure to the specific matter of a human being. According to Aristotle, all human beings are the same in respect to form (that which constitutes them as human), and their individual differences are to be accounted for by reference to the matter in which this common form is variously instantiated (just as the different properties of golf and squash balls are derived from the materials of which they are made, while their common geometrical properties are related to their similar size and shape). This being so, it is impossible for an individual human soul to have any existence separate from the body. Reincarnation is thus ruled out as a metaphysical impossibility. Further, the physical differences between men and women become philosophically significant, the sex of a person becoming a crucial part of his or her identity.

Although Plato and Aristotle gave a different metaphysical status to forms, their role in promoting and giving point to investigations of human nature was very similar. They both agreed that it is necessary to have knowledge of human nature in order to determine when and how human life flourishes. It is through knowledge of shared human nature that we become aware of the ideals at which we should aim, achieved by learning what constitutes fulfillment of our distinctively human potential and the conditions under which this becomes possible. These ideals are objectively determined by our nature. But we are privileged in being endowed with the intellectual capacities that make it possible for us to have knowledge of this nature. Development of our intellectual capacities is thus a necessary part and precondition of a fulfilled human existence.

Medieval period. Western medieval culture was dominated by the Christian Church. This influence was naturally reflected in the philosophy of the period. Theology, rather than metaphysics, tended to be given primacy, even though many of the structures of Greek philosophy, including its metaphysics, were preserved. The metaphysics of form and matter was readily assimilable into Christian thought, where forms became ideas in the mind of God, the patterns according to which he created and continues to sustain the universe. Christian theology, however, modified the positions, requiring some sort of compromise between Platonic and Aristotelian views. The creation story in the book of Genesis made man a creature among other creatures, but not a creature like other creatures; man was the product of the final act of divine initiative, was given responsibility for the Garden of Eden, and had the benefit of a direct relationship with his creator. The Fall and redemption, the categories of sin and grace, thus concern only the descendants of Adam, who were given a nature radically different from that of the animals and plants over which they were given dominion. Man alone can, after a life in this world, hope to participate in an eternal life that is far more important than the temporal life that he will leave. Thus, belief in a life after death makes it impossible to regard man as wholly a natural being and entails that the physical world now inhabited by man is not the sole, or even the primary, reality. Yet, the characteristically Christian doctrine of the resurrection of the body also entails that the human body cannot be regarded as being of significance only in the mortal, physical world.

Against the background of these constraints, Christian philosophy first, through the writings of St. Augustine, gave prominence to Platonic views. But this emphasis was superseded in the 12th century by the Aristotelianism of St. Thomas Aquinas. Augustine's God is a wholly immaterial, supremely rational, transcendent creator of the universe. The twofold task of the Christian philosopher, a lover of wisdom, is to seek knowledge of the nature of God and of his own soul, the human self. For Augustine the soul is not the entire man but his better part. There remains a Platonic tendency to regard the body as a prison for the soul and a mark of man's fallen state. One of the important consequences of Augustine's own pursuit of these two endeavours was the emphasis he came to place on the significance of free will. He argued that since the seat of the will was reason, when people exercise their will, they are acting in the image of God, the supreme rational being. Thomas Aquinas, while placing less emphasis on the will, also regarded man as acting in the image of God to the extent that he exercises and seeks to fulfill his intelligent nature. But he rejected the Platonic tendency to devalue the body, insisting that it is part of the concept of man that he have flesh and bone, as well as a soul.

Man as a creature of God

But whatever the exact balance struck in the relation between the mind and body, the view of man was first and foremost as a creature of God; man was privileged by having been created in the image of God and given the gift of reason in virtue of which he also has free will and must take the burden of moral responsibility for his own actions. In order to fulfill his distinctively human nature man must thus order his thoughts and actions in such a way as to reflect the supremacy of religious values.

In popular medieval culture there was also, however, a strong undercurrent of thoroughly fatalistic thought. This

was reflected in the popularity of astrology and alchemy, both of which appealed to the idea that events on Earth are governed by the influence of the heavenly bodies.

Renaissance. It was in the cultural context of the Renaissance, and in particular with the Italian humanists and their imitators, that the centre of gravity of reflective thought descended from heaven to earth, with man, his nature, and his capacities and limitations becoming a primary focus of philosophical attention. This gave rise to the humanism that constitutes philosophical anthropology in the second sense. Man did not thereby cease to view himself within the context of the world, nor did he deny the existence of God; he did, however, disengage himself sufficiently from the bonds of cosmic determination and divine authority to become a centre of interest in his own eyes. In ancient literature the educated people of the West rediscovered a clear conscience instead of the guilty conscience of Christianity; at the same time, the great inventions and discoveries suggested that man could take pride in his accomplishments and regard himself with admiration. The themes of the dignity and excellence of man were prominent in Italian humanist thought and can be found clearly expressed in Giovanni Pico della Mirandola's influential *De hominis dignitate oratio* (Oration on the Dignity of Man), written in 1486. In this work Pico expresses a view of man that breaks radically with Greek and Christian tradition: what distinguishes man from the rest of creation is that he has been created without form and with the ability to make of himself what he will. Being without form or nature he is not constrained, fated, or determined to any particular destiny. Thus, he must choose what he will become. (In the words of the 20th-century existentialists, man is distinguished by the fact that for him existence precedes essence.) In this way man's distinctive characteristic becomes his freedom; he is free to make himself in the image of God or in the image of beasts.

Rise of humanism

This essentially optimistic view of man was a product of the revival of Neoplatonist thought. Its optimism is based on a view of man as at least potentially a nonnatural, godlike being. But this status is now one that must be earned; man must win his right to dominion over nature and in so doing earn his place beside God in the life hereafter. He must learn both about himself and about the natural world in order to be able to achieve this. This was, however, only one of two streams of humanist thought. The other (more Aristotelian) was essentially more pessimistic and skeptical, stressing the limitations on man's intellectual capacities. There is an insistence on the need to be reconciled to the fact of man's humanity rather than to persist in taking seriously his superhuman pretensions and aspirations. These two differently motivated movements to focus attention on man himself, on his nature, his abilities, his earthly condition, and his relation to his material environment became more clearly articulated in the 16th and 17th centuries in the opposition between the rationalist and empiricist approaches to philosophy.

Two views of man in relation to God

THE 16TH AND 17TH CENTURIES: THE RISE OF SCIENTIFIC THOUGHT

Rationalism versus skepticism. The thought of Michel de Montaigne, the 16th-century French skeptical author of the *Essais* (1580–95; *Essays*), represented one of the first attempts at anthropological reflection (*i.e.*, reflection centred on man, which explores his different aspects in a spirit of empirical investigation that is freed from all ties to dogma). Skepticism, the adoption of an empirical approach, and liberation from dogmatic authority are linked themes stemming from the more pessimistic views of man's capacity for knowledge. The emphasis on man's humanity—on the limited nature of his capacities—leads to a denial that he can, even by the use of reason, transcend the realm of appearances; the only form of knowledge available to him is experimental knowledge, gained in the first instance by the use of the senses. The effect of this skeptical move was twofold. The first effect was a liberation from the dogmatic authority of claims to knowledge of a reality behind appearances and of moral codes based on them; skeptical arguments were to the effect that human beings are so constituted that such knowledge must always be unavailable to them. The second effect was a renewal of attention to and interest in the everyday world of appearances, which now becomes the only possible object of human knowledge and concern. The project of seeking knowledge of a reality behind appearances must be abandoned because it is beyond the scope of human understanding. And this applies as much to man himself as to the rest of the natural world; he can be known only experientially, as he appears to himself.

The thought of Montaigne

The anthropology of Montaigne began with a turning in upon himself; it gave priority to that reality which was within. Montaigne, however, was also witness to a renewal of knowledge brought about by numerous discoveries that made the horizons of the traditional universe expand greatly. For him, self-awareness already reflected an awareness of the surrounding world; it wondered about the "savages" of America and about the cannibals that were so different from him and yet so near; it compared the intelligence of man with that of beasts and accepted the idea of a relationship between animal existence and human existence. The idea that moral codes are the work of man, rather than reflective of an objective order, opened up the possibility of recognizing the legitimate existence of a plurality of codes and thus of the empirical study—rather than an immediate condemnation and rejection—of the customs of others.

Work of Descartes. By contrast, the work of the 17th-century French philosopher René Descartes represented a continuation of the theme of optimism about man's capacities for knowledge. Descartes explicitly set out, in his *Meditations* (first published in 1641), to beat the skeptics at their own game. He used their methods and arguments in order to vindicate claims to be able to have nonexperimental knowledge of a reality behind appearances. The *Meditations* thus also begins with a turning in of Descartes upon himself but with the aim of finding there something that would lead beyond the confines of his own mind.

Cartesianism occupies a key position in the history of modern Western philosophy; Descartes is treated as a founding father by most of its now diverse traditions. His work is characteristic of the philosophical effort of the 17th century, which was engaged in a struggle to achieve a synthesis between old established orders and the newly proclaimed freedoms that were based on a skeptical rejection of the older orders. There are undeniable tensions in the philosophy of this period that are the product of various unsuccessful attempts to reconcile two very different views of man in relation to God and the world.

The first, the authoritarian view, was that inherited from medieval philosophy and from Thomist theology. It derived its ideal of human freedom from the Stoic conception of the wise man, who, in the 17th century was called a man of *honestas* (the French concept of *honnêteté*). The man of *honestas* seeks freedom in the discovery of and obedience to the order and law on which the world is grounded. He believes that there is such a law, that he has a "place" in the scheme of things, and that he is bound to his fellow human beings by that nature through which he participates in this higher order. He tends to look to the authorities—whether these be church, state, or classical texts—for knowledge of this order, for it is not to be found at the level of experience; it is a "higher" order. His worldview is derived from a mixture of Platonic and Aristotelian (realist) metaphysics.

The second, the libertarian view, was that of the skeptical humanists—individualists and freethinkers, skeptical of any preestablished order, or at least of man's ability to know what it is or might be. The skeptical humanist is therefore untrammeled by it. He deploys skeptical arguments to release the individual from the constraints and demands of outer authorities. He is free to do what he wills or desires and to make his own destiny, for there can be no knowledge of objective norms. Human knowledge is limited to experience, to what is sensed, and people must therefore make their own order within experience. His view is descended from the *via moderna* of the medieval philosopher William of Ockham and the nominalists.

The synthesis sought was a position that would incorporate recognition of the individual and of his freedom under

universal principles of order, a reconciliation of will with reason. This was sought via a nonauthoritarian conception of objective knowledge, which was the same conception that gave rise to modern science. This required, on the one hand, arguments to combat those of the skeptical freethinkers—arguments that demonstrated that there was an objective order external to human thought and that humans have the capacity not merely to know of its existence but also to discover something of its nature. On the other hand, it was necessary to establish, against the authorities, that each individual, insofar as he is rational, has the capacity to acquire knowledge for himself, by the proper use of his reason. It is this second requirement that produced numerous treatises on the scope and limits of human understanding and on the method of acquiring knowledge. The focus was now firmly fixed on the nature of human thought and on the procedures available to it.

Descartes utilized the skeptic's own arguments to urge a meditative turning inward. This inward journey was designed to show that each human being can come to knowledge of his intellectual self and that as he does so he will find within himself the idea of God, the mark of his creator, the mark that assures him of the existence of an objective order and of the objective validity of his rational faculties. The foundation and starting point of Cartesian knowledge is, for each individual, within himself, in his experience of the certainty that he must have of his own existence and in the idea of a perfect, infinite being, in other words, an idea that he finds within himself, of a being whose essence entails God's existence, and of whose existence man can thus be assured on the basis of his idea of God.

Descartes thus preserved and built on Montaigne's emphasis on self-consciousness, and this is what marks the changed orientation in philosophy that constitutes philosophical anthropology in the stricter, second sense. As the French scientist and religious philosopher Blaise Pascal realized, the question had now become one of whether man finds within himself the basis of loyalty to a universal order of reason and law with which his own thought and will is continuous, or whether he finds, by inner examination, that order, at least insofar as it can be known, is relative to his feeling, desire, and will.

Failure of Descartes's proofs of the existence of God

The attempt to regain an objective order by looking inward apparently fails with the failure of Descartes's proofs of the existence of God, proofs that his contemporaries (even those who, like Gottfried Wilhelm Leibniz, were sympathetic to many aspects of the project) were quick to criticize. Reaction to this failure was twofold. In the work of rationalist philosophers, such as Spinoza, Leibniz, and Malebranche, there is a return to the classical Greek approach to philosophy through metaphysics. Empiricists, such as Locke, Condillac, and Hume, on the other hand, retain the Cartesian, introspective basis seeking what Hume calls a mitigated skepticism. This is a position that recognizes essential limitations placed on human cognitive capacities by assuming that experience is the only source of knowledge, but that affirms the value of the knowledge so gained and seeks to define the project of natural science as a quest for objective order within this domain.

Work of Locke. John Locke, for instance, argued that while man cannot prove that the material world exists, his senses give him evidence affording all the certainty that he needs. Locke's position is, however, essentially dualist: mind and body remain distinct even though pretensions to intellectual transcendence are given up. Moreover, Locke regarded it as in principle impossible for humans to have any understanding of the relation between mind and body. All perceptions of one's own body, as of the rest of the material world, are ideas in one's mind. It is impossible to adopt any vantage point outside oneself from which to observe the correlation between a condition of one's body and one's perception of this condition. Where other people are concerned, their bodies and behaviour can be observed but an observer can have no direct perception of what is going on in their minds. There is thus a bifurcation in the study of man. The mind and its contents are known to each person by introspection; it is presumed that the minds of all people work in basically the same way so that introspection provides evidence for human psychology. Other people, their bodies, and their behaviour are known by observation in exactly the same way that knowledge of any other natural object is obtained. One infers from their behaviour that they have minds like one's own and on this basis attributes psychological states to them.

In keeping with this bifurcation Locke distinguished between the terms "man" and "person," reserving "man" for the animal species, an object of study for natural historians. "Person" is used to denote the moral subject, the being who can be held responsible for his actions and thus praised, blamed, or punished. According to Locke, what constitutes a person is a characteristic continuity of consciousness, which is not merely rational thought but the full range of mental states accessible to introspection. Just as a tree is a characteristic organization of life functions sustained by exchanges of matter, so a person is a characteristic organization of mental functions continuing through changes in ideas (the matter of thought). A person can be held responsible for an action only if he acknowledges that action as one which he performed; *i.e.*, one of which he is conscious and remembers having performed.

The empiricist position thus opens up the possibility of empirical studies both of man as a natural and as a moral being and puts these studies on a par with the natural sciences. But it does so in such a way that the resulting picture lacks any integral unity, for man is an incomprehensible union of body and mind.

Discovery of the anthropoid apes

Development of anthropological studies. A renewed study of the natural history of man was stimulated by European encounters with the great anthropoid apes of Africa (Angola) and Asia (the Sunda Islands) at the beginning of the 16th century. Until then Europe had known only the smaller monkeys, which were too far removed from the human species to present any confusion. The discovery of the chimpanzee and the orangutan (meaning "man of the woods" in Malay) raised such questions as whether the anthropoid, who resembles man, is an animal or a man, and why it should be considered an ape and not a man. In the climate of opinion—typified by Locke and fostered by the Royal Society of London, with its enthusiasm for empirical observation—these questions prompted the detailed observational studies of a leading member of the society, Edward Tyson.

Work of Tyson. Tyson had the opportunity to study the remains of a young chimpanzee (named Pygmie) from Angola that had died in London several months after its arrival. His research was published by the Royal Society in 1699 under the title *Orang-Outang, sive Homo Sylvestris: or, The Anatomy of a Pygmie Compared with That of a Monkey, an Ape, and a Man.* This treatise, a landmark in anthropology and comparative anatomy, is remarkable for the empirical approach used in the investigation. Tyson's precise measurements, his complete exploration of the external and internal structures of the animal, and his minutely detailed sketches permitted him to pose what is perhaps the central problem of physical anthropology: whether it is possible to find among the anatomical or physiological characteristics of the ape the justification for asserting a radical difference between ape and man, notwithstanding all their similarities. He analyzed in great detail the similarities and dissimilarities between a chimpanzee and a man. He emphasized the fact that the ape is a quadrumane (having four hands) rather than a quadruped (having four feet); unlike the human foot, its foot has an opposable, and thus thumblike, big toe. The arrangement of the internal organs allows the erect posture that makes the ape similar to man. But on an analysis of the form and mass of the brain and speech apparatus, Tyson concluded that he was unable to determine, from a strictly anatomical point of view, why the ape is incapable of thinking and speaking.

Integral to the empiricism that forms the philsophical background to Tyson's work was a rejection of the whole notion of forms or essences as objectively determining fixed and strict demarcations within the natural world. Classification was the work of man imposed upon a natural continuum, which replaced the older ladderlike conception of the Chain of Being. This encouraged a quest for

"missing links," examples of intermediary forms between those already recognized. For example, zoophytes (invertebrate animals resembling plants, such as sponges) were said to form the link between the vegetable order and the animal order. For Tyson, the chimpanzee was the missing link between animal and man.

Discovery of other civilizations

Emergence of cultural anthropology. If physical anthropology was born out of Western man's encounter with the anthropoid apes, cultural anthropology was made necessary by his encounter with people in the rest of the world during the great voyages of discovery begun in the 15th century. Cultural anthropology became the product of the confrontation between the classical values of the West and the opposing values and customs of newly discovered civilizations.

The "savage" appeared to manifest a style of humanity that was a contradiction of the certainties that had sustained Europeans for centuries. The shock was such that the naked Indian and the cannibal were at first assumed not to belong to the human race; this approach enabled Europeans to avoid the problem. This solution was, however, rejected by Pope Paul III in 1537 in his bull, or decree, *Sublimus Deus* ("The Transcendent God"), according to which Indian savages were human beings; they had souls and, as such, could be initiated into the Christian religion. This left the problem of how to reconcile the increasingly manifest human diversity with the theological requirement of human unity. One solution was to account for diversity in terms of environment, including cultural environment, and to regard the "savage" as a "primitive," as a "man of nature," who remained close to an initial state from which a privileged part of humanity had been able to remove itself by a continued effort at community and individual advancement. A study of the history of man endeavoured to bring to light the successive stages through which the human species had passed along the way to the present civilized societies. The themes of "civilization" and "progress" were among the principal preoccupations of the Enlightenment.

THE 18TH-CENTURY ENLIGHTENMENT

What has come to be known as the Enlightenment is characterized by an optimistic faith in the ability of man to develop progressively by using reason. By coming to know both himself and the natural world better he is able to develop morally and materially, increasingly dominating both his own animal instincts and the natural world that forms his environment. However, the divergence between rationalist and empiricist traditions continues, giving rise to rather different interpretations of this theme.

The natural history of man. The writings of the Scottish philosopher David Hume give a clear statement of the implications of empiricist epistemology for the study of man. Hume argued first that scientific knowledge of the natural world can consist only of conjectures as to the laws, or regularities, to be found in the sequence of natural phenomena. Not only must the causes of the phenomenal regularities remain unknown but the whole idea of a reality behind and productive of experience must be discounted as making no sense, for experience can afford nothing on the basis of which to understand such talk. Given that this is so, and given that man also observes regularities in human behaviour, the sciences of man are possible and can be put on exactly the same footing as the natural sciences. The observed regularities of human conduct can be systematically recorded and classified, and this is all that any science can or should aim to achieve. Explanation of these regularities (by reference to the essence of man) is not required in the sciences of man any more than explanation of regularities is required in the natural sciences.

Man thus becomes an object of study by natural history in the widest possible sense. All observations—whether of physiology, behaviour, or culture—contribute to the empirical knowledge of man. There is no need, beyond one of convenience, to compartmentalize these observations, since the method of study is the same whether marital customs or skin colour is the topic of investigation; the aim is to record observations in a systematic fashion making generalizations where possible. Such investigations into the natural history of man were undertaken by Linnaeus, Buffon, and Blumenbach, among others.

Work of Linnaeus, Buffon, and Blumenbach

In his *Systema Naturae* (1735), the Swedish naturalist Carolus Linnaeus (Carl von Linné) gave a very precise description of man, placing him among the mammals in the order of primates, alongside the apes and the bat. But the distinguishing characteristic of man remains his use of reason; something that is not dependent on any physiological characteristics. Moreover, the variations that are to be found within the genus *Homo sapiens* are the product of culture and climate. In later editions of *Systema Naturae,* Linnaeus presented a summary of the diverse varieties of the human species. The Asian, for example, is "yellowish, melancholy, endowed with black hair and brown eyes," and has a character that is "severe, conceited, and stingy. He puts on loose clothing. He is governed by opinion." The African is recognizable by the colour of his skin, by his kinky hair, and by the structure of his face. "He is sly, lazy, and neglectful. He rubs his body with oil or grease. He is governed by the arbitrary will of his masters." As for the white European, "he is changeable, clever, and inventive. He puts on tight clothing. He is governed by laws." Here mentality, clothes, political order, and physiology are all taken into account.

The French naturalist Georges Leclerc, comte de Buffon, devoted two of the 44 volumes of his *Histoire naturelle, général et particulière* (1749–1804) to man as a zoological species. Buffon criticized Linnaeus' system and all other systems of classification that depended only on external characteristics; to force individual objects into a rational set of categories was to impose an artificial construct on nature. He was echoing arguments that Locke had used, arguments based on the conception of the Great Chain of Being as a continuum, not as a sequence of discrete steps. An artificial taxonomy came from the mind, not from nature, and achieved precision at the expense of verisimilitude. Buffon's answer was to determine species not by characteristics but by their reproductive history. Two individual animals or plants are of the same species if they can produce fertile offspring. Species as so defined necessarily have a temporal dimension: a species is known only through the history of its propagation. This means that it is absurd to use the same principles for classifying living and nonliving things. Rocks do not mate and have offspring, so the taxonomy of the mineral kingdom cannot be based on the same principles as that of the animal and vegetable kingdom. Similarly, according to Buffon, there is "an infinite distance" between animal and man, for "man is a being with reason, and the animal is one without reason." Thus, "the most stupid of men can command the most intelligent of animals . . . because he has a reasoned plan, an order of actions, and a series of means by which he can force the animal to obey him." The ape, even if in its external characteristics it is similar to man, is deprived of thought and all that is distinctive of man. Ape and man differ in temperament, in gestation period, in the rearing and growth of the body, in length of life, and in all the habits that Buffon regarded as constituting the nature of a particular being. Most important, apes and other animals lack the ability to speak. This is significant in that Buffon saw the rise of human intelligence as a product of development of an articulated language. But this linguistic ability is the primary manifestation of the presence of reason and is not merely dependent on physiology. Animals lack speech not because they cannot produce articulated sound sequences, but because, lacking minds, they have no ideas to give meaning to these sounds.

The German scholar Johann Friedrich Blumenbach is recognized as the father of physical anthropology for his work *De Generis Humani Varietate Nativa* ("On the Natural Variety of Mankind"), published in 1775 or 1776. He also regarded language as an important distinguishing characteristic of man, but added that it is only man who is capable of laughing and crying. Perhaps most important is the suggestion, also made by the American statesman Benjamin Franklin, that it is only man who has hands that make him capable of fashioning tools. This was a suggestion that broke new ground in that it opened up the

possibility of speculating on a physiological origin for the development of intellectual capacities.

The thought of Kant

Man the rational subject. The great German philosopher Immanuel Kant credited Hume with having wakened him from his dogmatic slumbers. But while Kant concurred with Hume in rejecting the possibility of taking metaphysics as a philosophical starting point (dogmatic metaphysics), he did not follow him in dismissing the need for metaphysics altogether. Instead he returned to the Cartesian project of seeking to find in the structure of consciousness itself something that would point beyond it.

Thus, Kant started from the same point as the empiricists, but with Cartesian consciousness—the experience of the individual considered as a sequence of mental states. But instead of asking the empiricists' question of how it is that man acquires such concepts as number, space, or colour, he enquired into the conditions under which the conscious awareness of mental states—as states of mind and as classifiable states distinguished by what they purport to represent—is possible. The empiricist simply takes the character of the human mind—consciousness and self-consciousness—for granted as a given of human nature and then proceeds to ask questions concerning how experience, presumed to come in the form of sense perceptions, gives rise to all of man's various ideas and ways of thinking. The methods proposed for this investigation are observational, and thus the study is continuous with natural history. The enterprise overlaps with what would now be called cognitive psychology but includes introspection regarded simply as self-observation. But this clearly begs a number of questions, in particular, how the empiricist can claim knowledge of the human mind and of the character of the experience that is the supposed origin of all ideas.

Even Hume was forced to admit that self-observation, or introspection, given the supposed model of experience as a sequence of ideas and impressions, can yield nothing more than an impression of current or immediately preceding mental states. Experiential self-knowledge, on this model, is impossible. The knowing subject, by his effort to know himself, is already changing himself so that he can only know what he was, not what he is. Thus, any empirical study, whether it be of man or of the natural world, must be based on foundations that can only be provided by a nonempirical, philosophical investigation into the conditions of the possibility of the form of knowledge sought. Without this foundation an empirical study cannot achieve any unified conception of its object and never will be able to attain that systematic, theoretically organized character that is demanded of science.

Use of critical reflection

The method of such philosophical investigation is that of critical reflection—employing reason critically—not that of introspection or inner observation. It is here that the origin of what has come to be regarded as philosophical anthropology in the stricter, third sense (*i.e.*, 20th-century humanism) can be identified, since there is an insistence that studies of the knowing and moral subject must be founded in a philosophical study. But there remain questions about the humanity of Kant's subject. Kant's position was still firmly dualist; the conscious subject constitutes itself through the opposition between experience of itself as free and active (in inner sense) and of the thoroughly deterministic, mechanistic, and material world (in the passive receptivity of outer sense). The subject with which philosophy is thus concerned is finite and rational, limited by the constraint that the content of its knowledge is given in the form of sense experience rather than pure intellectual intuition. This is not a differentiated individual subject but a form of which individual minds are instantiations. The ideals regulating this subject are purely rational ideals. This tendency is even more marked in the philosophies of Fichte, Schelling, and Hegel.

Humanist thought is anthropocentric in that it places man at the centre and treats him as the point of origin. There are different ways of doing this, however, two of which are illustrated in the works of Locke and Kant, respectively. The first, realist, position assumes at the outset a contrast between an external, independently existing world and the conscious human subject. In this view man is presented as standing "outside" of the physical world that he observes. This conception endorses an instrumental view of the relation between man and the nonhuman, natural world and is therefore most frequently found to be implicit in the thought of those enthusiastic about modern technological science. Nature, from this viewpoint, exists for man, who by making increasingly accurate conjectures as to the laws governing the regular succession of natural events is able to increase his ability to predict them and so to control his environment.

The second, idealist position, argues that the world exists only in being an object of human thought; it exists only by virtue of man's conceptualization of it. In the form in which Kant expressed this position the thought that constitutes the material, physical world, is that of a transcendent mind, of which the actual minds of humans are merely vehicles.

There is also a third, dialectical, form of anthropocentrism, which, although it did not emerge fully until the 19th century, was prefigured in the works of Vico and Herder. From this standpoint the relation between man and nature is regarded as an integral part to the dynamic whole of which it is a part. The world is what it is as a result of being lived in and transformed by human beings, while people, in turn, acquire their character from their existence in a particular situation within the world. Any thought about the world is concerned with a world as lived through a subject, who is also part of the world about which he thinks. There is no possibility of transcendence in thought to some external, non-worldly standpoint. Such a position wants both to grant the independent existence of the world and to stress the active and creative role of human beings within it. It is within this relatively late form of humanism—which arose from a synthesis of elements of the Kantian position, with the insights of the Italian Giambattista Vico and the German Johann Gottfried von Herder—that philosophical anthropology in the third sense can be located.

Vico's *Scienza nuova*

Vico's *Scienza nuova* (1725; *The New Science of Giambattista Vico*) announced not so much a new science as the need to recognize a new form of scientific knowledge. He argued (against empiricists) that the study of man must differ in its method and goals from that of the natural world. This is because the nature of man is not static and unalterable; a person's own efforts to understand the world and adapt it to his needs, physical and spiritual, continuously transform that world and himself. Each individual is both the product and the support of a collective consciousness that defines a particular moment in the history of the human spirit. Each epoch interprets the sum of its traditions, norms, and values in such a way as to impose a model for behaviour on daily life as well as on the more specialized domains of morals and religion and art. Given that those who make or create something can understand it in a way in which mere observers of it cannot, it follows that if, in some sense, people make their own history, they can understand history in a way in which they cannot understand the natural world, which is only observed by them. The natural world must remain unintelligible to man; only God, as its creator, fully understands it. History, however, being concerned with human actions, is intelligible to humans. This means, moreover, that the succession of phases in the culture of a given society or people cannot be regarded as governed by mechanistic, causal laws. To be intelligible these successions must be explicable solely in terms of human, goal-directed activity. Such understanding is the product neither of sense perception nor of rational deduction but of imaginative reconstruction. Here Vico asserted that, even though a person's style of thought is a product of the phase of culture in which he participates, it is nonetheless possible for him to understand another culture and the transitions between cultural phases. He assumed that there is some underlying commonality of the needs, goals, and requirement for social organization that makes this possible.

Herder denied the existence of any such absolute and universally recognized goals. This denial carried the disturbing implication that the specific values and goals pursued by various human cultures may not only differ but also may not all be mutally compatible. Hence, not

only may cultural transitions not all be intelligible, but conflict may not be an attribute of the human condition that can be eliminated. If this is so, then the notion of a single code of precepts for the harmonious, ideal way of life, which underlies mainstream Western thought and to which—whether they know it or not—all human beings aspire, could not be sustained. There will be many ways of living, thinking, and feeling, each self-validating but not mutually compatible or comparable nor capable of being integrated into a harmonious pluralistic society.

THE 19TH CENTURY

The 19th century was a time of greatly increased activity in the sciences of man. There was a correspondingly rapid development of various disciplines, but this was accompanied by increasing specialization within disciplines. Perhaps the most significant theme, common to all branches of science, was the declining influence of religon. The philosophers of the Enlightenment had concurred in thinking that the transcendence of God doomed to failure any attempt to encompass him within the framework of human discourse. Theological discourse was thus only human discourse. Herder had stated, "It is necessary to read the Bible in a human manner, for it is a book written by men for men." Even so, he insisted, "The fact that religion is integrally human is a profound sign in recognition of its truth." But with human truth the only available truth, such a line was hard to maintain, and by the late 19th century the German philosopher Friedrich Nietzsche had announced that God was dead.

Declining influence of religion

But the death of God also meant that the essence of God in every man was dead—that which was common to all and that in virtue of which the individual transcended the natural, material world and his purely biological nature. Also dead was the part of a person that recognized universal God-given ideals of reason and truth, goodness and beauty. There thus emerged views of man that, while integrating him more thoroughly with the natural world—treating his incarnation as an essential aspect of his condition—had to come to terms with the consequences for science, morality, and the study of man himself of the removal of a transcendent support for belief in absolute standards or ideals.

The presumption of a fixed human nature was undercut at the level of natural history by the emergence and eventual acceptance of evolutionary biology. This added a historical, developmental dimension to the natural history of man, which complimented developmental views of culture and of man as a culturally constituted being. But more importantly, evolutionary biology made man a direct descendant of nonhuman primates and suggested that the gift of reason, which so many had seen as establishing a gulf between man and animal, might too have developed gradually and might indeed have a physiological basis.

Even though Buffon had tied classification to the ability to reproduce, and had thus introduced a temporal dimension into the characterization of species, he had retained the idea of stable species. But a static classification could not explain the dynamic relations between isolated species. A primitive time line of natural history thus developed. The relationship of families led to the idea of filiation between them according to an order of succession. The interpretation of fossils aroused impassioned debates. From them have arisen concepts of mutation (the process by which the genetic material of a cell is altered), transformism (the theory that one species is changed into another), and evolution. These concepts, already being formulated in the 18th century, were clarified in the work of Lamarck and Darwin.

The evolutionary theory of Charles Darwin's *Origin of Species* (1859) differed from that of Jean-Baptiste de Lamarck in that it proposed a mechanistic, nonpurposive account of evolution as the product of the natural selection of randomly produced genetic mutations (survival of the fittest). Advantageous characteristics acquired by an individual were not, as Lamarck had thought, inherited and therefore could not play a role in evolutionary development.

The theme of continuity with the rest of the natural world was one that was also to be found in the very different, antiscientific thought of Romanticism, which was one of the reactions to the rise of the doctrine of mechanism and to the Industrial Revolution for which it was held responsible. The experience of the Industrial Revolution was crucial to most 19th-century thought about man. Reactions to this experience can be put into three broad categories. There were those who saw in industrialization the progressive triumph of reason over nature, making possible the march of civilization and the moral triumph of reason over animal instinct. This was a view that continued the spirit of the Enlightenment, with its confidence in reason and the ability to advance through science. Into this category can be put the English philosopher John Stuart Mill, a stout defender of liberal individualism. Mill's philosophy was in many respects a continuation of that of Hume but with the addition of Jeremy Bentham's utilitarian view that the foundation of all morality is the principle that one should always act so as to produce the greatest happiness of the greatest number. This ethical principle gives a prominent place to the sciences of man (which are conceived as being parallel in method to the natural sciences), their study deemed necessary for an empirical determination of the social and material conditions that produce the greatest general happiness. This is a non-dialectical, naturalistic humanism, which gives primacy to the individual and stresses the importance of his freedom. For Mill, all social phenomena, and therefore ultimately all social changes, are products of the actions of individuals.

Reactions to the Industrial Revolution

The humanist opponents of capitalist industrialization fall into two groups, both presuming some form of dialectical humanism: those who, like Marx, retained a faith in the scientific application of reason and those who, like Goethe and Schiller, fundamentally questioned the humanity of mechanistic science and the technology it spawned.

The Romantics questioned the instrumental conception of the relation between man and nature, which is fundamental to the thinking behind much technological science. They insisted on an organic relation between man and the rest of nature. It is not man's place outside of nature that is emphasized but his situation within it. Equally central to this view was a recognition of the historicity of human culture and a rejection of any conception of a fixed, determined human nature on which a science of man parallel in structure to the natural sciences (*i.e.*, a science with laws, whether empirical or rational, that determine the actions and the historical development of mankind) could be based. There was a continued commitment to the perspective of the individual, and his creative relation with the world, an orientation that was carried over into the philosophical anthropology of 20th-century phenomenologists and existentialists, with their critiques of modern industrial science.

Marxist opposition to capitalism

The Marxist opposition to capitalist industrialization is not to industrialization as such but to capitalist forms of it. This opposition is founded on socialism, which stresses the role of social structures; it is at the level of society—its structures and its economic base of production—that the course of history can be understood. Marx emphasized the importance of labour and work in man's relation both to the natural and to the social worlds in which he finds himself and which condition his ability to realize himself through these relationships. He deplored the loss of humanity associated with capitalist industrialization, which was manifest in the alienating conditions under which members of the working class were treated as objects and thus deprived of their full status as human subjects by their industrial masters. Nonetheless, he retained a faith in scientific knowledge and in the possibility of a scientific understanding of history by integrating its economic, social, and political aspects. Marx argued, however, that it was not reason but revolution that would cause the overthrow of the capitalist system.

Common to all of these reactions is that whether they privileged reason or not they did not seek to validate the claims of reason—and hence the claims of science—by reference to a rational God. But with this transcendent guarantor removed, the question of the objectivity of rational standards and of the commonality of human

thought structures became pressing. The Cartesian starting point focused attention on thought as a sequence of ideas, knowable only to the individual concerned. Animals, even if capable of uttering structured sound sequences, were denied linguistic abilities on the ground that these sound sequences could not be the expressions of thoughts and could not have meaning; lacking minds, animals also lack ideas, the thoughts that give words their meaning. According to this view, words are simply conventionally established vehicles for the communication of thoughts that exist prior to, and independent of, their linguistic expression. However, if it is not assumed that human minds are all instantiations of a single transcendent mind, or that although individual they were created from a common pattern, this account of linguistic communication must appear inadequate. Since according to Cartesianism introspection is the only route to awareness of ideas, each person can only ever be aware of his own ideas, never of those of another. He could never know that his attempts to communicate succeed in calling up in another person's mind ideas similar to those in his own. Some new way of looking at linguistic communication was required, and this could be nothing short of a new starting point, a new way of thinking about thought itself.

THE 20TH CENTURY: EMERGENCE OF PHILOSOPHICAL ANTHROPOLOGY

The mood of the late 19th century, which has also dominated 20th-century philosophy, can be characterized as anti-psychologistic—a rejection of introspective, idea-oriented ways of thinking about thought, which presume that thought is prior to language. This fundamental re-orientation had implications for every other aspect of the study of man. The writers of the late 19th and early 20th centuries who most influenced subsequent philosophical thought about man were Gottlob Frege, Edmund Husserl, Ferdinand de Saussure, and Sigmund Freud. Each helped to transform one of the three reactions to the Industrial Revolution outlined above, to bring it into accord with the new, anti-psychologistic orientation: Frege influenced the empiricist, scientific reaction; Husserl the Romantic; and Saussure and Freud the scientific Socialist.

Frege and empiricist anthropology. Frege argued that if language is to be a vehicle for the expression of objective, scientific knowledge of the world, then the meaning (cognitive content) of a linguistic expression must be the same for all users of the language to which it belongs and must be determined independently of the psychological states of any individual. A word may call up a variety of ideas in the mind of an individual user, but these are not part of its meaning. Such associations may be important to the poet but are irrelevant to the scientist. The function of language in the expression of scientific knowledge is to represent an independently existing world. The meanings of linguistic expressions must thus derive from their relation to the world, not from their relation to the minds of language users. Similarly, logic—embodying the principles of reasoning and the standards of rationality—must be concerned not with laws of human thought, but with laws of truth. The principles of correct reasoning must be justified by reference to the function of language in representing the world correctly or incorrectly rather than by reference to human psychology.

Frege's work on formal logic

It is for his work on formal logic, which stemmed from these ideas, that Frege is renowned, because it opened the way for the mechanical reproduction of reasoning processes, which was crucial to the development of information processing by computers and for devices capable of artificial intelligence. Frege argued that the principles of deductive reasoning are purely formal principles, which means that their correct application does not depend on an ability to understand the sentences involved, so long as they have been put into the correct logical form. To give an account of the meaning of a sentence requires that it be analyzed so as to reveal its logical form. The logical analysis of everyday and scientific language thus becomes a primary focus of philosophical activity, hence the name "analytic philosophy" for the tradition, predominating in Great Britain, North America, and Australasia that can be regarded as post-Fregean philosophy. In this tradition the focus is on the analysis of rational, human thought, where it is presumed that the only correct way to do this is to analyze the logical structure of language.

Thus language has replaced God as the locus of rationality and of principles of reason; and the language–world relation has taken over many of the roles previously played by the God–world relation. The individual participates in a rationality that is independent of him to the extent that he is a language user. The position assumes that standards of rationality are absolute, since they are seen as necessarily governing the meaning structures of all languages. The linguist Noam Chomsky proposed a thesis that was regarded as being complimentary to this philosophical position, namely that of a universal grammar—a formal structure that underlies all languages, no matter how diverse their grammatical forms seem on the surface. Moreover, he suggested that all humans have the same innate capacity to learn language, which explains why it is that they all structure their languages, and hence their thought, in the same way.

A further assumption (christened the "principle of charity" by the American philosopher Donald Davidson) is that all humans are rational and that the majority of human behaviour is to be explained as rational, given the beliefs and desires of the person concerned. This, together with the view that language is the locus of rationality and the embodiment of thought, leads to the view that the primary objective of the sciences of man is to interpret the language of a community under study so as to attribute beliefs and desires to its members on the basis of what they say, and so give some explanation of their behaviour. The interpretation is deemed incorrect if the attributed beliefs and desires result in too much behaviour being portrayed as irrational. There will then be a mutual adjustment between language interpretation and the explanation of behaviour in which there can be no final separation of the two and no such thing as a uniquely correct interpretation. There is thus no hope of finding laws linking psychological states of belief or desire to physiological states, even though, by maintaining that each mental event is just a physical event under a different description, a dualism of mind and body is denied. What remains is an irreducible dualism between physiological and psychosocial studies of man. The situation is frequently explained by utilizing a computer analogy (for the computer is, in this view, man creating a machine in his own image). The relation between the structures of thought and the body is likened to the relation between computer software and hardware; the same hardware may be used to run different software, and the same software may be run on different hardware. The two descriptions of computer functioning are thus relatively independent.

In this account the consciousness of the individual plays little explicit role, but a model of man is nevertheless implicit in the whole approach. It is still basically the model employed by Hume, with experience consisting of sensory stimuli. Experience of other people is thus limited to observation of their physical and behavioral characteristics. It is on the basis of such observations that we have to make conjectures about their mental states. What has changed is the method of making such attributions. It is not sufficient to argue by analogy from introspection; any attribute of rational or mental faculties must go via an analytic interpretation of the language spoken. But with the assumption that all languages must share a common logical structure in virtue of their function in representing the world, there is also an inbuilt presumption of a uniformity in the rational structure of all human thought.

Husserl and philosophical anthropology. Husserl is regarded as the founder of phenomenology. He, like Frege, wished to avoid the so-called psychologism of idea-based discussions of thought and rejected naturalistic approaches to the study of the mind and of what passes for rational thought. He, too, believed that laws of reasoning needed to be validated by reference to the objects of thought, but he did not agree that logic could be made purely formal and independent of the particular subject matter in hand, nor did he agree that the primary focus should be on language.

Indeed, he rejected the position from which Frege started, namely, the assumption that there is a clear separation between the knowing subject and an independently existing reality that is the object of his knowledge. This assumption, Husserl argued, reveals a blindness to the conditions, or presuppositions, involved in all knowledge and already analyzed in part by Kant. Husserl adopted Kant's strategy but in a more radical form that was designed to restore the in-the-worldness of the human subject.

Husserl's phenomenology

The program of phenomenology aimed at rigorous understanding of the life-world. Kant had explored the conditions of the possibility of experience, and in so doing he had presumed that this experience was always that of an "I," a subject. Husserl also asked after the conditions for the possibility of a consciousness that is always potentially self-conscious. He claimed that all consciousness is intentional; *i.e.,* is consciousness of something. The method pursued was a phenomenal investigation of the "contents of consciousness." This required the investigator to "bracket off" all theories, presuppositions, and evidence of existence, including his own existence. There could be no dogmas. The implication was still that the individual can, in principle, abstract from every influence of culture and environment by abstracting also from that element of consciousness that involves awareness of self. It was presumed that consciousness as such had structures that would then be revealed. It is only self-conscious thought that is culturally constituted; for Husserl, each human individual is by necessity socially and historically conditioned by his environment. But even so it has to be doubted whether the required abstraction from self is possible in the sort of consciousness—*i.e.,* reflective rational thought—that is required of a rigorous phenomenological analysis.

Descartes and his successors had taken the self, the individual subject, for granted and in the process inevitably had assigned to the knowing subject a position outside, beyond, or transcending the world of which he sought knowledge. Husserl, by putting the individual subject into the field of philosophical investigation, paved the way for investigations of the human condition that start with the concrete, with man's being-in-the world. In this respect he can also be regarded as the founder of philosophical anthropology in the narrowest sense of the term: the personal unity of the human being becomes both the point of departure and the goal of philosophical reflection. The use of philosophical anthropology to characterize this approach emerged in the first half of the 20th century with the tendency both in Germany and in France to treat the problems of anthropology as the centre of all philosophical studies. Its emergence at this time may be seen as a reaction to the totalitarian systems of the 20th century: Italian Fascism, Soviet Communism under Stalin, and German Nazism were powerful ideologies calling for the annihilation of the individual character of the person. The philosophical protests of the German phenomenologist Max Scheler, of the Russian existentialist Nikolay Berdyayev, of the Jewish philosophical theologian Martin Buber, and of the French personalist Emmanuel Mounier offered answers to this challenge; the philosophies of the person and of existence present to each individual the means to centre himself upon himself.

Reaction to totalitarian ideologies

Work of Heidegger. Husserl's work not only gave rise to phenomenology but also to the existentialist ideas of Martin Heidegger and Jean-Paul Sartre. Heidegger adopted the method of phenomenology but rejected Husserl's refusal to allow existence to feature in the phenomenological starting point. Heidegger argued for a philosophy in which man's being-in-the-world is registered, and where this being (existence) precedes any determination of what man is (his essence).

In his *Brief über den "Humanismus"* (1947; *Brief Letter on Humanism*), Heidegger wrote:

> Are we really on the right track toward the essence of man as long as we set him off as one living creature among others in contrast to plants, beasts and God? . . . when we do this we abandon man to the essential realm of *animalitas* but attribute a specific difference to him. In principle we are still thinking of *homo animalitas*—even when *anima* (soul) is posited as *animus sive mens,* and this in turn is later posited as subject, person, or spirit (*geist*). Such positing is in the manner of metaphysics.

Naturalistic definitions of man fail, because like all traditional metaphysical definitions they naively assume that we know what we mean when we say of something that it is; *i.e.,* when we ascribe being to it.

Humanity and the world form a whole in which neither is privileged. The focus shifts from intentional objects of consciousness to the world itself, a world of objects that appear (and hence exist as individualized objects) only insofar as they have meaning and significance for human beings, by virtue of the way in which they relate to human projects. A fallen tree branch is noticed as firewood only by one who is in search of fuel. Similarly, events are noticed and recorded and so become historical events but only in the light of the meaning that they have for the historian. This means that neither history nor the study of man can be objective and purely factual history. History is always a story about the past from someone who has a specific vantage point within history.

Work of Sartre and other existentialists. Sartre, in *L'Être et le néant* (1943; *Being and Nothingness*), tried to tread a middle line between Husserl and Heidegger, retaining the concrete in-the-worldness of Heidegger while restoring a place for intentional consciousness. He hoped to provide an account of, as he put it, intentional-consciousness-in-the-world-as-it-is-lived. Sartre's driving belief was in human freedom, the ability to choose not only a course of action but also what one would become. Neither Husserl, with his already structured and regulated consciousness, nor Heidegger, with his world that is already given meaning, left enough room for freedom.

Sartre's belief in human freedom

Sartre insisted on the dualism of being (thingness) and consciousness (no-thingness) and of the individual in itself and for itself. The disjunction between these is absolute: no state of the world can determine human action, even to the extent of providing a motive, or reason, for action. If man is truly free, the world, whether material or social, can place no constraints on him, not even to the extent of determining what would or would not be good reasons for following a given course of action. He must create his own values and his own morality and take responsibility for his choices. Sartre's critics pointed out, however, that this total freedom dissolves into arbitrariness and randomness. An action that is selected at whim, chosen without (or beyond) reason, and that recognizes no rational constraints, is more an abandonment to fate than an assertion of freedom; where there is no basis for decision there is simply the necessity to choose.

In his later writings, and in particular the *Critique de la raison dialectique* (1960; *Critique of Dialectical Reason*), which attempts a reconciliation between existentialism and Marxism, Sartre came to recognize that there are constraints on the exercise of human freedom. He first acknowledged that man is a creature with biological needs, who must eat, drink, shelter, and clothe himself as a condition of being able to engage in other kinds of activity; and, second, he saw the struggle against need as conditioned by the fact that it takes place in conditions of scarcity. This means that there is competition for resources and thus the ever-present likelihood that the realization of an individual's freedom will limit that of another. Each individual in these conditions experiences others as possible threats to his own freedom (*i.e.,* he experiences alienation).

Individuals whose freedom is in this way conditioned, not just by naturally occurring material conditions but by the materiality of human practice, are (as Marx had said) both "subjects" and "objects" of history. But Sartre insisted that history is only intelligible because it records a process brought into being by human action. This rules out an understanding of history based on a "dialectic of nature," adopted by some Marxists whom Sartre criticized as being dogmatists. Sartre thus rejected the idea that there could be any naturalistic science of humanity—a science that proceeds by discovering laws without reference to the consciousness of individuals. History is neither a mere process (without a subject) nor the product of some form of social, collective "subject." But this does not mean that individuals can be treated as wholly independent units

that can be understood without taking into account the formative and conditioning role of their material and social situation. It is in this work that Sartre was still facing up to, and grappling with, the problem of the reconciliation of the demands of freedom and reason, but in an altogether more practical and concrete way than was done by his 17th-century predecessors.

Sartre's abandonment of the radical freedom of *Being and Nothingness* owed much to the French phenomenologist Maurice Merleau-Ponty's criticism of it—in *Sens et non-sens* (1948; *Sense and Non-Sense*)—as being still a dualist philosophy of consciousness and for failing to put man truly in the world. In *Sense and Non-Sense* he also expressed his view of the relation between existentialism and Marxism:

> Marx gives us an objective definition of class in terms of the effective position of individuals in the production cycle, but he tells us elsewhere that class cannot become a decisive historical force and revolutionary factor unless individuals become aware of it, adding that this awareness itself has social motives, and so on. As a historical factor, class is therefore neither a simple objective fact, nor is it, on the other hand, a simple value arbitrarily chosen by solitary consciousnesses.

In Merleau-Ponty's writing there is also a clear statement of the human presupposition that forms the basis of philosophical anthropology in this third sense:

> I am not the result or the intersection of multiple causalities that determine my body or my "psychism"; I cannot conceive of myself as nothing but a part of the world, as the simple object of biology, psychology, and sociology, nor close over myself the universe of science. Everything that I know of the world, even through science, I know from a viewpoint that is my own . . .

One effect of the insistence that it is concrete, lived experience that must form the starting point of philosophical anthropology is that not only must class and its experience enter into such accounts, but so too must sex and gender. Once the human subject, as a focus of philosophical attention, is no longer a mind whose relation to a body is at best obscure, is no longer a pure consciousness, but is essentially embodied and immersed in human culture, the biological differences between the sexes and the socially constituted role differentiation between male and female must play a part in the account of humanity.

Deuxième Sexe of de Beauvoir

In Simone de Beauvoir's *Deuxième Sexe* (1949; *The Second Sex*), she used the categories provided by Sartre to argue that to be a woman—as distinct from a man—is to be robbed of one's subjectivity, to be treated as an object by men, and to have one's conception of oneself as female defined by men. To assert her subjectivity a woman must thus negate her femininity, to reject the status of object for men that constitutes the feminine. A woman is thus placed in a condition of self-alienation, with which a man does not have to contend. In this way de Beauvoir revealed the need for a philosophy of "man" that is also a philosophy of "woman," a viewpoint that generally has been acknowledged only by female writers.

Philosophical anthropology and theology. Just as class and gender determine the way in which one lives in the world and is related to the world, so too may religion. Even for those not brought up in any religion, Western culture is still one in which religion is significant. Philosophical anthropology must thus take the phenomenon of religious experience seriously, in a way that empiricist anthropology does not. But its starting point is with the constitution of a religious consciousness, and with the conditions of the possibility of the forms of religion encountered; it does not start with theology. There is room once again for dispute over the possibility of any kind of transcendence. The 19th-century Danish philosopher Søren Kierkegaard thought that man's existence has meaning only in the experience of grace, which inexplicably raises man up from his worthlessness. The anguish and loneliness of mortal existence, the "wretchedness of man without God," is only overcome by a form of experience that confers faith in the existence of God and hence the ultimate possibility of human transcendence.

Philosophical anthropology in its narrowest (third) sense is founded on an insistence that the only knowledge available to man is knowledge from his human perspective, conditioned, as he himself is, by his situation in the world. God cannot be invoked as a source of absolute standards of truth or of absolute values nor to give content to the supposition that there are any. If God exists, then the thought that there are such standards and values, even if we cannot know of them, remains possible. This possibility was denied with Nietzsche's proclamation of the death of God; the attempt to come to terms with this view defines the scope of most philosophical anthropology. The view of religion that reflects the inversion which takes place was expressed by Ludwig Feuerbach in *Das Wesen des Christentums* (1841; *The Essence of Christianity*), when he declared that "man is not a shadow of God; it is God who is the shadow of man, an illusory phantasm that man nourishes out of his own substance."

Saussure, Freud, and antihumanism. There were also those, however, who saw the death of God as heralding the death of man as the focus and starting point for philosophy. Saussure, in his *Cours de linguistique général* (1915; *Course in General Linguistics*), held, like Frege, that the meaning of a linguistic sign, that which gives it a value for the purposes of communication, could not be an idea in the mind of an individual. But unlike Frege he did not concentrate on the relation between language and an external world. Rather, he argued that the meaning of any one linguistic sign is dependent on its relation to other signs in the language to which it belongs; thus, the meaning of one sign is determined by its place in the overall structure that constitutes a language. A consequence of this view is that language becomes a closed, autonomous system. Linguistic signs do not depend for their meaning on anything external to language. Moreover, Saussure argued that the present meaning of a word could not be revealed by tracing its etymology. It is only by reference to present language structures that current meanings are determined. The language structures that become the focus of attention are thus to be treated as autonomous from their history (*i.e.,* as if they had no history).

Thought of Saussure

With this focus on structures and the method of studying them, Saussure can be considered to be one of the founding figures of structuralism. This view of meaning came to be extended from linguistic signs to all kinds of human actions to which a conventional meaning, or significance, is attributed. It has been used as the framework for anthropological investigations of cultures, their customs, etc., as, for example, in the work of Claude Lévi-Strauss or in the interpretation of dreams and the structures of the unconscious in the works of Jacques Lacan.

It is significant that, again, meaning is studied without reference to the consciousness of individual language speakers. Man is treated as essentially not just a language speaker but as a user and interpreter of signs, and the significance of these signs is determined without reference to any relation to the individual. A language, or sign-system, takes over the role of providing the framework of reason in which significance is given, but this framework transcends the individual. Such systems of codification regulate all human experience and activity and yet lie beyond the control of either individual or social groups. Indeed, since there is no meaning or understanding outside of a given sign-system, it is only from the meaning of the signs he "uses" that the individual comes to learn what it is that he means by his action, and hence what he thinks. This is why such views of language can readily be grafted onto Freud's theory of the unconscious.

Freud's theory of the unconscious

Freud treated the realm of the mind as one that is as lawgoverned as is the natural world; nothing that a person does or says is haphazard or accidental, for everything can in principle be traced to causes that are somehow in the person's mind, although many of these are not accessible to consciousness. Freud's view of the human mind is thus very different from Descartes's. For Freud, the part of the mind that is accessible to consciousness is but the tip of a large iceberg; the hidden remainder, which influences the conscious, is the unconscious. Thus, for instance, there are unconscious desires that can cause someone to do things that he cannot explain rationally to others, or even to himself. In this there is a parallel between Freud and

Marx, for both hold views on which human consciousness, far from being perfectly free and rational, is really determined by causes of which man is not aware; but whereas Marx says that these causes are social and economic in nature, Freud claims that they are individual and mental. In both cases the implications for the study of man are anti-psychologistic in that they turn attention away from the individual consciousness. On both views a scientific understanding of man is only to be gained by examining the factors that determine consciousness rather than the level of the individual subject of consciousness.

The id, ego, and superego

In his later expositions (those given in the 1920s) Freud assigned to the mind a tripartite structure: the id, which contains all the instinctual drives seeking immediate satisfaction; the ego, which deals with the world outside the person, mediating between it and the id; and the superego, a special part of the ego that contains the conscience, the social norms acquired in childhood. Whatever can become conscious is in the ego, although even in it there may be things that remain unconscious, whereas everything in the id is permanently unconscious. The instincts or drives contained in the id are the motivating forces in the mental apparatus, and all of the energy of the mind comes from them. Freud included a sexual instinct as one of the basic instincts and thus gave sexuality a much wider scope in psychology and in the study of man than had previously been the case. Freud's account of individual human character is also developmental. He held that particular "traumatic" experiences, although apparently forgotten, could continue to exercise a harmful influence on a person's mental health. The fully fledged theory of psychoanalysis generalizes from this and asserts the crucial importance, for the adult character, of the experiences of infancy and early childhood. Freud also held that the first five or so years of life are the time in which the basis of an individual's personality is laid down; one cannot fully understand a person, therefore, until he comes to know the psychologically crucial facts about that person's early childhood. Freud produced detailed theories of the stages of development that are concerned specifically with the development of sexuality, in which this concept is widened to include any kind of pleasure obtained from parts of the body. Freud's view was that individual well-being, or mental health, depends on a harmonious relationship between the various parts of the mind and between the person and the real world in which he must live. Neurosis results from the frustration of basic instincts, either because of external obstacles or because of internal mental imbalance. The work of the analyst is to interpret the behaviour and speech of a patient in such a way as to give insight into the unconscious, to be able to explain what is inexplicable at the conscious level, and in this way to try to give the patient an understanding of himself. Here there is a need for a theory of signs and of interpretation in which its notion of meaning, or significance, does not rest on either reference to the physical world or on ideas in an individual consciousness; structuralist theories provide one such possibility.

From the point of view of either Freudian theory or of the non-existentialist reading of Marx, any attempt to provide a study of man—of human behaviour and history—that starts from the individual consciousness must seem misguided. This will include the empiricist approaches, which assume that all human behaviour is to be explained in terms of the conscious mental states (*i.e.*, beliefs and desires) of individuals. Such approaches seem to fail to acknowledge that the causes of human actions include factors of which they are not consciously aware. A scientific account, one that is concerned with providing causal explanations, must not be confined to the subjectivity of the individual consciousness but must adopt an objective standpoint, from which these factors can be recognized and studied. Equally as important, however, the sort of arguments used by phenomenologists and existentialists to query the availability of objective viewpoints can be reapplied here. Thus, structuralism gives place to the post-structuralism of Derrida and Deleuze, according to which neither a scientific nor a philosophical anthropology is possible.

BIBLIOGRAPHY. General works include LESLIE STEVENSON, *Seven Theories of Human Nature* (1974), which gives short introductory sketches of the views of Plato, Christian philosophers, Marx, Freud, Skinner, Sartre, and Lorenz; BERNARD GROETHUYSEN, *Anthropologie philosophique* (1953, reprinted 1980), a series of historical sketches of human personality from antiquity to the Renaissance; MICHAEL LANDMANN, *De Homine: Man in the Mirror of His Thought* (1979; originally published in German, 1962), philosophical rather than anthropological; J.S. SLOTKIN (ed.), *Readings in Early Anthropology* (1965), a good selection of important historical texts, mainly of Anglo-Saxon thinkers; GEORGES GUSDORF, *Les Sciences humaines et la pensée occidentale* (1966–), a general history of the sciences of man on the basis of an anthropological philosophy—12 of 13 vol. had appeared to 1986; ERNST CASSIRER, *An Essay on Man: An Introduction to the Philosophy of Human Culture* (1944, reprinted 1974), a useful and accurate sketch; and A.L. KROEBER (ed.), *Anthropology Today: An Encyclopedic Inventory* (1953, reissued 1965).

On the history of the philosophy of man in the Western tradition, see PRUDENCE ALLEN, *The Concept of Woman: The Aristotelian Revolution, 750 BC–AD 1250* (1985), which contains an excellent bibliography and numerous quotations from historical sources on human nature and the relation between male and female. Insights into the spirit of Renaissance thinking about man are given by ERNST CASSIRER, *The Individual and the Cosmos in Renaissance Philosophy*, trans. by MARIO DOMANDI (1963, reissued 1972; originally published in German, 1927); and DOROTHY KOENIGSBERGER, *Renaissance Man and Creative Thinking: A History of Concepts of Harmony, 1400–1700* (1979). The problems faced by 17th-century philosophers are outlined in LEROY E. LOEMKER, *Struggle for Synthesis: The Seventeenth Century Background of Leibniz's Synthesis of Order and Freedom* (1972). ERNST CASSIRER, *The Philosophy of the Enlightenment*, trans. by FRITZ C.A. KOELLN and JAMES P. PETTEGROVE (1951, reissued 1979; originally published in German, 1932), provides a general discussion of the Enlightenment. On Hegel's philosophy and its impact, see CHARLES TAYLOR, *Hegel and Modern Society* (1979). A useful comparison of two very different 19th-century views of man and society is provided by GRAEME DUNCAN, *Marx and Mill* (1973, reprinted 1977). CHARLES COULSTON GILLISPIE, *Genesis and Geology: A Study in the Relations of Scientific Thought, Natural Theology, and Social Opinion in Great Britain, 1790–1850* (1951, reissued 1959), discusses the impact of science on religious conceptions of man and his place in the order of nature in the decades before Darwin; and MARY MIDGLEY, *Beast and Man; The Roots of Human Nature* (1978, reissued 1980).

Post-Fregean, analytic philosophical thinking about man is conveyed in SAMUEL GUTTENPLAN (ed.), *Mind and Language* (1975, reprinted 1977); AMÉLIE OKSENBERG RORTY (ed.), *The Identities of Persons* (1976); JOHN SEARLE, *Minds, Brains, and Science* (1984); and DONALD DAVIDSON, *Essays on Actions and Events* (1980). The controversy over whether linguistic ability is a distinctively human trait is discussed in EUGENE LINDEN, *Apes, Men and Language* (1975, reprinted 1981); and NOAM CHOMSKY's *Language and Mind*, enl. ed. (1972).

Post-Hegelian philosophy that constitutes philosophical anthropology in the strict, third sense, together with reactions against it, is discussed in KATE SOPER, *Humanism and Anti-Humanism* (1986); and MARK POSTER, *Existential Marxism in Postwar France* (1975, reprinted 1977). Works with the orientation characteristic of philosophical anthropology in this sense include HANNAH ARENDT, *The Human Condition* (1958, reprinted 1974); EDMUND HUSSERL, *Ideas: General Introduction to Pure Phenomenology*, trans. by W.R. BOYCE GIBSON (1931, reissued 1972; originally published in German, 1913); MARTIN HEIDEGGER, *Being and Time*, trans. by JOHN MACQUARRIE and EDWARD ROBINSON (1962, reissued 1973; originally published in German, 7th ed., 1953); JEAN-PAUL SARTRE, *Being and Nothingness*, trans. by HAZEL E. BARNES (1956, reissued 1978; originally published in French, 1943), and *Critique of Dialectical Reason*, trans. by ALAN SHERIDAN-SMITH (1976, reissued 1982; originally published in French, 1960); M. MERLEAU-PONTY, *Phenomenology of Perception*, trans. by COLIN SMITH (1962, reprinted 1981; originally published in French, 1945); and SIMONE DE BEAUVOIR, *The Second Sex*, trans. by H.M. PARSHLEY (1953, reprinted 1983; originally published in French, 2 vol., 1949).

Opposition to this orientation can be found in, among others, CLAUDE LÉVI-STRAUSS, *The Savage Mind* (1966, reissued 1972; originally published in French, 1962), and *Structural Anthropology*, 2 vol. (1963–76; originally published in French, 1958–73); LOUIS ALTHUSSER, *For Marx*, trans. by BEN BREWSTER (1969, reissued 1979; originally published in French, 1965); and JACQUES LACAN, *The Four Fundamental Concepts of Psycho-Analysis*, trans. by ALAN SHERIDAN (1977, reissued 1981; originally published in French, 1973).

(G.P.G./M.E.T.)

Western Philosophical Schools and Doctrines

It is the purpose of this article to examine in detail major schools of Western philosophy, ancient, medieval, and modern. Several of the ancient schools presented here developed around the philosophical systems of eponymous "founders." In the cases of Platonism and Aristotelianism, the treatments herein exclude the doctrines of Plato and Aristotle themselves, focussing rather on the survival and metamorphosis of these doctrines in subsequent thought; the original Platonic and Aristotelian systems are treated in full, along with the lives of the respective philosophers, in the *Macropædia* articles PLATONISM and ARISTOTELIANISM. While the philosophy of Thomas Aquinas is treated briefly below in the context of medieval Scholasticism, full treatment appears with the biography of Aquinas in the article THOMISM. The central systems of the major modern philosophers are likewise treated in conjunction with *Macropædia* biographies (namely, CARTESIANISM; HEGELIANISM; KANTIANISM; MARXISM), and the discussion of their philosophies is, therefore, excluded from the present section on modern schools of thought. Contemporary schools and doctrines are treated in the *Macropædia* article PHILOSOPHY, THE HISTORY OF WESTERN.

It should be noted that several of the schools here classified as ancient underwent revivals or reformulations in later periods (*e.g.*, Platonism, Skepticism); further, it could be argued that no philosophical school regarded as modern is without its precedent in the ancient tradition. The present material, therefore, is drawn together to facilitate comparing and contrasting what might be called the careers of these ideas. (Ed.)

This article is divided into the following sections:

Ancient and medieval schools 563
- Aristotelianism 563
  - The Hellenistic Age and Neoplatonism
  - Aristotelianism in Arabic philosophy
  - Aristotelianism in Jewish philosophy
  - The Christian East
  - The Christian West
  - Conclusion
- Atomism 565
  - The basic nature of Atomism
  - Various senses of Atomism
  - Diverse philosophical characterizations of Atomism
  - History and major representatives of the various Atomisms
  - Foundational issues posed by Atomism
- Eleaticism 571
  - The Eleatic school vis-à-vis rival movements
  - The rigorous ontologism of Parmenides and Melissus
  - The paradoxes of Zeno
  - The decline of Eleaticism
- Epicureanism 573
  - The nature of Epicureanism
  - History of Epicureanism
  - Criticism and evaluation
- Platonism 576
  - Aristotle's account of Platonism
  - The Academy after Plato: the rise of Neoplatonism
  - Influence of Platonism on Christian thought
- Pythagoreanism 578
  - General features of Pythagoreanism
  - Major concerns and teachings
  - History of Pythagoreanism
  - Evaluation
- Realism 581
  - Nature and scope of Realism
  - History of Western Realism
  - Major issues and evaluation of Realism
- Scholasticism 585
  - Nature and significance
  - History and issues
  - Enduring features
- Skepticism 590
  - Various senses and applications
  - Ancient Skepticism
  - Medieval Skepticism
  - Modern Skepticism
  - Criticism and evaluation
- Sophists 593
  - History of the name
  - The 5th-century Sophists
  - Nature of Sophistic thought
  - Writings
  - Particular doctrines
  - The Second Sophistic movement
- Stoicism 596
  - Nature and scope of Stoicism
  - Ancient Stoicism
  - Stoicism in medieval and modern philosophy

Modern schools 600
- Analytic and Linguistic philosophy 600
  - General viewpoint of Analytic philosophy
  - Early history of Analytic philosophy
  - Later history of the movement
- Empiricism 608
  - Various meanings of Empiricism
  - History of Empiricism
  - Criticism and evaluation
- Existentialism 611
  - Nature of Existentialist thought and manner
  - Historical survey of Existentialism
  - Methodological issues in Existentialism
  - Substantive issues in Existentialism
  - Social and historical projections of Existentialism
- Idealism 617
  - Approaches to understanding Idealism
  - Types of philosophical Idealism
  - Criticism and appraisal
- Materialism 621
  - Types of Materialist theory
  - History of Materialism
  - Substantive issues in Materialism
- Phenomenology 625
  - Characteristics of Phenomenology
  - Origin and development of Husserl's Phenomenology
  - Later developments
  - Dissemination of Phenomenology
  - Conclusion
- Positivism and Logical Empiricism 630
  - The Social Positivism of Comte and Mill
  - The Critical Positivism of Mach and Avenarius
  - Logical Positivism and Logical Empiricism
  - Current criticisms and controversies
- Pragmatism 636
  - Major theses of philosophic Pragmatism
  - History of Pragmatism
  - Evaluation of Pragmatism
- Rationalism 640
  - Types and expressions of Rationalism
  - History of Rationalism
  - Status of Rationalism
- Utilitarianism 644
  - The nature of Utilitarianism
  - Historical survey
  - Effects of Utilitarianism in other fields
  - Summary and evaluation

Bibliography 646

ANCIENT AND MEDIEVAL SCHOOLS

Aristotelianism

Aristotle's work, constituting the ancient world's greatest encyclopaedia, has exerted an immense influence over the succeeding centuries. It is proposed in the context of the present section to trace the course of the several streams of thought which had their source in Aristotle.

THE HELLENISTIC AGE AND NEOPLATONISM

The school founded by Aristotle in the Lyceum in Athens in 335 BC long survived his death. Its members became known as the Peripatetics. Aristotle's immediate disciples, Theophrastus of Eresus and Eudemus of Rhodes, devoted themselves to maintaining and to developing his teaching without altering either its content or its spirit; but after them the school fell rapidly into a decline as far as philosophy was concerned, and thenceforward until the middle or later decades of the 1st century BC no one taught as Aristotle had done. Then at last Andronicus of Rhodes made it his business to bring to light the long-sequestered treatises of Aristotle, to classify them according to their subject matter, and to publish them. His edition started a revival of interest in Aristotelian philosophy, and numerous commentaries on these texts were produced in the last centuries of the Hellenistic Age.

Andronicus of Rhodes

Andronicus himself interpreted a series of Aristotle's treatises, especially those of the *Organon,* and his example was followed. One of the most important commentators was Alexander of Aphrodisias, who taught in Athens from AD 198 to AD 211 and was known as the Second Aristotle because of the clarity of his exposition. Commentaries by him on part of the *Organon,* on *Metaphysics* I–IV, on the *Meteorologica,* and on the treatise *On Sensation and the Sensible* are extant; and in an original work of his, *On the Soul,* he gives a materialist interpretation of Aristotle's psychology of man, at the same time identifying the "active intellect" with God.

After Alexander, the Peripatetic school was absorbed by Neoplatonism, under which Platonic doctrines were resuscitated amid strong currents of Aristotelian influence. The last Greeks of the ancient world to write commentaries on Aristotle were all Neoplatonists, as follows:

Porphyry (234–*c.* 305), a pupil of Plotinus, wrote a very important *Isagoge,* or introduction, to the *Categories,* as well as a commentary on that treatise.

Themistius (*c.* 320–390), who taught in Constantinople, left commentaries on several works of Aristotle's, notably on the treatise *On the Soul.*

Ammonius Hermiae, who after studying under Proclus in Athens was head of the school of Alexandria toward the end of the 5th century, lectured on several of Aristotle's treatises: transcripts are extant of notes taken down at his lectures on the *Categories,* on the treatise *On Interpretation,* on the *Prior Analytics,* and also on Porphyry's *Isagoge.*

Simplicius, who was a pupil of Ammonius in Alexandria, wrote ample commentaries on the *Categories,* on the *Physics,* and on the treatise *On the Heavens.*

Finally John Philoponus, another pupil of Ammonius but a member of the Christian community in Alexandria, wrote against certain doctrinal errors that he detected in Aristotle: fragments of his treatise *Against Aristotle* are extant; so is most of his book *De aeternitate mundi,* which attacks the thesis of the eternity of the world as elaborated by the pagan Neoplatonist Proclus; and a series of Aristotelian commentaries is ascribed to him, some of which consist, however, of notes taken down at Ammonius' lectures and simply filled out by John Philoponus.

To sum up, Aristotle's philosophy can hardly be said to have been maintained in its entirety among the Greeks of the ancient world after the first generation of his disciples. Andronicus launched a revival in the 1st century BC, but from the 4th century AD onward Aristotelianism was submerged in Neoplatonism, which accommodated to its own peculiar view of the universe whatever Aristotelian doctrine it cared to take up.

ARISTOTELIANISM IN ARABIC PHILOSOPHY

Aristotelianism was to have a highly distinguished history in the world of Islām; but the Arabic philosophers, who owed their first acquaintance with it to the Neoplatonists' commentaries, never presented it in its purity or disengaged it from the Neoplatonic context in which it had been transmitted to them. This is readily understandable: whereas Aristotle's own metaphysic was too imperfect to satisfy Islāmic monotheism, the Neoplatonist metaphysic of Plotinus supplied an invaluable complement, to which Muslim thinkers always, to a greater or lesser extent, had recourse. The tendency to combine Aristotelianism proper with Neoplatonism was moreover strengthened by the diffusion of a work known as the *Theology of Aristotle:* this, originally compiled in Syriac by a Christian monk availing himself of extracts from the *Enneads* of Plotinus, was translated into Arabic *c.* 840 and was commonly ascribed to Aristotle himself, with the result that the latter came to be credited with metaphysical doctrines characteristic of Plotinus. Much later, probably in the 12th century, another pseudo-Aristotelian work—of Arabic origin this time—was circulated in Spain, namely the work well known from its prompt translation into Latin as the *Liber de causis* ("On Causes"): a commentary on propositions selected from Proclus' *Elements of Theology,* it was monotheistic in inspiration and served further to confirm the habit of attributing to Aristotle creationist doctrines wholly foreign to him.

The combination of Aristotelianism with Neoplatonism is already realized in the writings of the first Arabic philosopher, al-Kindī, who flourished in 9th-century Baghdad. A century later, al-Fārābī (died *c.* 950), who likewise taught mainly in Baghdad, similarly linked Aristotle's doctrines with the metaphysic of the last Alexandrian Neoplatonists, stressing however the independence of philosophy from religion.

The first major thinker of Islām was the Iranian philosopher and physician Avicenna (properly Ibn Sīnā, 980–1037). Besides his personal writings, many of which are lost, he produced a great encyclopaedia of philosophy, *Kitāb ash-shifā* ("Book of Healing"), which consists largely of a paraphrase of Aristotelian writings but is capped with an emanationist metaphysic derived from the so-called *Theology of Aristotle* and from other Neoplatonic sources.

Avicenna

Aristotle's influence in the Arabic world reached its zenith with the work of Averroës (properly Ibn Rushd, 1126–98), of Córdoba in Andalusia, who professed a boundless admiration for the Greek master and regarded him as sent by God to teach men true philosophy. Often in reaction against Avicenna, Averroës meant to restore Aristotelianism in its integrity and composed three series of commentaries on Aristotle's treatises: (1) the "little commentaries," short compendiums or epitomes providing a brief analysis of the treatises; (2) the "middle commentaries," explaining the texts literally; and (3) the "great commentaries," a more advanced and more profound literal exegesis. While he is faithful, on the whole, to Aristotle's thought, Averroës nonetheless, perhaps unwittingly, gives it an undue extension by endowing the Aristotelian "prime mover" with the characteristics of the Plotinian and Islāmic transcendent God, the universal First Cause. Furthermore, he often enough supplements Aristotle by advancing his own interpretation of obscure passages or by developing doctrines that Aristotle scarcely considered at all. The typical instance is where he propounds his own "monopsychism": forcing Aristotle to follow his metaphysical principles to their logical conclusion, Averroës maintains that the two human intellects, namely the passive or receptive intellect and the active intellect, being immaterial, cannot be multiplied in individuals; that consequently both are single substances, entering by their own operation into relation with human individuals, as the passive intellect thinks by means of the ideas that the active intellect abstracts from images in the human brain; and that the human individual is only a superior

kind of animal, altogether mortal. If Averroës won few followers among the Arabs, his interpretation of Aristotle and, particularly, his monopsychism were taken up with great interest by Jewish and even more so by Christian thinkers.

ARISTOTELIANISM IN JEWISH PHILOSOPHY

Jewish speculative thought in the Muslim world was long dominated by Neoplatonism, but all Jewish philosophers from the time of Isaac Israeli (9th–10th century) to the end of the Middle Ages were subject, more or less, to the Aristotelian influence. A firmer orientation toward Aristotelianism is discernible in the famous *Guide of the Perplexed* of Moses Maimonides (1135–1204), a contemporary of Averroës and, like him, a native of Córdoba. But here again there is no occasion to speak of pure Aristotelianism: Maimonides adopts a large measure of Neoplatonic theology (emanation and the *via negativa* or "negative way" to knowledge of God), adopts also much of the Plotinian system of ethics, and furthermore differs from Aristotle on the question of the eternity of the world. Jewish philosophy in Spain, in France, and in Italy was influenced by Maimonides throughout the remainder of the Middle Ages and even longer.

Maimonides' *Guide*

A Jewish opposition to Aristotelianism had already manifested itself most distinctly in the first decades of the 12th century, when Judah ha-Levi denounced the current philosophy. After Maimonides this anti-Aristotelian reaction persisted, notably among the Kabbalists.

THE CHRISTIAN EAST

During the patristic period, some Aristotelian doctrines infiltrated Eastern Christianity through Neoplatonic channels: they can be shown to have affected St. Gregory of Nyssa and Nemesius of Emesa in the 4th century; Pseudo-Dionysius the Areopagite and John Philoponus (whose work as a commentator has already been mentioned) in the 6th; and St. John of Damascus in the 8th. In the 9th century the Byzantine patriarch Photius and his disciple Arethas took an interest in Aristotelian logic, and from the 11th century onward there was a major revival of Aristotelian studies in Constantinople, exemplified particularly by Michael of Ephesus (late 11th century) and by Eustratius of Nicaea (*c.* 1050–1120), who both wrote commentaries on parts of the *Organon* and on the *Nicomachean Ethics.* Further commentaries appeared in the 13th and 14th centuries. In the 15th century Cardinal Bessarion argued for the ultimate concordance of Aristotelianism with Platonism.

THE CHRISTIAN WEST

Aristotle's influence during the patristic period was even more restricted in Western Christianity than in Eastern: St. Augustine of Hippo, for instance, was acquainted with nothing Aristotelian except the *Categories.*

Boethius

The fountainhead of Aristotelianism in the Christian West was the Roman philosopher Boethius (*c.* 480–524). With the intention of demonstrating the profound harmony between Plato and Aristotle, he set himself to translate the works of those two great masters into Latin. He is known to have succeeded at least in translating all the treatises of the *Organon* except the *Analytica posteriora,* together with Prophyry's *Isagoge;* and he also produced commentaries on the *Categories,* on the treatise *On Interpretation,* and on the *Isagoge.* Though his achievement fell short of his project, his work was of capital importance for the transmission of Greek philosophical thought and of Aristotelian logic to Latin Christendom: the *Categories,* the treatise *On Interpretation,* and the *Isagoge* constituted the principal textbooks of logic for the early Middle Ages and so came, later, to be known as the *Ars vetus* or *Logica vetus* ("old technique" or "old logic") when the other parts of the *Organon,* known as the *Ars nova* or *Logica nova,* had been rediscovered and published in Latin versions. Boethius therefore well deserves to be remembered as "the Teacher of the West," particularly because of the essential role played by Aristotelian logic in the intellectual formation of the new peoples who arrived in western Europe as barbarian invaders and remained there to develop a civilization replacing that of the Roman Empire. With his works on logic Boethius paved the way for the elaboration of the scholastic method, and in his commentary on Porphyry's *Isagoge* he posed the problem of universals, which was to figure so prominently in the controversies of a later age.

Charlemagne's educational policy confirmed the status of Aristotelian logic in the scholastic curriculum during the literary and scientific revival of the Carolingian Renaissance: "dialectic," or logic, was in fact the only philosophic discipline to be admitted among the seven "liberal arts" that represented secular, or "profane," learning in the program of teaching authorized by Charlemagne's capitulary of the year 778. Subsequently, between the 9th and the 12th centuries, more and more work on logic was undertaken; Aristotle's influence grew continuously as the nature of knowledge was discussed; and with Peter Abelard (1079–1142) the moderate realism of the Aristotelians was vindicated against the extreme realism of the Platonists. With Abelard likewise, and especially in his book *Sic et non,* the scholastic method was perfected. The scholastic method, a product of Aristotelian logic, contributed much not only to the development of speculative theology but also to the progress of the deductive sciences and to the grammatical organization of the European languages.

The triumph of Aristotelianism in the epistemology and in the logic of the 12th-century scholastics prepared the ground for the Aristotelian domination of the universities in the 13th century. The middle of the 12th century had seen the start of a massive penetration of Aristotle's works into western Europe, in Latin translations first from Arabic versions, then from Greek texts. Most of Aristotle's treatises were known by the beginning of the 13th century, but it was only gradually, in the course of the next 100 years, that the consequences of this flood of pagan philosophy became clear (Aristotle's works were accompanied by those of other pre-Christian Greeks, as well as by the commentaries of the Muslims). In the universities the "faculties" of arts (successors of the earlier schools of the liberal arts) enlarged their curricula from the start of the 13th century, and Aristotelianism became more and more firmly implanted, both at Paris and at Oxford, despite opposition from some of the ecclesiastical authorities. The Paris faculty of arts decided on March 19, 1255, that its students should attend lectures on every known treatise of Aristotle's, and indeed by that time every faculty of arts was turning into a faculty of philosophy teaching Aristotelianism. Already, moreover, in the 1220s, Aristotelianism had broken into the faculties of theology; and thenceforward until the end of the Middle Ages (or even later in some establishments) it was to remain fundamental to the structure of scholasticism, both philosophically and theologically. Of all the Aristotelian revivals, the most dynamic was that which the 13th century witnessed in the Christian West.

Aristotelianism in the medieval universities

The revived Aristotelianism of the Christian Middle Ages was no purer, however, than that of the Arabs or that of the Jews had been: various complementary or corrective elements were always present, whether religious or philosophical. Philosophically the main influence was derived, once again, from Neoplatonism.

Up to 1250, Latin Aristotelianism remained very eclectic and, for the most part, Avicennian, so that Aristotle's doctrine, albeit preponderant, was compounded with secondary importations. Avicenna's paraphrases were found useful for the interpretation of difficult texts; the Jewish Neoplatonist Ibn Gabirol (Avicebron) was highly esteemed; Proclus was also available; and in the 1230s the work of Averroës became known to Christian Europe. For theology, Aristotelianism was combined with traditional doctrines (derived mainly from St. Augustine and from the Pseudo-Dionysius) or with the teaching of the 12th-century masters.

After 1250, the Aristotelian influence becomes perceptibly stronger, though at the same time it branches out in various directions. The several schools of thought can be distinguished from one another by the differences in their attitude toward Aristotle, but all remain basically Aristotelian (the Augustinian school in the strict sense of

the name did not come into existence until *c.* 1270, as a reaction against heterodox Aristotelianism and Thomism). Thus the Aristotelianism of St. Bonaventura is of an Augustinian tendency; that of St. Albertus Magnus is more Neoplatonic, being strongly affected by Proclus, by the Pseudo-Dionysius, and by Avicenna; that of St. Thomas Aquinas is so profoundly recogitated as to be converted into a distinct system, Thomism; and that of Siger of Brabant is heterodox, as it accepts doctrines imcompatible with Christianity.

The great doctrinal controversies of the 13th century were largely disputes between champions of the various sorts of Aristotelianism. Thus, when St. Thomas Aquinas held his ground against the majority of the Paris theologians (*c.* 1270), the conflict was not so much between Aristotelianism and Augustinianism as between the eclectic Aristotelianism for which Alexander of Hales and William of Auvergne had stood and the more consistent and vigorous Aristotelianism which Thomas was maintaining. Similarly the conflict between Thomas and Siger can be regarded as one between the Christian and the heterodox or pagan varieties of Aristotelianism.

From the end of the 13th century Aristotelianism in philosophy was upheld chiefly by the logicians, metaphysicians, psychologists, and ethical theorists teaching in the faculties of arts and was usually "moderate"; *i.e.,* orthodox with respect to Christian doctrines. Thomism became the established system of the Dominicans and even won adherents outside their order. The Neo-Augustinianism that had come into being in the reaction against the nascent Thomism found its definitive expression in Scotism, which in fact was marked by a reversion to Aristotelianism in certain fields and remained in many respects dependent on Aristotle. Finally an Averroist Aristotelianism was launched in Paris by John of Jandun in the first quarter of the 14th century and was taken up in Italy by Taddeo da Parma and by Angelo d'Arezzo. This Latin Averroism was still flourishing in Italy in the 16th century, though it was opposed alike by the Platonism of the humanist Renaissance and by the rival Aristotelianism of the Alexandrists, who revived the doctrines of Alexander of Aphrodisias to interpret Aristotle's psychology.

The 14th century, however, saw also new currents of thought running counter to the influence of Aristotelianism. On the one hand, the Aristotelian system of physics was challenged both at Paris and at Oxford; on the other hand, moderate realism was battered successively by Nominalism, by phenomenalism, by Skepticism, and by agnosticism, which, to a greater or lesser degree, questioned the validity of knowledge and, in particular, the possibility of metaphysics. Failing to disengage itself soon enough from the obsolete physics, Scholasticism was brought farther and farther into disrepute by the successes of the new. One of the reproaches that the men of the humanist Renaissance cast most frequently at the Scholastics was that of being excessively obsequious to Aristotle. Nevertheless, Aristotelianism was still the standard doctrine of some universities down to the end of the 18th century and even longer. In the 19th and 20th centuries there was a great revival of Aristotelian studies, most notably in England (Oxford), in Germany, in France, and in Belgium (Louvain). For some scholars, the interest was chiefly historical: they saw Aristotle as one of the most brilliant products of Greek culture and, indeed, of human culture in general. For the promoters of the Thomist revival, on the other hand, the interest in Aristotelianism was essentially doctrinal, St. Thomas Aquinas having taken him as the main source of his philosophy.

Modern revivals

CONCLUSION

Very different judgments have been passed on Aristotelianism in the course of history, and its value as a philosophy or as an instrument of theological speculation is still debated. Defenders of religion, whether Jewish, Muslim, or Christian, have often denounced Aristotelianism as tending toward empiricism, as defective in its metaphysic, and as limited to earthly life in its ethic, and they have accused its followers of being naturalists or rationalists. Platonists and idealists also object to Aristotelianism on the grounds that it is empiricist, that it gives a central place to natural philosophy in its scheme of things, and that it makes excessive use of discursive reason and of abstract concepts. Finally, many modern thinkers regard the majority of Aristotle's philosophical categories as out of date. On the contrary side, Thomists maintain that many of the notions advanced by Aristotle, not only in metaphysics but also in physics, in psychology, and in ethics, are still really valuable, quite independently of the evident fact that his "science" is now altogether obsolete.

(F.V.Sn.)

Atomism

In the broadest sense the term Atomism refers to any doctrine that explains complex phenomena in terms of aggregates of fixed particles or units. This philosophy has found its most successful application in natural science: according to the atomistic view, the material universe is composed of minute particles, which are considered to be relatively simple and immutable and too small to be visible. The multiplicity of visible forms in nature, then, is based upon differences in these particles and in their configurations; hence any observable changes must be reduced to changes in these configurations.

THE BASIC NATURE OF ATOMISM

Atomism is in essence an analytical doctrine. It regards observable forms in nature not as intrinsic wholes but as aggregates. In contrast to holistic theories, which explain the parts in terms of qualities displayed by the whole, Atomism explains the observable properties of the whole by those of its components and of their configurations.

Core concepts

In order to understand the historical development of Atomism and, especially, its relation to modern atomic theory, it is necessary to distinguish between Atomism in the strict sense and other forms of Atomism. Atomism in the strict sense is characterized by three points: the atoms are absolutely indivisible, qualitatively identical (*i.e.,* distinct only in shape, size, and motion), and combinable with each other only by juxtaposition. Other forms of Atomism are less strict on these points.

Atomism is usually associated with a "realistic" and mechanistic view of the world. It is realistic in that atoms are not considered as subjective constructs of the mind employed for the sake of getting a better grip upon the phenomena to be explained; instead, atoms exist in actual reality. By the same token, the mechanistic view of things, which holds that all observable changes can be reduced to changes of configuration, is not merely a matter of employing a useful explanatory model; the mechanistic thesis holds, instead, that all observable changes are caused by motions of the atoms. Finally, as an analytic doctrine Atomism is opposed to organismic doctrines, which teach that the nature of a whole cannot be discovered by dividing it into its component parts and studying each part by itself.

VARIOUS SENSES OF ATOMISM

The term Atomism is derived from the Greek word *atoma*—"things that cannot be cut or divided."

Two basic types of Atomism. The history of Atomism can be divided into two more or less distinct periods, one philosophical and the other scientific, with a transition period between them (from the 17th to the 19th century). This historical fact justifies the distinction between philosophical and scientific Atomism.

Philosophical Atomism. In philosophical Atomism, which is as old as Greek philosophy, attention was focussed not on the detailed explanation of all kinds of concrete phenomena but on some basic general aspects of these phenomena and on the general lines according to which a rational explanation of these aspects was possible. These basic aspects were the existence in nature of a manifold of different forms and of continuous change. In what way could these features be explained? Philosophical Atomism offered a general answer to that question. It did not, however, strictly confine itself to the general problem of explaining the possibility of change and multiplicity—not even in

Forces and change

ancient Greek Atomism, for in Greek thought philosophy and science still formed a unity. Consequently, Atomists also tried to give more detailed explanations of concrete phenomena, such as evaporation, though these explanations were meant more to endorse the general doctrine of Atomism than to establish a physical theory in the modern sense of the word. Such a theory was not yet possible, because a physical theory must be based upon indirect or direct information about the concrete properties of the atoms involved, and such information was not then available.

Modern atomic theory. With the development of a scientific atomic theory, the general philosophical problems gradually disappear into the background. All attention is focussed on the explanation of concrete phenomena. The properties of the atoms are determined in direct relationship with the phenomena to be explained. For this reason the chemical atomic theory of the 19th century supposes that each identified chemical element has its own specific atoms and that each chemical compound has its own molecules (fixed combinations of atoms). What particles act as unchanged and undivided units depends upon what kind of process is involved. Some phenomena, such as evaporation, are explained by a process in which the molecules remain unchanged and identical. In chemical reactions, however, the molecules lose their identity. Their structures are broken up, and the composing atoms, while retaining their own identity, are rearranged into new molecules. With nuclear reactions a new level is reached, on which the atoms themselves are no longer considered as indivisible: more elementary particles than the atoms appear in the explanations of nuclear reactions.

Extensions to other fields. Whereas classical Atomism spoke mainly of material atoms (*i.e.,* of particles of matter), the success of the atomic doctrine encouraged the extension of the general principles of Atomism to other phenomena, more or less removed from the original field of application. Rather plausible, for example, was the extension of Atomism to the phenomena of electricity. There were reasons to suppose the existence of an elementary charge of electricity associated with an elementary material particle, the electron (19th century). A second fruitful extension concerned energetic processes (20th century). Some experimental data suggested the hypothesis that energy can exist only in amounts that are whole multiples of an elementary quantity of energy. Extensions of the idea of an atomic structure to amounts of gravitation and even to time have been attempted but have not been sufficiently confirmed.

Logical Atomism, monadology, associationism

More removed from the original field of application of Atomism is a theory known as Logical Atomism (developed by the eminent philosopher and logician Bertrand Russell and by the philosopher of language Ludwig Wittgenstein), which supposes that a perfect isomorphism exists between an "atom" of language (*i.e.,* an atomic proposition) and an atomic fact; *i.e.,* that for each atomic fact there is a corresponding atomic proposition. An atomic proposition is one that asserts that a certain thing has a certain quality; *e.g.,* "this is red." An atomic fact is the simplest kind of fact and consists in the possession of a quality by some individual thing.

Another application of Atomism (albeit in a moot sense) lies in the monadology of the philosopher-scientist G.W. Leibniz. According to Leibniz the atoms of Democritus, who provides the paradigm case of ancient Greek Atomism, were not true unities; possessing size and shape, they still were divisible in principle. The ultimate constituents of things must, therefore, be points, said Leibniz—not mathematical but metaphysical points; *i.e.,* points of real existence. They are indeed a kind of soul, which he came to call "monads."

In psychology, Atomism is a doctrine about perception. It holds that what man perceives is a mosaic of atomic sensations, each independent and unconnected with any other sensation. According to the early modern Empiricist David Hume and the pre-World War I father of experimental psychology Wilhelm Wundt, the fact that man nevertheless experiences an ordered whole formed from the unordered "atoms" of perception is caused by the mind's capacity to combine them by "association."

DIVERSE PHILOSOPHICAL CHARACTERIZATIONS OF ATOMISM

One being: Lemaître and Parmenides

The intrinsic nature of the atoms. In 1927 the Belgian astronomer Georges Lemaître formulated the hypothesis that the present high degree of differentiation of matter in space and the complexity of forms displayed by the various astronomical objects must have resulted from a violent explosion and subsequent dispersal of an originally highly compressed homogeneous material, a kind of "primitive atom," containing all of the matter that exists. From the philosophical viewpoint this hypothesis is interesting. By its attempt to reduce the manifold to unity it recalls the beginning of Greek philosophy, which was also inspired by a thesis of the unity of being, propounded by the Eleatic Parmenides. Even apart from their respective contexts, there is, of course, a great difference between Lemaître's and Parmenides' conceptions of the unity of being, for the latter combined the thesis of the unity of being with that of the immutability of being.

Although it would be wrong to classify Parmenides among the Atomists, it is nonetheless appropriate that in an introduction to the diverse forms of Atomism, his conception of reality as just one being should be mentioned. Parmenides' thesis is not only historically but also logically the cornerstone of atomistic thought. Any atomic theory can be interpreted as an attempt to reconcile the thesis of the unity and immutability of being with the fact that the senses observe multiplicity and change. The different ways in which the unity and immutability are understood characterize the different forms of Atomism.

Atoms as lumpish corpuscles. As corpuscles (minute particles), atoms can either be endowed with intrinsic qualities or be inherently qualityless.

Inherently qualitied atoms

The most striking basic differences in the material world, which lead to a first classification of substances in nature, are those between solids, liquids, gases, and fire. These differences are an observed datum that must be accounted for by every scientific theory of nature. It is, therefore, only natural that one of the first attempts to explain the phenomena of nature was based upon these differences and proclaimed that there are four qualitatively different primitive constituents of everything, namely, the four elements: earth, water, air, and fire (Empedocles, 5th century BC)—a theory that dominated physics and chemistry until the 17th century.

Although the theory of the four elements is not necessarily an atomistic theory, it obviously lends itself to interpretation in atomistic terms, namely, when the elements are conceived as smallest parts that are immutable. In this case, all observable changes are reduced to the separation and commingling of the primitive elementary substances. Thus Parmenides' thesis that being is immutable is maintained, whereas the absolute unity of being is abandoned. Yet, the fact that the infinite variety of forms and changes in nature is reduced to just one type of process between only four elementary kinds of atoms shows its affinity with the thesis of the unity of all being.

Notwithstanding the great disparity between the theory of the four elements and modern chemistry, it is clear that modern chemistry with its approximately 100 qualitatively different atoms falls into the same class of atomic theories as that of Empedocles. There are differences, of course, but these will be deferred for later discussion.

More removed from the original thesis of Parmenides was the theory of his contemporary Anaxagoras of Clazomenae, which assumed as many qualitatively different "atoms" as there are different qualitied substances in nature. Inasmuch as these atoms, which Anaxagoras called "seeds," were eternal and incorruptible, this theory still contains an idea borrowed from Parmenides. A special feature of Anaxagoras' theory was that every substance contains all possible kinds of seeds and is named after the kind of seed that predominates in it. Since the substance contains also other kinds of seed, it can change into something else by the separation of its seeds.

Another interesting form of Atomism with inherently qualitied atoms, also based on the doctrine of the four elements, was proposed by Plato. On mathematical grounds he determined the exact forms that the smallest parts of

the elements must have. Fire has the form of a tetrahedron, air of an octahedron, water of an icosahedron, and earth of a cube. Inasmuch as he characterized the atoms of the four elements by different mathematical forms, Plato's conception can be considered as a transition between the qualitative and quantitative types of Atomism.

Qualityless atoms

The most significant system of Atomism in ancient philosophy was that of Democritus (5th century BC). Democritus agreed with Parmenides on the impossibility of qualitative change but did not agree with him on that of quantitative change. This type of change, he maintained, is subject to mathematical reasoning and therefore possible. By the same token, Democritus also denied the qualitative multiplicity of visible forms but accepted a multiplicity based on purely quantitative differences. In order to reduce the observable qualitative differences to quantitative differences, Democritus postulated the existence of invisible atoms, characterized only by quantitative properties: size, shape, and motion. Observed qualitative changes are based upon changes in the combination of the atoms, which themselves remain intrinsically unchanged. Thus Democritus arrived at a position that was defined above as Atomism in the strict sense. In order to make the motion of atoms possible, this Atomism had to accept the existence of the void (empty space) as a real entity in which the atoms could move and rearrange themselves. By accepting the void and by admitting a plurality of beings, even an infinite number of them, Democritus seemed to abandon—even more than Empedocles did—the unity of being. Nevertheless, there are sound reasons to maintain that, in spite of this doctrine of the void, Democritus' theory remained close to Parmenides' thesis of the unity of being. For Democritus' atoms were conceived in such a way that almost no differences can be assigned to them. First of all, there are no qualitative differences; the atoms differ only in shape and size. Secondly, the latter difference is characterized by *continuity:* there are no privileged shapes and no privileged sizes. All shapes and sizes exist, but they could be placed in a row in such a manner that there would be no observable difference between successive shapes and sizes. Thus not even the differences in shape and size seem to offer any ground explaining why atoms should be different. By accepting an infinite number of atoms, Democritus retained as much as possible the principle that being is *one.* With respect to the acceptance of the void, it must be stressed that the void in the eyes of Democritus is more nonbeing than being. Thus even this acceptance does not seriously contradict the unity of being.

Atoms as sheer extension. Democritus had declared quantitative differences to be intelligible, because they were subject to mathematical reasoning. Precisely this relationship between quantitative differences and mathematics made it impossible for Descartes (17th century) to think along the atomistic lines of Democritus. If the only thing that is clearly understandable in matter is mathematical proportions, then matter and spatial extension are the same—a conclusion that Descartes did not hesitate to draw. Consequently, he rejected not only the idea of indivisible atoms but also that of the void. In his eyes the concept "void" is a contradiction in terms. Where there is space, there is by definition extension and, therefore, matter.

Descartes's vortices

Yet, however strange it may seem in view of his identification of matter with extension, Descartes offered nonetheless a fully developed theory of smallest particles. To the questions that arise immediately as to how these particles are separated and distinct from each other, Descartes answered that a body or a piece of matter is all of that which moves together. In the beginning of the world all matter was divided into particles of equal size. These particles were in constant motion and filled all of space. As, however, there was no empty space for moving particles to move into, they could only move by taking the places vacated by other particles that, however, were themselves in motion. Thus the motion of a single particle involved the motion of an entire closed chain of particles, called a vortex. As a result of the original motion, some particles were gradually ground into a spherical form, and the resulting intermediary space became filled with the surplus splinters or "grindings." Ultimately, three main types of particles were formed: (1) the splinter materials, which form the finest matter and possess the greatest velocity; (2) the spherical particles, which are less fine and have a smaller velocity; and (3) the biggest particles, which originated from those original particles that were not subject to grinding and became united into larger parts.

Thus Descartes could construct an atomic theory without atoms in the classical sense. Although this theory as such has not been of great value for the scientific atomic theory of modern times, its general tendency was not without importance. However arbitrarily and speculatively Descartes may have proceeded in the derivation of the different kind of corpuscles, he finally arrived at corpuscles characterized by differences in mass, velocity, amount of motion, etc.—properties that could be treated mathematically.

Atoms as centres of force: dynamic particles. Most systems of Atomism depict the action between atoms in terms of collision—*i.e.,* as actual contact. In Newton's theory of gravitation, however, action between bodies is supposed to be action at a distance—which means that the body in question acts everywhere in space. As its action is the expression of its existence, it is difficult to confine its existence to the limited space that it is supposed to occupy according to its precise shape and size. There is, therefore, no reason for a sharp distinction between occupied and empty space. Consequently, the mind finds it natural to consider the atoms not as extended particles but as point-centres of force. This conception was worked out by the Dalmatian scientist R.G. Boscovich (1711–87), who attempted to account for all known physical effects in terms of action at a distance between point-particles, dynamic centres of force.

Boscovich's point-centres and Leibniz's monads

Atoms as psychophysical monads. The idea of applying the atomistic conceptions not only to material but also to psychical phenomena is as old as Atomism itself. Democritus had spoken of the atoms of the soul. According to the principles of his doctrine these atoms could differ only quantitatively from those of the body: they were smoother, rounder, and finer. This made it easy for them to move into all parts of the body. Basically, however, the atoms of the soul were no less material than other atoms.

In Leibniz's monadology the situation was quite different. Leibniz did not first conceive of material atoms and then only later interpret the soul in terms of these atoms; from the beginning he conceived his "atoms," the monads, in terms of an analogy with the soul. A monad is much more a spiritual than a material substance. Monads have no extension; they are centres of action but not, first of all, in the physical sense. Each of the monads is gifted with some degree of perception; each mirrors the universe in its own way. Monads differ from each other, however, in the degree of perception of which they are capable.

The immutability of atoms. By their nature all atomic theories accept a certain degree of immutability of the atoms. For without any fixed units no rational analysis of complex phenomena is possible. At least with respect to the stable factors in the analysis involved, the atoms have to be considered as immutable. According to Atomism in the strict sense, this immutability had to be interpreted in an absolute way.

Absolute vs. relative

The same absolute interpretation appeared in classical chemistry, although its atomic theory deviated from Atomism in the strict sense by assuming qualitatively different atoms and by assuming molecules (rather stable aggregates of atoms). The decisive point, however, is that molecules were formed by mere juxtaposition of atoms without any intrinsic change of the qualities of the atoms. Modern atomic theory, in contrast, gives a less rigid interpretation of the immutability of elementary particles: the particles that build up an atom do not retain their identity in an absolute way.

In some philosophical atomistic theories, the immutability of the atoms has been understood in a highly relative sense. This interpretation arose mainly in the circles of those Aristotelian philosophers who tried to combine atomistic principles with the principle of Aristotle that elements changed their nature when entering a chemical

compound. The combination of both principles led to the doctrine known as the *minima naturalia* theory, which holds that each kind of substance has its specific *minima naturalia*, or smallest entities in nature. *Minima naturalia* are not absolutely indivisible: they can be divided but then become *minima naturalia* of another substance; they change their nature. In a chemical reaction the *minima* of the reagents change into the *minima* of the substances that result from the reaction.

Other differences. Atomisms also differ regarding the number of atoms, whether they occupy a void, and how they relate to one another.

Number of atoms. As has already been mentioned, Democritus introduced the hypothesis that the atoms are infinite in number. Although one may question whether the term infinite has to be taken in its strict sense, there is no doubt that by using this term Democritus wanted not merely to express the triviality that, on account of their smallness, there had to be an enormous quantity of atoms. Democritus also had a strong rational argument for postulating a strictly infinite quantity of atoms: only thus could he exclude the existence of atoms that specifically differed from each other.

When in modern science the problem of the number of atoms arises, the situation is quite different from that of the Greek Atomists. There is now much more detailed information about the properties of the atoms and of the elementary particles, and there is also in astrophysical cosmology some information about the universe as a whole. Consequently, the attempt to calculate the total number of atoms that exist is not entirely impossible, although it remains a highly speculative matter. In a time (around 1930) when all chemical atoms were supposed to be composed of electrons and protons, the pioneering joint-relativity-quantum astrophysicist A.S. Eddington calculated the number of these elementary particles to be $2 \times 136 \times 2^{256}$, or approximately 10^{79}, arguing that, since matter curves space, this is just the number of particles required barely to close the universe up into a hypersphere and to fill up all possible existence states.

Eddington's cosmical number

The existence of the void. To Democritus the existence of the void was a necessary element in atomistic theory. Without the void the atoms could not be separated from each other and they could not move. In the 17th century Descartes rejected the existence of the void, whereas Newton's conception of action at a distance was in perfect harmony with the acceptance of the void and the drawing of a sharp distinction between occupied and nonoccupied space.

The success of the Newtonian law of gravitation was one of the reasons that atomic theories came to prevail in the 18th century. Even with respect to the phenomena of light, the corpuscular and hence atomic theory of Newton, which held that light is made of tiny particles, was adopted almost universally, in spite of Huygens' brilliant development of the wave hypothesis.

When in the beginning of the 19th century the corpuscular theory of light in its turn was abandoned in favour of the wave theory, the case for the existence of the void had to be reopened, for the proponents of the wave theory did not think in terms of action at a distance; the propagation of waves seemed to presuppose, instead, a medium with not only geometrical properties but with physical ones as well. At first the physical properties of the medium, the ether, were described in the language of mechanics; later they were described in that of the electromagnetic field theory of J.C. Maxwell. Yet, to a certain extent the old dichotomy between occupied and nonoccupied space continued to exist. For, according to the ether theory, the atoms moved without difficulty in the ether, whereas the ether pervaded all physical bodies.

In contemporary science this dichotomy has lost its sharpness, owing to the fact that the distinction between material phenomena, which were supposed to be discontinuous, and the phenomena of light, which were supposed to be continuous, appears to be only a relative one. In conclusion it can be claimed that although modern theories still speak of space and even of "empty" space, this "emptiness" is not absolute: space has come to be regarded as the seat of the electromagnetic field, and it certainly is not the void in the sense in which the term was used by Democritus.

Atoms in external aggregation versus in internal relationship. In most forms of Atomism it is a matter of principle that any combination of atoms into a greater unity can only be an aggregate of these atoms. The atoms remain intrinsically unchanged and retain their identity. The classical atomic theory of chemistry was based upon the same principle: the union of the atoms into the molecules of a compound was conceived as a simple juxtaposition. Each chemical formula (*e.g.*, H_2O, H_2SO_4, NaCl, etc.) reflects this principle through the tacit implication that each atom is still an H, O, or S, etc., even when in combination to form a molecule.

Chemistry had a twofold reason for adopting this principle. One reason was an observational, the other a philosophical one. The fact that some of the properties of a chemical compound could, by simple juxtaposition, be derived from those of the elements (the molecular weight, for example, equals the simple sum of the respective atomic weights) was a strong factual argument in favour of the principle. Many properties of the components, however, could not be determined in this way. In fact, most chemical properties of compounds differed considerably from those of the composing elements. Consequently, the principle of juxtaposition could not be based on factual data alone. It was in need of a more general support. This support was offered by the philosophical idea that inspired all Atomism, viz., that if complex phenomena cannot be explained in terms of aggregates of more elementary factors, they cannot be explained at all.

Holistic tendencies

For the evaluation of this idea, the development of the scientific atomic theory is highly interesting, especially with respect to the interpretation of the concept of an aggregate. Is the only interpretation of this concept that of an assemblage in which the components preserve their individuality—like, for instance, a heap of stones?

Modern atomic theory offers an answer to this question. This theory still adheres to the basic principle that a complex structure has to be explained in terms of aggregates of more elementary factors, but it interprets the term "aggregate" in such a way that it is not limited to a mere juxtaposition of the components. In modern theories atomic and molecular structures are characterized as associations of many interacting entities that *lose* their own identity. The resulting aggregate originates from the converging contributions of all of its components. Yet, it forms a new entity, which in its turn controls the behaviour of its components. Instead of mere juxtaposition of components, there is an internal relationship between them. Or, expressed in another way: in order to know the properties of the components, one has to study not only the isolated components but also the structures into which they enter. To a certain extent modern atomic theory has bridged the gap between atomistic and holistic thought.

HISTORY AND MAJOR REPRESENTATIVES OF THE VARIOUS ATOMISMS

Philosophical Atomism. From the ancient Greeks through the 16th century, Atomism remained mainly philosophical.

Ancient Greek Atomism. It is characteristic of the importance of Greek philosophy that, already in the foregoing exposition of the different aspects of Atomism, several Greek philosophers had to be introduced. Not only the general idea of Atomism but also the whole spectrum of its different forms originated in ancient Greece. As early as the 5th century BC Atomism in the strict sense (Leucippus and Democritus) is found, along with various qualitative forms of Atomism: that of Empedocles, based on the doctrine of the four elements, and that of Anaxagoras, with as many qualitatively different atoms as there are different substances.

Criticisms by Plato and Aristotle

Yet, in spite of its successful start, Atomism did not gain preeminence in Greek thought. This is mainly because Plato and Aristotle were not satisfied with the atomistic solution of the problems of change as a *general* solution. They refused to reduce the whole of reality, including

man, to a system that knew nothing but moving atoms. Even with respect to the problems of the material world, Atomism seemed to offer no sufficient explanation. It did not explain the observable fact that, notwithstanding continual changes, a total order of specific forms continued to exist. For this reason Aristotle, with Plato, was more interested in the principle of order than in that of the material elements. In his own analysis of change, which resulted in the matter-form doctrine, Aristotle explicitly rejected the thesis of Democritus that in a chemical reaction the component parts retain their identity. According to Aristotle, the elements that entered into a composite with each other did not remain what they were but became a compound. Although there is some indication that in Aristotle's chemical theory smallest particles played a role, it was certainly not a very important one.

Meanwhile, atomistic ideas remained known in Greek thought. Their opponents paid much attention to them, and there were also a few adherents of Democritean Atomism in later times, such as the Greek hedonist Epicurus (*c.* 341–279 BC) and the Roman poet Lucretius Carus (*c.* 95–55 BC) who, through his famous didactic poem *De rerum natura* ("On the Nature of Things"), introduced Atomism into the Latin world.

The elachista of the early Aristotelian commentators. Empedocles had suggested an Atomism with qualitatively different atoms, based upon the doctrine of the four elements. Aristotle adopted the latter doctrine but without its atomistic suggestion. Certain Greek commentators on the works of Aristotle, however, viz., Alexander of Aphrodisias (2nd century AD), Themistius (4th century AD), and Philoponus (6th century AD), combined the Aristotelian theory of chemical reactions with atomistic conceptions. In their systems the atoms were called *elachista* ("very small" or "smallest"). The choice of this term is connected with the Aristotelian rejection of the infinite divisibility of matter. Each substance had its own minimum of magnitude below which it could not exist. If such a minimum particle were to be divided, then it would become a minimum of another substance.

The minima naturalia of the Averroists. The Latin commentators on Aristotle translated the term *elachista* into its Latin equivalent *minima* or also into *minima naturalia; i.e., minima* determined by the nature of each substance. In fact, for most medieval Aristotelians the *minima* acquired little more reality than the theoretical limit of divisibility of a substance; and in their descriptions of physical and chemical processes, they paid no attention to the *minima*. With the Averroists—followers of the Arab Aristotelian Averroës (1126–98)—an interesting development occurred. Agostino Nifo (1473–1538), for example, explicitly stated that in a substance the *minima naturalia* are present as *parts;* they are physical entities that actually play a role in certain physical and chemical processes. Because the *minima* had acquired more physical reality, it then became necessary to know how the properties of the *minima* could be connected with the sensible properties of a substance. Speculations in this direction were developed by the Italian physician, philosopher, and litterateur Julius Caesar Scaliger (1484–1558).

Modern scientific Atomism: early pioneering work. Modern Atomism arose with the flowering of science in the present sense of the word.

The 17th century. In the history of Atomism the 17th century occupies a special place for two reasons: it saw the revival of Democritean Atomism, and it saw the beginning of a *scientific* atomic theory.

The revival of Democritean Atomism was the work of the ambiguous Epicureo-Christian thinker Pierre Gassendi (1592–1655), who made his contemporaries not only better acquainted with Atomism but also succeeded in divesting it of the materialistic interpretation with which it was hereditarily infected. This reintroduction of Democritus was well timed. Because of its quantitative character Democritus' Atomism invited for its elucidation the application of mathematics and mechanics, which in the 17th century were sufficiently developed to answer this invitation. In point of fact, the 17th century was more interested in the possibilities that Atomism offered for a physical theory than it was in the philosophical differences between the different atomistic systems. For this reason it saw, for example, hardly any difference between the systems of Gassendi and Descartes, although the latter explicitly rejected some of the fundamentals of Democritus, such as the existence of the void and the indivisibility of the atoms, as noted above (see *Atoms as sheer extension*).

Gassendi, Sennert, and Boyle

In the case of scientists mainly interested in the chemical aspects, the same shift of emphasis from philosophical to scientific considerations can be discerned. According to the physician and philosopher of nature Daniel Sennert (1572–1637), Democritus' Atomism and the *minima* theory really amounted to the same thing. As far as philosophy was concerned, Sennert was only interested in the general idea of Atomism; the precise content of an atomic doctrine in his view ought to be a matter of chemical experimentation. His own experience as a chemist taught him the specific differences existing between the atoms. In this respect Sennert continued the *minima* tradition. His own contribution to the chemical atomic theory lay in the clear distinction that he made between elementary atoms and the *prima mista,* or atoms of chemical compounds.

The early modern experimentalist Robert Boyle (1627–91) followed the same line of thought as Sennert, but he was much more aware of the discrepancy between Democritus' Atomism and an atomic theory suitable for chemical purposes. Boyle's solution to this problem was the thesis that the atoms of Democritus were normally associated into primary concretions, which did not easily dissociate and which acted as elementary atoms in the chemical sense. These primary concretions can combine to form compounds of a higher order, which may be compared to Sennert's *prima mista* and to the molecules of modern chemistry.

Founding of modern Atomism. The 17th century had laid the theoretical foundations for a scientific atomic theory. For its further development it was in need of better chemical insights, especially concerning the problem of what substances should be considered as chemical elements. Boyle had shown conclusively that the traditional four "elements" were certainly not elementary substances, but at the same time he confessed that he did not yet see any satisfactory method to determine which substances were true elements. This method was provided by another of the principal founders of modern chemistry A.-L. Lavoisier (1743–94): a chemical element is a substance that cannot be further analyzed by known chemical methods.

Lavoisier and Dalton

John Dalton (1766–1844), usually regarded as the father of modern atomic theory, applied the results of Lavoisier's chemical work to atomistic conceptions. When Dalton spoke of elementary atoms, he did not have a merely theoretical idea in mind but the chemical elements as determined by Lavoisier. Dalton held that there are as many different kinds of elementary atoms as there are chemical elements. The atoms of a certain element are perfectly alike in weight, figure, etc.; and the same point applies to the atoms of a certain compound. As weight was the decisive characteristic in Lavoisier's theory, Dalton stressed the importance of ascertaining the relative weights of atoms and the number of elementary atoms that constituted one compound "atom." As to the question of the way in which the atoms are combined in a compound, Dalton conceived this combination as a simple juxtaposition with each atom under the influence of Newtonian forces of attraction. The atoms retain their identity through a chemical reaction. In this one point the founder of the chemical atomic theory did not differ from Democritus.

Recent and contemporary scientific Atomism. Until its development in the third decade of the 20th century, the scientific atomic theory did not differ philosophically very much from that of Dalton, although at first sight the difference may appear large. Dalton's atoms were no longer considered to be immutable and indivisible; new elementary particles sometimes appeared on the scene; and molecules were no longer seen as a mere juxtaposition of atoms—when entering into a compound atoms became ions. Yet, these differences were only accidental; the atoms revealed themselves as composed of more elementary particles—protons, neutrons, and electrons—

but these particles themselves were considered then as immutable. Thus the general picture remained the same. The material world was still thought to be composed of smallest particles, which differed in nature and which in certain definite ways could form relatively stable structures (atoms). These structures were able to form new combinations (molecules) by exchanging certain component parts (electrons). The whole process was ruled by well-known mechanical and electrodynamic laws.

In contemporary atomic theory the differences from Dalton are much more fundamental. The hypothesis of the existence of immutable elementary particles has been abandoned: elementary particles can be transformed into radiation and vice versa. And when they combine into greater units, the particles do not necessarily preserve their identity; they can be absorbed into a greater whole.

Atomism in the thought of India. It is interesting to note that atomistic conceptions are not restricted to Western philosophy and science. Examples of qualitative Atomism, based upon the doctrine of the four elements, are also found in Indian philosophy. In some Indian systems the atoms are not absolutely indivisible but only relatively so. In certain aspects Indian Atomism is, therefore, more related to the *minima* doctrine than to the Atomism of Democritus. Indian Atomism has, however, not developed into a scientific theory.

FOUNDATIONAL ISSUES POSED BY ATOMISM

Atomism as a metaphysical system. In discussing Atomism, one particular system, that of Democritus, has been here distinguished as Atomism in the strict sense because of the fact that in no other system have the foundational issues of Atomism been so clearly expressed. Atomism in the strict sense is not merely one of the historical forms of Atomism, one of the many possible scientific attempts at explaining certain physical phenomena; it is, first of all, a metaphysical system: it has been presented as the only possible explanation of change and multiplicity. And as a metaphysic it is rationalistic, mechanistic, and realistic.

Rationalistic, mechanistic, realistic

It is *rationalistic* because it has so much confidence in reason that, in order to explain observed phenomena, it does not hesitate to postulate the existence of unobservable atoms: *i.e.*, of things that are in principle unobservable by the human senses and can be known only by a process of reasoning. Atomists go even further, for they not only are convinced of the existence of atoms but they also think it possible to deduce in a rational way their fundamental properties. Moreover, the description of these properties in *mechanistic* terms is not just a matter of convenience; it is supposed to be the adequate expression of *reality*.

This rationalistic and mechanistic metaphysics is not only characteristic of Democritus' Atomism but also of the early forms of scientific Atomism. The clearest expression of this metaphysics is found in Descartes. For Democritus mechanistic concepts are clear and distinct ideas, so that any further experimental investigation is superfluous. It should be stressed that the atomistic assumption that the human mind, by mere reasoning, can know the properties of the atoms is a necessary consequence of the idea that atoms are not subject to internal change; for the changeless can never be a subject of experimentation. The great weakness of the mechanistic concept of immutable atoms was that it forced the analyzing experiments to stop at the atoms; but this weakness could reveal itself only after, in the course of the further development of science, the fundamentally experimental character of human knowledge had become evident.

This weakness, in fact, was precisely one of the reasons why Aristotle rejected the Atomism of Democritus, viz., that the latter had postulated atoms that were not subject to change. For Aristotle the very essence of matter was its being subject to change; hence to him the concept of immutable atoms was a contradiction in terms.

Kant's anti-realism and scientific models

Aristotle's criticism of Atomism was clearly directed against its mechanistic metaphysics, not against its realism. This latter characteristic was the target, however, of an attack launched by the incomparable 18th-century epistemologist Immanuel Kant. In a famous argument, known as the antinomy of the continuum, Kant tried to prove that the acceptance of the reality of spatial extension, the cornerstone of Atomism, led to contradictions. His argument can be summarized as follows: It is possible to prove that any compound must be composed of simple things (for if not, there would be nothing but composition). On the other hand, it also is possible to prove that no material thing can be simple, for the very reason that a part of an extended being is always extended and is thus open to division. Hence, every allegedly simple part is at once simple and nonsimple. Consequently, spatial extension cannot be real. Extension is therefore not a property of the material world itself; it is a form imposed upon reality by man's perception.

By his argument Kant did not intend to reject atomic theories as such; he rejected only their realistic pretensions. Kant was deeply convinced that man had to *think* of nature by way of analogy with a mechanism, but he denied that knowledge construed in such a fashion could reach reality itself.

In the 19th century, scientists were, as a rule, hardly impressed by Kant's attack on the realistic pretensions of human knowledge. Scientists had already learned to go their own way and no longer worried about philosophical considerations. Only when an internal crisis in science itself arose were they prepared to reflect upon their presuppositions. Such a crisis occurred in the 20th century when science was forced to accept the relativity of both classical models: the wave and the particle models.

To a certain extent, the problem of whether a scientific model is nothing but a subjective construct in which the scientist unites his experience is the same as the problem that Kant had in mind. One of the differences, however, is that in Kant's time science was still rather exclusively *theory*. Its close connection with praxis (practice, doing) had not yet been discovered. For this reason the Kantian epistemological (or human knowledge) problem could centre on such a question as: what guarantee does the *knowing* subject have that his "models" of reality reflect reality itself? Inasmuch as, in an exclusively theoretical science, the only contact that one has with reality is afforded by means of one's knowledge, the problem seems to be insoluble.

The development of science from a theoretical to an experimental discipline forces philosophy to view the epistemological problem in a new way. For in an experimental science man is in a twofold contact with reality, viz., by his knowledge and by his experimental praxis. Modern atomic theory is one of the best examples to illustrate this point. It was this theory that was most directly confronted with the problem of the realistic value of its models. It could take up this challenge because of the theory's effectiveness for experimental praxis, which is the final judge of the realistic value of the theoretical models. It has confirmed the audacious rational speculations of ancient Atomism; but at the same time it has revealed that, in order to be really effective, reason is in need of experimental assistance.

Ancient Greek Atomism versus contemporary scientific Atomism. In comparing Greek Atomism and modern atomic theories, it should be recalled that in Greek thought philosophy and science still formed a unity. Greek Atomism was inspired as much by the desire to find a solution for the problems of mutability and plurality in nature as by the desire to provide scientific explanations for specific phenomena. While it is true that some of the Greek Atomists' ideas can rightly be considered as precursors of later physics, the main importance of the old atomistic doctrines for modern science does not lie in these primitive anticipations. Much more important is the attempt to take seriously the variety and mutability discerned by sense experience and yet to reconcile it with the thesis of Parmenides about the unity and the immutability of matter. In its search for universal and unchangeable laws, modern science is to a great extent inspired by the same idea as Parmenides, since universal laws presuppose a certain unity in the material world and unchangeable laws cannot be established without the presupposition that something unchangeable must be hidden behind all changes. By the same token, without this latter presupposition experiments would not make any sense at all. For if the diversity of reactions occurring under different conditions is to re-

Parmenides and the modern quest for laws

veal anything, these reactions must be the expression of an immutable nature. The differences have to indicate something about that which remains the same. The great achievement of the Greek philosophers was, therefore, that they took a general view of nature as a whole that made a scientific attitude possible. To this, both the quantitative and the qualitative forms of Atomism contributed, the former drawing attention to the mathematical aspects of the problem, the latter to the observational.

A comparison of ancient Greek Atomism with scientific Atomism merely on the basis of their respective scientific contents would therefore do a great injustice to Greek Atomism; it would in fact misjudge its main value. Such a comparison would, however, also take too narrow a view of modern scientific Atomism. It would imply the philosophical irrelevance of the latter. It has here been shown, however, that the later development of the scientific atomic theory has clarified many philosophical problems that, as basic issues, divided Atomism in the strict sense from other forms of Atomism. To mention only a few examples: the development of the scientific atomic theory has deepened man's insights into the relationship between a whole and its components, into the relative character of ultimate particles, and into certain fundamental epistemological problems.

Evaluation of Atomism. The success of the atomic theory shows the value of the idea of Atomism: the explanation of the complex in terms of aggregates of fixed particles or units. Its history also shows, however, the inherent danger of this idea, namely, that of absolutism. History has corrected this absolutism: the unitary factors have to be conceived as ultimate only with respect to the complex under consideration, and their union into aggregates need not occur only by way of juxtaposition. (A.G.M.v.M.)

Eleaticism

One of the principal schools of ancient Pre-Socratic philosophy was Eleaticism, so called from its seat in the Greek colony of Elea (or Velia) in southern Italy. This school, which flourished in the 5th century BC, was distinguished by its radical monism; *i.e.,* its doctrine of the One, according to which all that exists (or is really true) is a static plenum of Being as such, and nothing exists that stands either in contrast or in contradiction to Being. Thus, all differentiation, motion, and change must be illusory. This monism is also reflected in its view that existence, thought, and expression coalesce into one.

Radical monism

The sources for the study of Eleaticism are both archaeological and literary. Archaeologists have ascertained that, at the time of Parmenides, the Rationalist who founded the school, Elea was a large town with many temples, a harbour, and a girdle of walls several miles long. They have also unearthed a site presumed to be that of the medical school that Parmenides founded and an inscription bearing Parmenides' name.

The literary sources consist of fragments preserved by later classical authors—19 from Parmenides, four from his pupil Zeno, renowned for his paradoxes of motion, and 10 from another pupil, Melissus, an admiral of Samos—of which all but three from Parmenides and two from Melissus are 10 lines or less in length. The interpretation is also affected by the weight of scholarly opinion that now holds (as against Karl Reinhardt, an early 20th-century philologist and historian of philosophy) that Parmenides reacted to Heracleitus and not vice versa. Moreover, the biases of the citing authors must be weighed. Heracleitans and Parmenideans of the second generation, for example, saw their masters, simplistically, only as the prophets of movement and immobility; and the ancient Skeptic Sextus Empiricus distorted Parmenides' thinking into problems of epistemology (theory of knowledge), because this is what his Skeptical eye saw in his writings.

THE ELEATIC SCHOOL VIS-À-VIS RIVAL MOVEMENTS

Each member of the Eleatic school espoused a distinctive variety of Eleaticism: Parmenides pursued a direct and logical course of thought and viewed Being as finite and timeless; Zeno concurred in Parmenides' doctrines but employed the indirect methods of reductio ad absurdum and infinite regress (see below); and Melissus modified the doctrines, viewing Being as infinitely extensive and eternally temporal. A fourth thinker, the Sicilian Sophist Gorgias of Leontini, though not an adherent to Eleaticism, employed the methods of the Eleatic Zeno to defend its opposite—a nihilism that affirmed Non-Being instead of Being.

Eleaticism represents a reaction against several tendencies of thought. Methodologically, it spurned the empirical (observational) approach taken by earlier cosmologists, such as the 6th-century Milesians Thales and Anaximenes, who discerned ultimate reality in water and in air (or breath), respectively, for these substances are materializations of Being—analogous to the materialization that occurs in Pythagoreanism in passing from an abstract line or plane or three-dimensional form to a solid perceptible body—rather than Being itself; or, at best (as some scholars hold), the substances are mythological representations of Being. The Eleatics, on the contrary, ignoring perceptual appearances, pursued a rationalistic—*i.e.,* a strictly abstract and logical—approach and thus found reality in the all-encompassing, static unity and fullness of Being and in this alone. Thus Parmenides was the father of pure ontology.

Contrasts with other movements

Ontologically—in its view of the nature of Being—the Eleatic school, holding to what Parmenides called the "way of truth" ("what is"), stood opposed to two other "ways of research" that were then current: first, to the "way of opinion" (or seeming; later developed at length by Plato in the *Sophist*), which held that a being comprises or is defined not only by what it is but also by what it is not—*i.e.,* by its contrast with other things; and, second, to a way recognizable as that of Heracleitus, a caustic and often cryptic philosopher then living in Ephesus, who maintained—still more radically—that the essence of a being lies in part in its involvement in, or even its identity with, its opposite. Finally, as an aspect of Parmenides' opposition to the way of opinion, he was in reaction also against Anaximander, another Milesian scientist and philosopher. Though Anaximander's basic principle, the *apeiron* ("boundless"), was duly abstract and not a part of the world itself (as were water and air), his philosophy depended, nonetheless, upon the world's contrast with the infinite *apeiron,* from which all things come and to which they return "in accordance with the ordinance of Time." This contrast—which, in a Pythagorean version, envisioned the world as breathing in voidness from the infinite outer breath in order to keep things apart or discrete—thus spawned a "many" that contradicts the Eleatic One.

THE RIGOROUS ONTOLOGISM OF PARMENIDES AND MELISSUS

For a long time Xenophanes of Colophon, a religious thinker and rhapsode of the 6th–5th century BC, was considered the founder of the Eleatic school and Parmenides' mentor. This ancient claim, however, has been successfully criticized by Reinhardt. It is even possible that, on the contrary, Xenophanes was an older pupil of Parmenides. In any case, his monistic view of a cosmic God, whom he may have equated pantheistically with Being itself, was Eleatic in its contention that God is one and ungenerated, that his seeing, thinking, and hearing are equally all-pervading (*i.e.,* he is not a composite), and that he "always remains in the same place, not moving at all."

Parmenides' poem *Peri physeōs* ("On Nature") is divided into three parts: (1) a proem (preface), in which his chariot ride through the heavens to the very seat of the goddess Alētheia (Truth) is described and their initial conversation is related, in which she announces that he is "to learn all things, both the unshaken heart of well-rounded truth and also what seems to mortals, in which is no true conviction"; (2) the "Way of Truth," the main part, in which the real and unique Being is depicted; and (3) the "Way of Opinion" (or Seeming), in which the empirical world—*i.e.,* the single things as they appear every day to every man—is presented.

Parmenides' poem and the two ways of research

Logical and linguistic approach. Thus, at the very heart of Parmenides' philosophy lies the distinction made by the goddess (in Fragment 2) between the two "ways of

research." As noted earlier, the first is the antinomy (or paradox) of those who think and say that everything is Being and who shun all assertions of Non-Being; and the second is that of those who think and say that something is in a way and is not in another way—that a book is a book, for example, and not a table. There is, however, also a third way that is far more erroneous and fallacious than the second: that of Heracleitus, who acknowledged, just as Parmenides did, the ontological antinomy of is and is not but reversed it, holding that the real way of understanding things is to grasp their essential contradiction, their intrinsic opposition to everything else. In this view, one must say that to be a table is also not to be just a table and that to be a chair is not to be just a chair but to be also a table, because not only opposite things but also things that are merely different are bound to each other. Thus, life is death to Heracleitus, death is life, and justice would be meaningless if it had no injustice to defeat.

In essence, then, the possible ways are three: (1) that of renouncing all contradictions whatsoever (truth); (2) that of contradicting oneself relatively (seeming); and (3) that of contradicting oneself completely and absolutely (Heracleitus). And Eleaticism chose the first, the absolutely noncontradictory way that says that only what is, Being, is really true.

Coalescence of Being, thought, and expression

Not-Being, in fact, can neither be recognized nor expressed, for, as Parmenides then added, "for the same thing can be thought and can exist." And—if one may guess at the words (now lost) that probably followed—what-is-not you can neither know nor say; thus, to think is indeed the same as to say that what you think is. To this coalescence of existing reality and the intellectual grasping of it, Parmenides also added the linguistic communication of such knowledge. Each way of research, in fact, is at the same time a way of speculation and a way of diction; *i.e.,* both a way of searching for truth with one's mental eyes and of expressing it in words. The primal source of the Eleatic philosophy thus lies in the archaic sense of language, according to which one cannot pronounce "yes" and "no" without deciding about the reality or unreality of the objects of the statements. Thus, "yes," or "is," becomes the name of the truth; and "no," or "it is not," becomes that of its opposite.

This Eleatic principle may be illustrated by a passage from Aeschylus, a leading Greek dramatist, who, in his *Hepta epi Thēbais* (Eng. trans., *Seven Against Thebes*), judged it very appropriate that Helen would have destroyed Troy, because her name—naïvely derived from *helein* ("destroy") and *naus* ("ship")—marked her as a destroyer of ships. Here *nomen est omen:* the language is not merely a symbol; it corresponds to reality in its very structure. Thus, the Eleatic could not imagine a truth that is only expressible but not thinkable nor one that is only thinkable but not expressible.

Monistic theory of Being. From the premise of the essential coalescence of language and reality follows Parmenides' theory of Being, which comprises the heart of his philosophy. The only true reality is Eōn—pure, eternal, immutable, and indestructible Being, without any other qualification. Its characterizations can be only negative, expressions of exclusions, with no pretense of attributing some special quality to the reality of which one speaks.

Conflict with Melissus on the eternity of Being

In Fragment 8, verse 5, Parmenides said that the absolute Being "neither was nor will be, because it is in its wholeness now, and only now." Thus, its presence lasts untouched by any variation in time; for no one can find a genesis for it, either from another being (for it is itself already the totality of Being) or from a Not-Being (for this does not exist at all).

Obviously, this Parmenidean conception of the eternal presence of the Being conflicts with Melissus' idea of the perpetual continuation of the Being in the past, in the present, and in the future. Thus, if Eleaticism had been founded by Melissus, no one could have really understood its actual doctrine. One could suspect in it only an aspiration to have things capable of being really enduring. But even then the theory would hardly be understandable, because what one wants is not stable things in general; one wants good things to be firm and stable and bad things to be ephemeral. The perpetual continuity of existence as espoused by Melissus was despised by Parmenides just because "will be" and "has been" are not the same as "is." Only "is" is the word of the reality—just because it is the right name for the right thinking of the right Being.

Among the consequences of this Eleatic conception is the rejection of every change (birth, movement, growth, death) as pertaining only to the second-rate reality, which is known and expressed through the second "way of research." Thus, the true and noncontradictory reality is extraneous to all of those happenings, great or small, that make the constant stuff of all history.

Secondly, the real Being has no difference, no lack, no variety whatsoever in itself. Melissus is here the true pupil of Parmenides, who said that the Eōn is so closely connected in itself that "all Being is neighbour of all Being": for Melissus developed this theory by the negation of every form of *kenon* ("void"): the Being is an absolute plenum just because every lack in its plentifulness would amount to a presence of some Not-Being.

THE PARADOXES OF ZENO

The position of the other great pupil of Parmenides, Zeno of Elea, was clearly stated in the first part of Plato's dialogue *Parmenides.* There Zeno himself accepted the definition of Socrates, according to which he did not really propose a philosophy different from that of Parmenides but only tried to support it by the demonstration that the difficulties resulting from the pluralistic presupposition of the *polla* (the multiple beings of man's daily experience) were far more severe than those that seemed to be produced by the Parmenidean reduction of all reality to the single and universal Being.

Paradoxes of plurality

The arguments by which Zeno upheld his master's theory of the unique real Being were aimed at discrediting the opposite beliefs in plurality and motion. There are several arguments against plurality. First, if things are really many, everything must be infinitely small and infinitely great—infinitely small because its least parts must be indivisible and therefore without extension and infinitely great because any part having extension, in order to be separated from any other part, needs the intervention of a third part; but this happens to such a third part, too, and so on ad infinitum.

Very similar is the second argument against plurality: if things are more than one, they must be numerically both finite and infinite—numerically finite because they are as many things as they are, neither more nor less, and numerically infinite because, for any two things to be separate, the intervention of a third thing is necessary, ad infinitum. In other words, in order to be two, things must be three, and, in order to be three, they must be five, and so on. The third argument says: if all-that-is is in space, then space itself must be in space, and so on ad infinitum. And the fourth argument says: if a bushel of corn emptied upon the floor makes a noise, each grain must likewise make a noise, but in fact this does not happen.

Paradoxes of motion

Zeno also developed four arguments against the reality of motion. These arguments may also be understood (probably more correctly) as proofs *per absurdum* of the inconsistency of any presupposed multiplicity of things, insofar as these things may be proved to be both in motion and not in motion. The first argument states that a body in motion can reach a given point only after having traversed half of the distance. But, before traversing half, it must traverse half of this half, and so on ad infinitum. Consequently, the goal can never be reached.

The second argument is that of "Achilles and the tortoise." If in a race the tortoise has a start on Achilles, Achilles can never reach the tortoise; for, while Achilles traverses the distance from his starting point to that of the tortoise, the tortoise will have gone a certain distance, and, while Achilles traverses this distance, the tortoise goes still further, ad infinitum. Consequently, Achilles may run indefinitely without overtaking the tortoise. This argument is fundamentally identical to the previous one, with the only difference being that here two bodies instead of one are moving.

The third argument is the strongest of them all. It says

the following: So long as anything is in a space equal to itself, it is at rest. Now, an arrow is in a space equal to itself at every moment of its flight; therefore, even the flying arrow is at rest all of the time. And the final argument says: Two bodies moving at equal speed traverse equal spaces in an equal time. But, when two bodies move at equal speed in opposite directions, one passes the other in half of the time that a moving body needs to pass a body that is at rest.

The difficulty with all these arguments is that of really understanding them in their historical frame, which neither Aristotle—who was mainly concerned to confute Zeno—nor many modern scholars—who are concerned with developing new theories for the calculation of infinitesimal quantities—really tried to do. Moreover, the role of the author of the paradoxes in the history of Greek philosophy is itself paradoxical, for many of the same arguments by which Zeno proved the self-contradictory nature of the unity considered as the smallest element of a pluralistic reality (the Many) were later similarly used by Gorgias and Plato to demolish the Parmenidean One-Totality itself.

THE DECLINE OF ELEATICISM

This problem is also connected with that of the correct interpretation of the second part of Plato's *Parmenides.* Here the discussion to which Zeno submits the young Socrates is meant as a serious exemplification of the logical training that Socrates still needs if he wants to make progress in philosophy. But the result is simply comic—a "fatiguing joke," according to Plato's *Parmenides*—because Zeno always starts from the mere principles of the pure Being or the One and arrives at absurd conclusions: everything is shown to be true as well as false and deducible and not deducible from everything else.

Gorgias' nihilism and Plato's *Sophist*

Such dialectical futility had been anticipated by the nihilism of Gorgias, presented in a work ironically entitled *Peri tou mē ontos ē peri physeōs* ("On That Which Is Not, or On Nature"), in which he said (1) that nothing exists; (2) that if anything exists, it is incomprehensible; and (3) that if it is comprehensible, it is incommunicable—and in so doing he applied Parmenides' coalescence of Being and thought and expression to Non-Being instead of to Being and thus signalled the decline of Eleaticism.

The serious discussion and criticism of the Eleatic philosophy, however, and the positive interpretation of every Non-Being as a *heteron* (*i.e.,* as a being characterized only by its difference from "another" being) is neither in Gorgias nor in the *Parmenides* but in the *Sophist* of Plato. There Plato argued that the antinomy between *On* and *Mē-On* (Being and Non-Being) does not really exist, the only real antinomy being that of *Tauton* and *Heteron; i.e.,* only that of a single object of consciousness in its present determination and all other things from which it is distinguished.

The real story of ancient Eleaticism thus ends with Plato and with Democritus, who said that Being exists no more than Non-Being, the thing no more than the no-thing. But many thinkers, and great thinkers at that—from Aristotle to Kant and from Hegel to Marx—have continued to work or to fight with the antinomy of Being and Non-Being.

(G.C./L.H.St.)

Epicureanism

In a strict sense, Epicureanism is the philosophy taught by Epicurus (341–270 BC); in a broad sense, it is a system of ethics embracing every conception or form of life that can be traced to the principles of his philosophy. In ancient polemics, as often since, the term was employed with an even more generic (and clearly erroneous) meaning as the equivalent of hedonism, the doctrine that pleasure or happiness is the chief good. In popular parlance, Epicureanism thus means devotion to pleasure, comfort, and high living, with a certain nicety of style.

THE NATURE OF EPICUREANISM

Several fundamental concepts characterize the philosophy of Epicurus. In physics, these are Atomism, a mechanical conception of causality, limited, however, by the idea of a spontaneous motion, or "swerve," of the atoms, which interrupts the necessary effect of a cause; the infinity of the universe and the equilibrium of all forces that circularly enclose its phenomena; the existence of gods conceived as beatified and immortal natures completely extraneous to happenings in the world. In ethics, the basic concepts are the identification of good with pleasure and of the supreme good and ultimate end with the absence of pain from the body and the soul—a limit beyond which pleasure does not grow but changes; the reduction of every human relation to the principle of utility, which finds its highest expression in friendship, in which it is at the same time surmounted; and, in accordance with this end, the limitation of all desire and the practice of the virtues, from which pleasure is inseparable, and a withdrawn and quiet life.

Atomism; pleasure; friendship

In principle, Epicurus' ethic of pleasure is the exact opposite of the Stoic's ethic of duty. The consequences, however, are the same: in the end, the Epicurean is forced to live with the same temperance and justice as the Stoic. Of utmost importance, however, is one point of divergence: the walls of the Stoic's city are those of the world, and its law is that of reason; the limits of the Epicurean's city are those of a garden, and the law is that of friendship. Though this garden can also reach the boundaries of earth, its centre is always a man.

HISTORY OF EPICUREANISM

Epicurus' predecessors were in physics Leucippus and Democritus and in ethics Antiphon Sophista, Aristippus of Cyrene, and Eudoxus of Cnidus, a geometer and astronomer. Epicurus differed from all of these in his systematic spirit and in the unity that he tried to give to every part of philosophy. In this respect, he was greatly influenced by the philosophy and teachings of Aristotle—taking over the essentials of his doctrines and pursuing the problems that he posed.

Epicurus' own life and teachings. In 306 BC, Epicurus established his school at Athens in his garden, from which it came to be known as The Garden.

His works. In accordance with the goal that he assigned to philosophy, Epicurus' teaching had a dogmatic character, in substance if not in form. He called his treatises *dialogismoi,* or "conversations." Since the utility of the doctrines lay in their application, he summarized them in *stoicheia,* or "elementary propositions," to be memorized. In this respect, Epicurus was the inventor of the catechetical method. The number of works produced by Epicurus and his disciples reveals an impressive theoretical activity. But no less important was the practical action in living by the virtues taught by him and in honouring the obligations of reciprocal help in the name of friendship. In these endeavours, continuous assistance was rendered by Epicurus himself, who, even when old and ill, was occupied in writing letters of admonishment, guidance, and comfort—everywhere announcing his gospel of peace and, under the name of pleasure, inviting to love.

Doctrine of Epicurus. Philosophy was, for Epicurus, the art of living, and it aimed at the same time both to assure happiness and to supply means to achieve it. As for science, Epicurus was concerned only with the practical end in view. If possible, he would have done without it. "If we were not troubled by our suspicions of the phenomena of the sky and about death," he wrote, "and also by our failure to grasp the limits of pain and desires, we should have no need of natural science." But this science requires a principle that guarantees its possibilities and its certainty and a method of constructing it. This principle and this method are the object of the "Canon," which Epicurus substituted for Logic. Since he made the "Canon" an integral introduction to the "Physics," however, his philosophy falls into two parts, the "Physics" and the "Ethics."

His "Canon"

The name canon, which means "rule," is derived from a special work entitled "On the Criterion, or Canon." It held that all sensations and representations are true and serve as criteria. The same holds for pleasure and pain, the basic feelings to which all others can be traced. Also

true, and included among the criteria, are what may be called concepts (*prolēpsis*), which consist of "a recollection of what has often been presented from without . . . ." Man, therefore, must always cling to that "which was originally thought" in relation to every single "term" and which constitutes its background. Since the truth attested by each of the criteria is reflected in the *phainomena,* man must cling to these, employing them as "signs," and must "conjecture" whatever "does not appear." With the use of signs and conjecture, however, the level of judgment is reached, and thought is well advanced into that sphere in which error is possible, a state that begins as soon as single terms are tied into a proposition. Error, which consists of what "our judgment adds" to the evidence, can be of two types, one relative to what is not an object of experience, the other relative to what is such an object but for which the evidence is dubious. Each type has its own method of proof. Following the principles and methods of the "Canon," Epicurus arrived at an Atomism that, like that of the ancient naturalist Democritus, taught that the atoms, the void space in which they move, and the worlds are all infinite. But in contrast to Democritus, who had followed the deductive route of the intellect, considering the knowledge of the senses to be spurious, Epicurus, following an inductive route, assigned truth to sensation and reduced the intellect to it. On the basis of the totality of problems as Aristotle posed them in his "Physics," Epicurus modified entirely the mechanical theory of causes and of motion found in Democritus and added the concept of a natural necessity, which he called nature, and that of free causality, which alone could explain the freedom of motion of man and animals. For this purpose he distinguished three forms of motion in the atoms: a natural one of falling in a straight line, owing to their weight; a forced one due to impacts; and a free motion of declination, or swerving from a straight line. Secondly, he made finite the number of forms of the atoms in order to limit the number of sensible qualities, since each form begets a distinctive quality, and he taught a mathematical as well as a physical Atomism. Lest an infinity of sensible qualities be generated, however, by an infinity of aggregations (if not of atomic kinds), Epicurus developed, from just this concept of infinity, the law of universal equilibrium of all the forces, or "isonomy." Upon it, enclosing the events in a circle, he founded a theory of cyclic returns.

Natural necessity and free causality

As part of his "Physics," Epicurus' psychology held that the soul must be a body. It is made of very thin atoms of four different species—motile, quiescent, igneous, and ethereal—the last, thinnest and the most mobile of all, serving to explain sensitivity and thought. Thus constituted, the soul is, from another perspective, bipartite: in part distributed throughout the entire body and in part collected in the chest. The first part is the locus of sensations and of the physical affects of pain and pleasure; the second (entirely dissociated from the first) is the *psychē* par excellence—the seat of thought, emotions, and will. Thought is due not to the transmission of sense motion but to the perception of images constituted by films that continuously issue from all bodies and, retaining their form, arrive at the *psychē* through the pores. The full autonomy and freedom of the *psychē* is assured, as, with an act of apprehension, it seizes at every moment the images it needs, meanwhile remaining master of its own feelings.

The two kinds of pleasure

The object of ethics is to determine the end and the means necessary to reach it. Taking his cue from experience, Epicurus looked to the animal kingdom for his answer. He concluded from this cue that the chief end is pleasure. He distinguished two kinds—a "kinetic" pleasure of sense and a "static" pleasure, consisting in the absence of pain—and taught that the pleasure of sense is good, though it is not good merely as motion but rather as a motion favourable to the nature of the receiving sense organ. In essence, pleasure is the equilibrium of the being with itself, existing wherever there is no pain.

Epicurus concluded that "freedom from pain in the body and from trouble in the mind" is the ultimate aim of a happy life. The damages and the advantages following the realization of any desire must be measured in a calculus in which even pain must be faced with courage if the consequent pleasure will be of longer duration.

Having thus given order to his life, however, the wise man must also provide himself with security. This he achieves in two ways—by reducing his needs to a minimum and withdrawing, far from human competition and from the noise of the world, to "live hidden"; and by adding the private compact of friendship to the public compact from which laws arise. To be sure, friendship stems from utility; but, once born, it is desirable in itself. Epicurus then added that "for love of friendship one has even to put in jeopardy love itself"; for every existence, being alone, needs the other. "To eat and drink without a friend," he wrote, "is to devour like the lion and the wolf." Thus, the utility sublimates itself and changes into love. But as every love is intrepid, the wise man, "if his friend is put to torture, suffers as if he himself were there" and, if necessary, "will die for his friend." Thus, into the bloody world of his time, Epicurus could launch the cry: "Friendship runs dancing through the world bringing to us all the summons to wake and sing its praises."

The two great fears

If man's unhappiness stemmed only from his own vain desires and from worldly dangers, this wisdom, founded upon prudence alone, would suffice. But besides these sources of unhappiness there are two great fears, fear of death and fear of the gods. If science, however, is effective in revealing the bounds of desire and (as already seen) in quelling the fear of the gods, it can also allay the fear of death. Regarding the soul as a body within another body, science envisions it as dissolving when the body dissolves. Death, then, "is nothing to us, so long as we exist, death is not with us; but when death comes, then we do not exist." But death is feared not only for what may be awaiting man in the beyond but also for itself. "I am not afraid of being dead," said the comic Epicharmus of Cos: "I just do not want to die." The very idea of not existing instills a fear that Epicurus considered to be the cause of all the passions that pain the soul and disorder men's lives. Against it Epicurus argued that if pleasure is perfect within each instant and "infinite time contains no greater pleasure than limited time, if one measures by reason the limits of pleasure," then all desire of immortality is vain. Thus, Epicurus' most distinguished pupil, Metrodorus of Lampsacus, could exclaim, "*bebiōtai*" ("I have lived"), and this would be quite enough. He who has conquered the fear of death can also despise pain, which "if it is long lasting is light, and if it is intense is short" and brings death nearer. The wise man has only to replace the image of pain present in the flesh with that of blessings enjoyed, and he can be happy even "inside the bull of Phalaris." The most beautiful example was set by Epicurus at the moment of his death:

> A happy day is this on which I write to you . . . . The pains which I feel . . . could not be greater. But all of this is opposed by the happiness which the soul experiences, remembering our conversations of a bygone time.

The ultimate concentration of all his wisdom is the *Tetrapharmacon,* preserved by Philodemus: "The gods are not to be feared. Death is not a thing that one must fear. Good is easy to obtain. Evil is easy to tolerate."

The Epicurean school. Epicurus' successor in the direction of the Garden was Hermarchus of Mitylene, and he was succeeded in turn by Polystratus, who was the last survivor to have heard Epicurus. Superior to both, however, were Metrodorus and Colotes, against whom a small work by Plutarch was directed. Among the Epicureans of the 2nd century BC, mention must be made of Demetrius of Lacon, of whose works some fragments remain, and Apollodorus, who wrote more than 400 books. Much was also written by his disciple Zeno of Sidon, who was heard by Cicero in 79 BC in Athens. After Zeno, there were Phaedrus, also a teacher of Cicero, who was in Rome in 90 BC, and Patro, the head of the school until 51 BC. Already famous as an epigram writer was Philodemus of Gadara (born 110 BC). In the papyri of Herculaneum, comprising the effects of Philodemus' library, there are sizable remains of almost all of his numerous works. Epicureanism had already been introduced in Rome, in the 2nd century BC. The first person to spread its doctrines in Latin prose was

Lucretius' *De rerum natura*

a certain Amafinius. At the time of Cicero, Epicureanism was in fact the philosophy in vogue; and the number of Romans subscribing to it was, according to Cicero, very large. Among the greatest was Titus Lucretius Carus (*c.* 95–55 BC), who, in the poem *De rerum natura* ("On the Nature of Things"), left an almost complete and amazingly precise exposition of Epicurus' "Physics." The extent to which Epicurus was still popular in the 1st century after Christ is demonstrated by Seneca, who cited and defended him. To the 2nd century AD belongs Diogenes of Oenoanda, who carved Epicurus' works on a portico wall. In the same century should perhaps be mentioned Diogenianus, fragments of whose polemic against the Stoic Chrysippus are found in the church historian Eusebius. Also Epicurean, between the 4th and 5th centuries, was the epigrammatist Palladas.

On account of its dogmatic character and its practical end, the philosophy of Epicurus was not subject to development, except in the polemic and in its application to themes that Epicurus either had treated briefly or had never dealt with at all. To be aware of this, it is sufficient to run through what remains of the representatives of his school and particularly of the works of Philodemus of Gadara. Epicurus' philosophy remained essentially unchanged. Once truth has been found, it requires no more discussion, particularly when it completely satisfies the end toward which man's nature tends. The main thing is to see this end; all of the rest comes by itself, and there is no longer anything to do but follow Epicurus, "liberator" and "saviour," and to memorize his "oracular words."

Epicureanism and egoism in modern philosophy. In the Middle Ages Epicurus was known through Cicero and the polemics of the fathers. To be an Epicurean at the time of Dante meant to be one who denied Providence and the immortality of the soul. In the 15th century, the notable humanist Lorenzo Valla—following brief hints by Petrarch—wrote, in the dialogue *De voluptate* (1431; "On Pleasure"), the first modern defense of the ethics of Epicurus, maintaining that the true good is pleasure and not virtue but concluding that the supreme pleasure is that which awaits man in heaven, which even the Bible calls *paradisum voluptatis*. In the 16th century, in terms of attitude and direction of thought, the first two great Epicureans were Michel de Montaigne in France and Francesco Guicciardini in Italy. Epicurean in everything, as man and as poet, was the early 16th-century classicist Ludovico Ariosto. But not until the 17th-century Provençal abbot Pierre Gassendi was the system of Epicurus to rise again in its entirety—this time, however, by approaching truth through faith. Gassendi in 1649 wrote a commentary on a book by the 3rd-century (AD) biographer Diogenes Laërtius. This comment, called the *Syntagma philosophiae Epicuri* ("Treatise on Epicurean Philosophy"), was issued posthumously at The Hague 10 years later. At the same time, in England, Thomas Hobbes, a friend of Gassendi, took up again the theory of pleasure and interpreted it in a dynamic sense, which was therefore closer to the doctrine of the ancient Cyrenaics. Starting from the premise that, in the natural state, "man is a wolf to man," he concluded that peace, without which there is no happiness, cannot be guaranteed by anything but force, and that this force must be relinquished, by common agreement, to the power of only one.

Gassendi's revived Epicureanism

During the 17th and 18th centuries, the European nation in which Epicureanism was most active was France, where its representatives were called libertines, among them moralists such as François, duc de La Rochefoucauld, and Charles de Saint-Évremonde; scientists such as Julien de La Mettrie, who believed that man could be explained as a machine, Claude-Adrien Helvetius, who reduced the ethic of the useful to a form of experimental science but who put public above private well-being, and Paul Henri Dietrich, baron d'Holbach, who gave particular importance to the physics of the atoms. The purely sensistic conception of knowledge had its most thoroughgoing theoretician in Étienne de Condillac. In England, Adam Smith, developing the ethical concepts of Hume (founded on sympathy), surmounted the egoism that is the basis of every act by using the principle of the impartial observer invoked to sympathize with one or another of the antagonists. After him, the jurist Jeremy Bentham, eliminating sympathy, reduced ethics to the pure calculus of the useful, which—in an entirely Epicurean formula—he defined as a "moral arithmetic." In the Epicurean stream lay also the Utilitarianism of the 19th century, of which the greatest representative was John Stuart Mill.

Epicureanism in contemporary philosophy. In contemporary times, the interpretation of pleasure as a psychic principle of action was initiated by Gustav Fechner, the founder of psychophysics, and developed toward the end of the century by Freud on the psychoanalytic level of the unconscious.

Pleasure as a lifestyle

Epicureanism and egocentric hedonism have few faithful representatives among 20th-century philosophers, though the viewpoint remains as a residue in some strains of popular thinking.

CRITICISM AND EVALUATION

In the first half of the 17th century, at a time when Pierre Gassendi was reviving atomistic Epicureanism, René Descartes, often called the founder of modern philosophy, offered arguments that tended to undercut Atomism. Reality is a plenum, he held, a complete fullness; there can be no such thing as a vacuous region, or the void of Atomism. Since matter is nothing but spatial extension, its only true properties are geometrical and dynamic. Because extension is everywhere, motion occurs not as a passage through emptiness, as Epicurus supposed, but as vortices, or "whirlpools," in which every motion sets up a broad area of movement extending indefinitely around itself—a view that has tended to be confirmed in contemporary gravitational and field theory.

Close to the heart of Epicureanism is the principle, which occurred also in Democritus, that denies that something can come from or be rooted in nothing. In a poem composed by an ancient monist, Parmenides of Elea (born *c.* 515 BC), this principle had been expressed in the two formulas: "Being cannot be Non-Being," and "Non-Being must be Non-Being." Though Epicurus had faithfully adhered to this principle almost throughout his system, he has been criticized for abandoning it at one point—in the swerves that he attributed to occasional atoms that take them aside from their normal paths. Epicurus abandoned the principle at this point in order to avoid espousing a physics that was inconsistent with the autonomy that he observed in the physical behaviour of men and animals. But to his Stoic critics, the swerves of the atoms were a scandal, since they implied that an event can occur without a cause. It has seldom been noted, however, that the swerve is merely a special case—a transposition into atomistic terms—of Aristotle's theory of accidents (*i.e.*, of properties that are not essential to the substances in which they occur), inasmuch as an accident, too, as Aristotle himself had stated (*Met.* I 3), is without a cause. Moreover, a similar view has been seriously advanced in modern times under the name of tychism by Charles Sanders Peirce, a philosopher of science.

To the Stoic charge that Epicurus lacked a doctrine of Providence (since he viewed the gods as being lazy), Epicurus answered that "mythical gods are preferable to the fate" posited by the Stoics. It has been suggested that he might equally well have added that the "immobile Prime Mover" of Aristotle's theology was hardly less lazy than Epicurus' gods. Epicurus' way of phrasing this issue, however, must not obscure the fact that the problem of Providence and its secularization has been a crucial one in the philosophy of history ever since ancient times.

Critique of egoistic hedonism

The effort of Epicurus to reduce the good to pleasure reflects the only criterion to which he would entrust himself, the "evidence of those passions immediately present," which give man the word of Nature. In the argument of psychological hedonism, here implied, the Epicurean holds that men as a matter of fact do take satisfaction in pleasure and decry pain, and he argues then to an egoistic ethical hedonism that identifies the (objective) good with pleasure. Most moralists, however, have felt that a thoroughgoing psychological hedonism cannot be defended; that desire is often, as a matter of fact, directed toward

an object with no thought at all about the pleasure that it will bring; that a mother's impulse to save her young from danger is more fundamental than any pleasure involved (which usually comes only afterward); that the tendency of a child to imitate his parents can be, in fact, quite painful; and that, as a 19th-century Utilitarian, Henry Sidgwick, has argued, in what he called the "hedonistic paradox," one of the most ineffective ways to achieve pleasure is to deliberately seek it out.

Some scholars have even argued that an Epicurean egoistic hedonism, however foresighted it may be, must logically be self-defeating. If the view is universalized, the egoist must advocate the maximization of his enemy's pleasure as well as of his own, which can lead to actions painful to himself. In consequence, the entire branch of ethics that covers the advising or judging of other agents is banned from consideration, and it may be questioned whether such a view can comprise an ethic at all.

On the other hand, it has been argued that man is subject to antinomies, or contradictions, that no system can escape; there are dimensions in his nature that transcend the rational level. Thus, whatever its rational credentials may be, Epicureanism, as an attitude toward life that was theorized in its purest form by Epicurus, nonetheless remains one of the important forms that human behaviour has often assumed; and, at its best, it has achieved a type of asceticism that, even in retirement and solitude, does not negate company but welcomes it, finding the purest joys of life in the unique richness of human encounters.

(C.D.)

Platonism

Since Plato refused to write any formal exposition of his own metaphysic, our knowledge of its final shape has to be derived from the statements of Aristotle, which are confirmed by scanty remains of the earliest Platonists preserved in the Neoplatonist commentaries on Aristotle. These statements can, unfortunately, only be interpreted conjecturally.

ARISTOTLE'S ACCOUNT OF PLATONISM

According to Aristotle (*Metaphysics* i, 987 *b* 18–25) Plato's doctrine of forms was, in its general character, not different from Pythagoreanism, the forms being actually called numbers. The two points on which Aristotle regards Plato as disagreeing with the Pythagoreans are (1) that, whereas the Pythagoreans said that numbers have as their constituents the unlimited and the limit, Plato taught that the forms have as constituents the one and the great and small; and (2) that, whereas the Pythagoreans had said that things are numbers, Plato intercalated between his forms (or numbers) and sensible things an intermediate class of mathematicals. It is curious that in connection with the former difference Aristotle dwells mainly on the substitution of the "duality of the great and small" for the "unlimited," not on the much more significant point that the one, which the Pythagoreans regarded as the simplest complex of unlimited and limit, is treated by Plato as itself the element of limit. He further adds that the "great and small" is, in his own technical terminology, the matter, the one, the formal constituent, in a number.

Platonic number theory

If we could be sure how much of the polemic against number-forms in *Metaphysics* xiii–xiv is aimed directly at Plato, we might add considerably to this bald statement of his doctrine, but unluckily it is certain that much of the polemic is concerned with the teaching of Speusippus and Xenocrates. It is not safe, therefore, to ascribe to Plato statements other than those with which Aristotle explicitly credits him. We have then to interpret, if we can, two main statements: (1) the statement that the forms are numbers; and (2) the statement that the constituents of a number are the great and small and the one.

Light is thrown on the first statement if we recall the corpuscular physics of the *Timaeus* and the mixture of the *Philebus*. In the *Timaeus*, in particular, the behaviour of bodies is explained by the geometrical structure of their corpuscles, and the corpuscles themselves are analyzed into complexes built up out of two types of elementary triangle, which are the simplest elements of the narrative of Timaeus. Now a triangle, being determined in everything but absolute magnitude by the numbers that express the ratio of its sides, may be regarded as a triplet of numbers. If we remember then, that the triangles determine the character of bodies and are themselves determined by numbers, we may see why the ultimate forms on which the character of nature depends should be said to be numbers and also what is meant by the mathematicals intermediate between the forms and sensible things. According to Aristotle, these mathematicals differ from forms because they are many, whereas the form is one, from sensible things in being unchanging. This is exactly how the geometer's figure differs at once from the type it embodies and from a visible thing. There is, for example, only one type of triangle whose sides have the ratios 3:4:5, but there may be as many pure instances of the type as there are triplets of numbers exhibiting these ratios; and again, the geometrical triangles that are such pure instances of the type, unlike sensible three-sided figures, embody the type exactly and unchangingly. A mathematical physicist may thus readily be led to what seems to be Plato's view that the relations of numbers are the key to the whole mystery of nature, as is actually said in the *Epinomis* (990 *e*).

Further departure from Pythagoreanism

We can now, perhaps, see the motive for the further departure from Pythagoreanism. It is clear that the Pythagorean parallelism between geometry and arithmetic rested upon the thought that the point is to spatial magnitude what the number 1 is to number. Numbers were thought of as collections of units, and volumes, in like fashion, as collections of points; that is, the point was conceived as a minimum volume. As the criticisms of Zeno showed, this conception was fatal to the specially Pythagorean science of geometry itself, since it makes it impossible to assert the continuity of spatial magnitude. (This, no doubt, is why Plato, as Aristotle tells us, rejected the notion of a point as a fiction.)

There is also a difficulty about the notion of a number as a collection of units, which must have been forced on Plato's attention by the interest in irrationals that is shown by repeated allusions in the dialogues, as well as by the later anecdotes that represent him as busied with the problem of doubling the cube or finding two mean proportionals. Irrational square and cube roots cannot possibly be reached by any process of forming collections of units, and yet it is a problem in mathematics to determine them, and their determination is required for physics (*Epinomis* 990 *c*–991 *b*).

This is sufficient to explain why it is necessary to regard the numbers that are the physicist's determinants as themselves determinations of a continuum (a great and small) by a limit and why at the same time the one can no longer be regarded as a blend of unlimited and limit but must be itself the factor of limit. (If it were the first result of the blending, it would reappear in all the further blends; all numbers would be collections of one and there would be no place for the irrationals.) There is no doubt that Plato's thought proceeded on these general lines. Aristotle tells us that he said that numbers are not really addable (*Metaphysics* xiii, 1083 *a* 34), that is, that the integer series is not really made by successive additions of 1; and the *Epinomis* is emphatic on the point that, contrary to the accepted opinion, surds are just as much numbers as integers. The underlying thought is that numbers are to be thought of as generated in a way that will permit the inclusion of rationals and irrationals in the same series. In point of fact there are logical difficulties that make it impossible to solve the problem precisely on these lines. It is true that mathematics requires a sound logical theory of irrational numbers and, again, that an integer is not a collection of units; it is not true that rational integers and real numbers form a single series.

The Platonic number theory was inspired by thoughts that have since borne fruit abundantly but was itself premature. We learn partly from Aristotle, partly from notices preserved by his commentators, that, in the derivation of the integer series, the even numbers were supposed to be generated by the dyad that doubles whatever it lays hold of, the odd numbers in some way by the one that limits

or equalizes, but the interpretation of these statements is, at best, conjectural. In the statement about the dyad there seems to be some confusion between the number 2 and the indeterminate dyad, another name for the continuum also called the great and small, and it is not clear whether this confusion was inherent in the theory itself or has been caused by Aristotle's misapprehension.

Nor, again, is it at all certain exactly what is meant by the operation of equalizing ascribed to the one. It would be improper here to propound conjectures that our space will not allow us to discuss. A collection and examination of the available evidence is given by L. Robin in his *Théorie platonicienne des idées et des nombres d'après Aristote* (1908), and an admirable exposition of the significance of the problem of the irrational for Plato's philosophy by G. Milhaud in *Les Philosophes-géomètres de la Grèce, Platon et ses prédécesseurs,* new ed. (1934). For a conjectural interpretation see A.E. Taylor, "Forms and Numbers," in *Mind,* new series, vol. 35 and 36 (1926, 1927).

THE ACADEMY AFTER PLATO: THE RISE OF NEOPLATONISM

Plato's Academy continued to exist as a corporate body down to AD 529, when the emperor Justinian, in his zeal for Christian orthodoxy, closed the schools of Athens and appropriated their emoluments. Plato's greatest scholar, Aristotle, had finally gone his own way and organized a school of his own in the Lyceum, claiming that he was preserving the essential spirit of Platonism while rejecting the difficult doctrine of the forms. The place of official head of the Academy was filled first by Speusippus, Plato's nephew (*c.* 347–339 BC), then by Xenocrates (*c.* 339–314 BC). Under Arcesilaus (*c.* 276–241 BC) the Academy began its long-continued polemic against the sensationalist dogmatism of the Stoics, which accounts both for the tradition of later antiquity that dates the rise of a New (some said Middle) and purely skeptical Academy from Arcesilaus and for the 18th-century associations of the phrase "academic philosophy."

In the 1st century BC the most interesting episode in the history of the school is the quarrel between its president, Philo of Larissa, and his scholar Antiochus of Ascalon, of which Cicero's *Academica* is the literary record. Antiochus, who had embraced Stoic tenets, alleged that Plato had really held views indistinguishable from those of Zeno of Citium and that Arcesilaus had corrupted the doctrine of the Academy in a skeptical sense. Philo denied this. The gradual rapprochement between Stoicism and the Academy is illustrated from the other side by the work of Stoic scholars such as Panaetius of Rhodes and Poseidonius of Apamea, who commented on Platonic dialogues and modified the doctrines of their school in a Platonic sense.

The history of the Academy after Philo is obscure, but since the late 1st century AD we meet with a popular literary Platonism of which the writings of Plutarch are the best example. This popular Platonism insists on the value of religion, in opposition to Epicureanism, and on the freedom of the will and the reality of human initiative, in opposition to the Stoic determinism; a further characteristic feature, wholly incompatible with the genuine doctrine of Plato, is the notion that matter is inherently evil and the source of moral evil.

The Neoplatonism of Plotinus

Genuine Platonism was revived in the 3rd century AD, in Rome, and independently of the Academy, by Plotinus. His Neoplatonism represents a real effort to do justice to the whole thought of Plato. Two aspects of Plato's thought, however, in the changed conditions inevitably fell into the background: the mathematical physics, and the politics. The 3rd century AD had no understanding for the former, and the Roman Empire under a succession of military chiefs no place for the latter. The doctrine of Plotinus is Platonism seen through the personal temperament of a saintly mystic and with the *Symposium* and the teaching of the *Republic* about the form of good always in the foreground. Plotinus lived in an atmosphere too pure for sectarian polemic, but in the hands of his successors Neoplatonism was developed in conscious opposition to Christianity. Porphyry, his disciple and biographer, was the most formidable of the anti-Christian controversialists; in the next century, "Platonists" were among the allies and counsellors of the emperor Julian in his attempts to invent a Hellenic counterpart to Christianity.

Early in the 5th century, Neoplatonism flourished for a short time in Alexandria (which disgraced itself by the murder of Hypatia in 415) and captured the Athenian Academy itself, where its last great representative was the acute Proclus (AD 410–485). The latest members of the Academy, under Justinian, occupied themselves chiefly with learned commentaries on Aristotle, of which those of Simplicius are the most valuable. The doctrine of the school itself ends with Damascius in mystical agnosticism.

INFLUENCE OF PLATONISM ON CHRISTIAN THOUGHT

Traces of Plato are probably to be detected in the Alexandrian *Wisdom of Solomon;* the thought of the Alexandrian Jewish philosopher and theologian Philo, in the 1st century AD, is at least as much Platonic as Stoic. There are, perhaps, no certain marks of Platonic influence in the New Testament, but the earliest apologists (Justin, Athenagoras) appealed to the witness of Plato against the puerilities and indecencies of mythology. In the 3rd century Clement of Alexandria and after him Origen made Platonism the metaphysical foundation of what was intended to be a definitely Christian philosophy. The church could not, in the end, conciliate Platonist eschatology with the dogmas of the resurrection of the flesh and the Last Judgment, but in a less extreme form the platonizing tendency was continued in the next century by the Cappadocians, notably St. Gregory of Nyssa, and passed from them to St. Ambrose of Milan. The main sources of the Platonism that dominated the philosophy of Western Christian theologians through the earlier Middle Ages, were, however, Augustine, the greatest thinker among the western Fathers, who had been profoundly influenced by Plotinus read in a Latin version before his conversion to Christianity; and Boethius, whose wholly Platonist vindication of the ways of Providence in his *De Consolatione Philosophiae* was the favourite serious book of the Middle Ages.

Augustine and Boethius

A further powerful influence was exerted by the writings of the Pseudo-Dionysius the Areopagite. These works are, in fact, an imperfectly Christianized version of the speculations of Proclus and cannot date before the end of the 5th century AD at the earliest, but they enjoyed an immense authority based on their attribution to an immediate convert of St. Paul. After their translation into Latin in the 9th century by John Scotus Erigena, they became popular in the West.

Apart from this theological influence, Plato dominated the thought of the earlier Renaissance that dates from the time of Charlemagne in another way. Since the West possessed the philosophical writings of Cicero, with the Neoplatonic comment of Macrobius on the *Somnium Scipionis,* as well as the Latin translation of the first two-thirds of the *Timaeus* by Chalcidius, with his commentary on the text, and versions, also, at least of the *Phaedo* and of the *Meno,* whereas nothing was known of the works of Aristotle except Latin versions of some of the logical treatises, the Middle Ages, between Charlemagne and the beginning of the 13th century, when the recovery of Aristotle's physics and metaphysics from Moors, Persians, and Jews began, was much better informed about Plato than about Aristotle; in particular, in the various encyclopaedias of this period, it is the *Timaeus* that forms the regular background.

The 13th century saw a change. Aristotle came to displace Plato as the philosopher, partly in consequence of the immediately perceived value of his strictly scientific works as a storehouse of well-digested natural facts, partly from the brilliant success of the enterprise carried through by St. Thomas Aquinas, the reconstruction of philosophical theology on an Aristotelian basis. Plato is, however, by no means supplanted in the Thomist system; the impress of Augustine on Western thought has been far too deep for that. Augustine's "exemplarism," that is, the doctrine of forms in the version, ultimately derived from Philo of Alexandria, which makes the forms creative thoughts of God, is an integral part of the Thomist metaphysics, though it is now denied that the exemplars are themselves

cognizable by the human intellect, which has to collect its forms, as best it can, from the data of sense.

Directly or through Augustine, the influence of Plato, not only on strictly philosophic thought but also on popular ethics and religion, has repeatedly come to the front in ages of general spiritual requickening and shows no signs of being on the wane.

Two revivals in particular are famous. The first is that of the 16th century, marked by the Latin translation of Marsilio Ficino and the foundation of Lorenzo de' Medici's fantastic Florentine Academy. What was revived then was not so much the spirit of Plato as that of the least sober of the Neoplatonists; the influence of the revival was felt more in literature than in philosophy or morals, but in literature its importance may be measured by the mere mention of such names as Michelangelo, Sir Philip Sidney, and Edmund Spenser.

In the 17th century, Plato, seen chiefly through the medium of Plotinus, supplied the inspiration of a group of noble thinkers who were vindicating a more inward morality and religion against the unspiritual secularism and Erastianism of Hobbes: namely the so-called Cambridge Platonists, Benjamin Whichcote, Henry More, Ralph Cudworth, and John Smith. In the 20th century, on the one hand A.N. Whitehead tried to work out a philosophy of the sciences that confessedly connected itself with the ideas of the *Timaeus;* and on the other the rise of totalitarian governments produced a number of publications confronting Plato with the theories (Communist, Fascist, etc.) inherent in their policies. Neo-Kantianism, Existentialism, and analytical philosophy produced their own interpretations of Plato. (A.E.Ta./Pp.M.)

The "Cambridge Platonists"

Pythagoreanism

The philosophical school and religious brotherhood known as Pythagoreanism is believed to have been founded by Pythagoras of Samos, who settled in Croton in southern Italy about 525 BC.

GENERAL FEATURES OF PYTHAGOREANISM

The character of the original Pythagoreanism is controversial, and the conglomeration of disparate features that it displayed is intrinsically confusing. Its fame rests, however, on some very influential ideas, not always correctly understood, that have been ascribed to it since antiquity. These ideas include those of (1) the metaphysic of number and the conception that reality, including music and astronomy, is, at its deepest level, mathematical in nature; (2) the use of philosophy as a means of spiritual purification; (3) the heavenly destiny of the soul and the possibility of its rising to union with the divine; (4) the appeal to certain symbols, sometimes mystical, such as the *tetraktys,* the golden section, and the harmony of the spheres (to be discussed below); (5) the Pythagorean theorem; and (6) the demand that members of the order shall observe a strict loyalty and secrecy.

By laying stress on certain inner experiences and intuitive truths revealed only to the initiated, Pythagoreanism seems to have represented a soul-directed subjectivism alien to the mainstream of Pre-Socratic Greek thought centring on the Ionian coast of Asia Minor (Thales, Anaximander, Anaxagoras, and others), which was preoccupied with determining what the basic cosmic substance is.

In contrast with such Ionian naturalism, Pythagoreanism was akin to trends seen in mystery religions and emotional movements, such as Orphism, which often claimed to achieve through intoxication a spiritual insight into the divine origin and nature of the soul. Yet there are also aspects of it that appear to have owed much to the more sober, "Homeric" philosophy of the Ionians. The Pythagoreans, for example, displayed an interest in metaphysics (the nature of Being), as did their naturalistic predecessors, though they claimed to find its key in mathematical form rather than in any substance. They accepted the essentially Ionian doctrines that the world is composed of opposites (wet-dry, hot-cold, etc.) and generated from something unlimited; but they added the idea of the imposition of limit upon the unlimited and the sense of a musical harmony in the universe. Again, like the Ionians, they devoted themselves to astronomical and geometrical speculation. Combining, as it does, a rationalistic theory of number with a mystic numerology and a speculative cosmology with a theory of the deeper, more enigmatic reaches of the soul, Pythagoreanism interweaves Rationalism and irrationalism more inseparably than does any other movement in ancient Greek thought.

MAJOR CONCERNS AND TEACHINGS

The problem of describing Pythagoreanism is complicated by the fact that the surviving picture is far from complete, being based chiefly on a small number of fragments from the time before Plato and on various discussions in authors who wrote much later—most of whom were either Aristotelians or Neoplatonists (see below *History of Pythagoreanism*). In spite of the historical uncertainties, however, that have plagued searching scholars, the contribution of Pythagoreanism to Western culture has been significant and therefore justifies the effort, however inadequate, to depict what its teachings may have been. Moreover, the heterogeneousness of Pythagorean doctrines has been well documented ever since Heracleitus, a classic early 5th-century Greek philosopher who, scoffing at Pythagoras' wide-ranging knowledge, said that it "does not teach one to have intelligence." There probably never existed a strictly uniform system of Pythagorean philosophy and religious beliefs, even if the school did have a certain internal organization. Pythagoras appears to have taught by pregnant, cryptic *akousmata* ("something heard") or *symbola.* His pupils handed these on, formed them partly into *Hieroi Logoi* ("Sacred Discourses"), of which different versions were current from the 4th century on, and interpreted them according to their convictions.

Religion and ethics. The belief in the transmigration of souls provided a basis for the Pythagorean way of life. Some Pythagoreans deduced from this belief the principle of "the kinship of all beings," the ethical implications of which were later stressed in 4th-century speculation. Pythagoras himself seems to have claimed a semidivine status in close association with the superior god Apollo; he believed that he was able to remember his earlier incarnations and, hence, to know more than others knew. Recent research has emphasized shamanistic traits deriving from the ecstatic cult practices of Thracian medicine men in the early Pythagorean outlook. The rules for the religious life that Pythagoras taught were largely ritualistic: refrain from speaking about the holy, wear white clothes, observe sexual purity, do not touch beans, and so forth. He seems also to have taught purification of the soul by means of music and mental activity (later called philosophy) in order to reach higher incarnations. "To be like your Master" and so "to come nearer to the gods" was the challenge that he imposed on his pupils. Salvation, and perhaps ultimate union with the divine cosmos through the study of the cosmic order, became one of the leading ideas in his school.

Significance of belief in transmigration of souls

The advanced ethics and political theories sometimes ascribed to Pythagoreanism may to some extent reflect ideas later developed in the circle of Archytas, the leading 4th-century Pythagorean. But a picture current among the Peripatetics (the school founded by Aristotle) of Pythagoras as the educator of the Greeks, who publicly preached a gospel of humanity, is clearly anachronistic. Several of the Peripatetic writers, Aristoxenus, Dicaearchus, and Timaeus, seem to have interpreted some principles—properly laid down only for esoteric use in the brotherhood—as though these applied to all mankind: the internal loyalty, modesty, self-discipline, piety, and abstinence required by the secret doctrinal system; the higher view of womanhood reflected in the admission of women to the school; a certain community of property; and perhaps the drawing of a parallelism between the macrocosm (the universe) and the microcosm (man), in which (for instance) the Pythagorean idea that the cosmos is an organism was applied to the state, which should thus mix monarchy, oligarchy, and democracy into a harmonic whole—these were all universalized.

Metaphysics and number theory. According to Aristot-

le, number speculation is the most characteristic feature of Pythagoreanism. Things "are" number, or "resemble" number. To many Pythagoreans this concept meant that things are measurable and commensurable or proportional in terms of number—an idea of considerable significance for Western civilization. But there were also attempts to arrange a certain minimum number of pebbles so as to represent the shape of a thing—as, for instance, stars in a constellation that seem to represent an animal. For the Pythagoreans even abstracted things "have" their number: "justice" is associated with the number four and with a square, "marriage" with the number five, and so on. The psychological associations at work here have not been clarified.

The harmony of the cosmos. The sacred decad in particular has a cosmic significance in Pythagoreanism: its mystical name, *tetraktys* (meaning approximately "fourness"), implies $1 + 2 + 3 + 4 = 10$; but it can also be thought of as a "perfect triangle," as in Figure 1.



Figure 1: The *Tetraktys* (see text).

Speculation on number and proportion led to an intuitive feeling of the *harmonia* ("fitting together") of the *kosmos* ("the beautiful order of things"); and the application of the *tetraktys* to the theory of music (see below *Music*) revealed a hidden order in the range of sound. Pythagoras may have referred, vaguely, to the "music of the heavens," which he alone seemed able to hear; and later Pythagoreans seem to have assumed that the distances of the heavenly bodies from the Earth somehow correspond to musical intervals—a theory that, under the influence of Platonic conceptions, resulted in the famous idea of the "harmony of the spheres." Though number to the early Pythagoreans was still a kind of cosmic matter, like the water or air proposed by the Ionians, their stress upon numerical proportions, harmony, and order comprised a decisive step toward a metaphysic in which form is the basic reality.

The doctrine of opposites. From the Ionians, the Pythagoreans adopted the idea of cosmic opposites, which they—perhaps secondarily—applied to their number speculation. The principal pair of opposites is the limit and the unlimited; the limit (or limiting), represented by the odd (3,5,7, . . .), is an active force effecting order, harmony, "cosmos," in the unlimited, represented by the even. All kinds of opposites somehow "fit together" within the cosmos, as they do, microcosmically, in an individual man and in the Pythagorean society. There was also a Pythagorean "table of ten opposites," to which Aristotle has referred—limit–unlimited, odd–even, one–many, right–left, male–female, rest–motion, straight–curved, light–darkness, good–evil, and square–oblong. The arrangement of this table reflects a dualistic conception, which was apparently not original with the school, however, or accepted by all of its members.

The Pythagorean number metaphysic was also reflected in its cosmology. The unit (1), being the starting point of the number series and its principle of construction, is not itself strictly a number; for, to be a number is to be even or odd, whereas, in the Pythagorean view, "one" is seen as *both* even and odd. This ambivalence applies, similarly, to the total universe, conceived as the One. There was also a cosmogonical theory (of cosmic origins) that explained the generation of numbers and number-things from the limiting-odd and the unlimited-even—a theory that, by stages unknown to scholars, was ultimately incorporated into Plato's philosophy in his doctrine of the derivation of sensed realities from mathematical principles.

Mathematics and science. Pythagorean thought was scientific as well as metaphysical and included specific developments in arithmetic and geometry, in the science of musical tones and harmonies, and in astronomy.

Arithmetic. Early Pythagorean achievements in mathematics are unclear and largely disputable, and the following is, therefore, a compromise between the widely divergent views of modern scholars.

Use of the *gnōmones*

In the speculation on odd and even numbers, the early Pythagoreans used so-called *gnōmones* (Greek: "carpenter's squares"). Judging from Aristotle's account, gnomon numbers, represented by dots or pebbles, were arranged in the manner shown in Figure 2. If a series of odd numbers

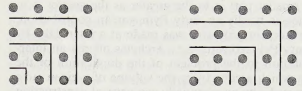


Figure 2: Gnomons of Pythagorean number theory (see text).

is put around the unit as gnomons, they always produce squares; thus, the members of the series 4, 9, 16, 25, . . . are "square" numbers. If even numbers are depicted in a similar way, the resulting figures (which offer infinite variations) represent "oblong" numbers, such as those of the series 2, 6, 12, 20 . . . . On the other hand, a triangle represented by three dots (as in the upper part of the *tetraktys*) can be extended by a series of natural numbers to form the "triangular" numbers 6, 10 (the *tetraktys*), 15, 21. . . . This procedure, which was, so far, Pythagorean, led later, perhaps in the Platonic Academy, to a speculation on "polygonal" numbers.

Probably the square numbers of the gnomons were early associated with the Pythagorean theorem (likely to have been used in practice in Greece, however, before Pythagoras), which holds that for a right triangle a square drawn on the hypotenuse is equal in area to the sum of the squares drawn on its sides; in the gnomons it can easily be seen, in the case of a 3,4,5–triangle for example (see Figure 3), that the addition of a square gnomon number

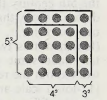


Figure 3: Gnomon for Pythagorean theorem. The marked off "carpenter's square"—comprising 3 groups of 3 dots each (3×3)—thus represents 3^2, which when added to 4^2 yields 5^2 (the total gnomon).

to a square makes a new square: $3^2 + 4^2 = 5^2$, and this gives a method for finding two square numbers the sum of which is also a square.

The problem of irrational numbers

Some 5th-century Pythagoreans seem to have been puzzled by apparent arithmetical anomalies: the mutual relationships of triangular and square numbers; the anomalous properties of the regular pentagon; the fact that the length of the diagonal of a square is incommensurable with its sides—*i.e.*, that no fraction composed of integers can express this ratio exactly (the resulting decimal is thus defined as irrational); and the irrationality of the mathematical proportions in musical scales. The discovery of such irrationality was disquieting because it had fatal consequences for the naive view that the universe is expressible in whole numbers; the Pythagorean Hippasus is said to have been expelled from the brotherhood, according to some sources even drowned, because he made a point of the irrationality.

In the 4th century, Pythagorizing mathematicians made a significant advance in the theory of irrational numbers, such as the-square-root-of-n ($\sqrt{n}$), n being any rational number, when they developed a method for finding progressive approximations to $\sqrt{2}$ by forming sets of so-called diagonal numbers.

Geometry. In geometry, the Pythagoreans cannot be credited with any proofs in the Euclidean sense. They were evidently concerned, however, with some speculation on geometrical figures, as in the case of the Pythagorean theorem, and the concept that the point, line, triangle, and tetrahedron correspond to the elements of the *tetraktys*, since they are determined by one, two, three, and four points, respectively. They possibly knew practical methods of constructing the five regular solids, but the theoretical basis for such constructions was given by non-Pythagoreans in the 4th century.

It is notable that the properties of the circle seem not to have interested the early Pythagoreans. But perhaps the tradition that Pythagoras himself discovered that the sum of the three angles of any triangle is equal to two right angles may be trusted. The idea of geometric proportions is probably Pythagorean in origin; but the so-called golden section—which divides a line at a point such

that the smaller part is to the greater as the greater is to the whole—is hardly an early Pythagorean contribution. Some advance in geometry was made at a later date, by 4th-century Pythagoreans; *e.g.,* Archytas offered an interesting solution to the problem of the duplication of the cube—in which a cube twice the volume of a given cube is constructed—by an essentially geometrical construction in three dimensions; and the conception of geometry as a "flow" of points into lines, of lines into surfaces, and so on, may have been contributed by Archytas; but on the whole the numerous achievements of non-Pythagorean mathematicians were in fact more conspicuous than those of the Pythagoreans.

Music. The achievements of the early Pythagoreans in musical theory are somewhat less controversial. The scientific approach to music, in which musical intervals are expressed as numerical proportions, originated with them, as did also the more specific idea of harmonic "means." At an early date they discovered empirically that the basic intervals of Greek music include the elements of the *tetraktys,* since they have the proportions 1:2 (octave), 3:2 (fifth), and 4:3 (fourth). The discovery could have been made, for instance, in pipes or flutes or stringed instruments: the tone of a plucked string held at its middle is an octave higher than that of the whole string: the tone of a string held at the ⅔ point is a fifth higher; and that of one held at the ¾ point is a fourth higher. Moreover, they noticed that the subtraction of intervals is accomplished by dividing these ratios by one another. In the course of the 5th century they calculated the intervals for the usual diatonic scale, the tone being represented by 9:8 (fifth minus fourth); *i.e.,* $3/2 \div 4/3$, and the semitone by 256:243 (fourth minus two tones); *i.e.,* $4/3 \div (9/8 \times 9/8)$. Archytas made some modification to this doctrine and also worked out the relationships of the notes in the chromatic (12-tone) scale and the enharmonic scale (involving such minute differences as that between A flat and G sharp, which on a piano are played by the same key).

Mathematical ratios of tones

Astronomy. In their cosmological views the earliest Pythagoreans probably differed little from their Ionian predecessors. They made a point of studying the stellar heavens; but—with the possible exception of the theory of musical intervals in the cosmos—no new contributions to astronomy can be ascribed to them with any degree of probability. Late in the 5th century, or possibly in the 4th century, a Pythagorean boldly abandoned the geocentric view and posited a cosmological model in which the Earth, Sun, and stars circle about an (unseen) central fire—a view traditionally attributed to the 5th-century Pythagorean Philolaus of Croton.

HISTORY OF PYTHAGOREANISM

Literary sources, especially Aristotelian

The life of Pythagoras and the origins of Pythagoreanism appear only dimly through a thick veil of legend and semihistorical tradition. The literary sources for the teachings of the Pythagoreans present extremely complicated problems. Special difficulties arise from the oral and esoteric transmission of the early doctrines, the profuse accumulation of tendentious legends, and the considerable amount of confusion that was caused by the split in the school in the 5th century BC. In the 4th century, Plato's inclination toward Pythagoreanism created a tendency—manifest already in the middle of the century in the works of his pupils—to interpret Platonic concepts as originally Pythagorean. But the radical skepticism as to the reliability of the sources shown by some modern scholars has on the whole been abandoned in recent research. It now seems possible to extract bits of reliable evidence from a wide range of ancient authors, such as Porphyry and Iamblichus (see below *Neo-Pythagoreanism*).

Most of these literary sources hark back ultimately to the environment of Plato and Aristotle; and here the importance of one of Aristotle's students has become obvious, viz., the musicologist and philosopher Aristoxenus, who in spite of his bias possessed firsthand information independent of the point of view of Plato's Academy. The role played by Dicaearchus, another of Aristotle's pupils, and by the Sicilian historian Timaeus, of the early 3rd century BC, is less clear. Recently, the reliability of Aristotle's account of Pythagoreanism has also been emphasized against the doubts that had been expressed by some modern scholars; but Aristotle's sources, in turn, hardly lead farther back than to the late 5th century (perhaps to Philolaus; see below *Two Pythagorean sects*). In addition, there are scattered hints in various early authors and in some not very substantial remains of 4th-century Pythagorean literature. The mosaic of reconstruction thus has to be to some extent subjective.

Four chief periods

Early Pythagoreanism. Within the ancient Pythagorean movement four chief periods can be distinguished: early Pythagoreanism, dating from the late 6th century BC and extending to about 400 BC; 4th-century Pythagoreanism; the Hellenistic trends; and Neo-Pythagoreanism, a revival that occurred in the mid-1st century AD and lasted for two and a half centuries.

Background. The background of Pythagoreanism is complex, but two main groups of sources can be distinguished. The Ionian philosophers (Thales, Anaximander, Anaximenes, and others) provided Pythagoras with the problem of a single cosmic principle, the doctrine of opposites, and whatever reflections of Oriental mathematics there are in Pythagoreanism; and from the technicians of his birthplace, the Isle of Samos, he learned to understand the importance of number, measurements, and proportions. Popular cults and beliefs current in the 6th century and reflected in the tenets of Orphism introduced him to the notions of occultism and ritualism and to the doctrine of individual immortality. In view of the shamanistic traits of Pythagoreanism, reminiscent of Thracian cults, it is interesting to note that Pythagoras seems to have had a Thracian slave.

Pythagorean communities. The school apparently founded by Pythagoras at Croton in southern Italy seems to have been primarily a religious brotherhood centred around Pythagoras and the cults of Apollo and of the Muses, ancient patron goddesses of poetry and culture. It became perhaps successively institutionalized and received different classes of esoteric members and exoteric sympathizers. The rigorism of the ritual and ethical observances demanded of the members is unparalleled in early Greece; in addition to the rules of life mentioned above, it is fairly well attested that secrecy and a long silence during the novitiate were required. The exoteric associates, however, were politically active and established a Crotonian hegemony in southern Italy. About 500 BC a coup by a rival party caused Pythagoras to take refuge in Metapontum, where he died.

During the early 5th century, Pythagorean communities, inspired by the original school at Croton, existed in many southern Italian cities, a fact that led to some doctrinal differentiation and diffusion. In the course of time the politics of the Pythagorean parties became decidedly antidemocratic. About the middle of the century a violent democratic revolution swept over southern Italy; in its wake, many Pythagoreans were killed, and only a few escaped, among them Lysis of Tarentum and Philolaus of Croton, who went to Greece and formed small Pythagorean circles in Thebes and Phlious.

Two Pythagorean sects. Little is known about Pythagorean activity during the latter part of the 5th century. The differentiation of the school into two main sects, later called *akousmatikoi* (Greek: *akousma,* "something heard," viz., the esoteric teachings) and *mathēmatikoi* (Greek: *mathēmatikos,* "scientific"), may have occurred at that time. The acousmatics devoted themselves to the observance of rituals and rules and to the interpretation of the sayings of the master; the "mathematics" were concerned with the scientific aspects of Pythagoreanism. Philolaus, who was rather a mathematic, probably published a summary of Pythagorean philosophy and science in the late 5th century.

Archytas and his school

4th-century Pythagoreanism. In the first half of the 4th century, Tarentum, in southern Italy, rose into considerable significance. Under the political and spiritual leadership of the mathematic Archytas, a friend of Plato, Tarentum became a new centre of Pythagoreanism, from which acousmatics—so-called Pythagorists who did not sympathize with Archytas—went out travelling as mendicant

ascetics all around the Greek-speaking world. The acousmatics seem to have preserved some early Pythagorean *Hieroi Logoi* and ritual practices. Archytas himself, on the other hand, concentrated on scientific problems, and the organization of his Pythagorean brotherhood was evidently less rigorous than that of the early school. After the 380s there was a give-and-take between the school of Archytas and the Academy of Plato, a relationship that makes it almost impossible to disentangle the original achievements of Archytas from joint involvements (but see above, *Geometry* and *Music*).

The Hellenistic Age. Whereas the school of Archytas apparently sank into inactivity after the death of its founder (probably after 350 BC), the Academics of the next generation continued "Pythagorizing" Platonic doctrines, such as that of the supreme One, the indefinite dyad (a metaphysical principle), and the tripartite soul. At the same time, various Peripatetics of the school of Aristotle, including Aristoxenus, collected Pythagorean legends and applied contemporary ethical notions to them. In the Hellenistic Age, the Academic and Peripatetic views gave rise to a rather fanciful antiquarian literature on Pythagoreanism. There also appeared a large and yet more heterogeneous mass of apocryphal writings falsely attributed to different Pythagoreans, as if attempts were being made to revive the school. The texts fathered on Archytas display Academic and Peripatetic philosophies mixed with some notions that were originally Pythagorean. Other texts were fathered on Pythagoras himself or on his immediate pupils, imagined or real. Some show, for instance, that Pythagoreanism had become confused with Orphism; others suggest that Pythagoras was considered a magician and an astrologist; there are also indications of Pythagoras "the athlete" and "the Dorian nationalist." But the anonymous authors of this pseudo-Pythagorean literature did not succeed in reestablishing the school, and the "Pythagorean" congregations formed in early imperial Rome seem to have had little in common with the original school of Pythagoreanism established in the late 6th century BC; they were ritualistic sects that adopted, eclectically, various occult practices.

Neo-Pythagoreanism. With the ascetic sage Apollonius of Tyana, about the middle of the 1st century AD, a distinct Neo-Pythagorean trend appeared. Apollonius studied the Pythagorean legends of the previous centuries, created and propagated the ideal of a Pythagorean life—of occult wisdom, purity, universal tolerance, and approximation to the divine—and felt himself to be a reincarnation of Pythagoras. Through the activities of Neo-Pythagorean Platonists, such as Moderatus of Gades, a pagan trinitarian, and the arithmetician Nicomachus of Gerasa, both of the 1st century AD, and, in the 2nd or 3rd century, Numenius of Apamea, forerunner of Plotinus (an epoch-making elaborator of Platonism), Neo-Pythagoreanism gradually became a part of the expression of Platonism known as Neoplatonism; and it did so without having achieved a scholastic system of its own. The founder of a Syrian school of Neoplatonism, Iamblichus, a pupil of Porphyry (who in turn had been a pupil of Plotinus), thought of himself as a Pythagorean sage and about AD 300 wrote the last great synthesis of Pythagoreanism, in which most of the disparate post-classical traditions are reflected. It is characteristic of the Neo-Pythagoreans that they were chiefly interested in the Pythagorean way of life and in the pseudoscience of number mysticism. On a more popular level, Pythagoras and Archytas were remembered as magicians. Moreover, it has been suggested that Pythagorean legends were also influential in guiding the Christian monastic tradition.

Amalgamation with Neoplatonism

Medieval and modern trends. In the Middle Ages the popular conception of Pythagoras the magician was combined with that of Pythagoras "the father of the quadrivium"; *i.e.*, of the more specialized liberal arts of the curriculum. From the Italian Renaissance onward, some "Pythagorean" ideas, such as the tetrad, the golden section, and harmonic proportions, became applied to aesthetics. To many Humanists, moreover, Pythagoras was the father of the exact sciences. In the early 16th century, Nicolaus Copernicus, who developed the view that the Earth revolves around the Sun, considered his system to be essentially Pythagorean or "Philolaic," and Galileo was called a Pythagorean. The 17th-century Rationalist G.W. Leibniz appears to have been the last great philosopher and scientist who felt himself to be in the Pythagorean tradition.

It is doubtful whether advanced modern philosophy has ever drawn from sources thought to be distinctly Pythagorean. Yet Platonic–Neoplatonic notions, such as the mathematical conception of reality or the philosopher's union with the universe and various mystical beliefs are still likely to be stamped as Pythagorean in origin. Even today a relatively uncritical admiration of Pythagoreanism is common.

EVALUATION

The history of the projection of Pythagoreanism into subsequent thought indicates how fertile some of its core concepts were. Plato is here the great catalyst; but it is possible to perceive behind him, however dimly, a series of Pythagorean ideas of paramount potential significance: the combination of religious esoterism (or exclusivism) with the germs of a new philosophy of mind, present in the belief in the progress of the soul toward the actualization of its divine nature and toward knowledge; stress upon harmony and order, and upon limit as the good; the primacy of form, proportion, and numerical expression; and in ethics, and emphasis upon such virtues as friendship and modesty. The fact that Pythagoras, to later ages, also became alternatively conceived of as a Dorian nationalist, a sportsman, an educator of the people, or a great magician is a more curious consequence of the productivity of his teaching.

(H.T.)

Realism

Understood in its broadest philosophical sense, Realism connotes any viewpoint that accords to the objects of man's knowledge an existence that is independent of whether he is perceiving or thinking about them. Though it may seem strange to the unphilosophical layman that the independent existence of objects "out there" should be questioned, the philosopher, faced with the many profound challenges that Idealists have posed against the independence of objects, knows that the problem of the existence of objects—whether in thought or in concrete form—is far from trivial.

The Idealist challenge

Clearly, Idealists have argued, musical tones such as middle C do not have existence as tones in the air; they appear, instead, to be qualities that the mind itself generates when the appropriate hair cells in the organ of Corti are stimulated. Nor does the colour purple have existence as a quality in the world outside of the mind; there can be, in fact, no such thing as a beam of pure (monochromatic) purple light inasmuch as purple is a unique kind of colour that is perceived when vibrations at the opposite extremes of the visual spectrum (red and violet) are mixed together in the same beam. At least in this one case, the colour seems created by the mind. But if this is so of purple, the Idealists ask, is it not true, also, of all colours? Similarly, under certain circumstances heat is felt as cold and rotation as oscillation. It is not surprising, therefore, that philosophers have asked what, if any, residual properties an object might have in and of itself after due allowance has been made for those qualities that the mind and perspective of the observer have imposed upon it; nor is it surprising that they have asked what, if anything, it would mean to insist on the objective existence of an object of which all of the qualities were mental. Realists, on the other hand, have held that, in spite of the foregoing considerations as proposed by the Idealists, there still remains a sense in which objects can have an existence that is independent of minds.

Realism exists, however, in several strikingly different versions: its objects may be, for example, either individual things (such as "the Moon"), or merely particular qualities of things (such as "roundness," "yellowness"), or species and genera of things (such as "moons," "planetary bodies"). In one way or another, however, whether it regards

things from the viewpoint of the things themselves or from that of the human activities related to them, Realism tends to stress some definite function of the independent existence of objects.

NATURE AND SCOPE OF REALISM

Realism and the problem of knowledge. One of the major problems confronting Realism involves the distinction between private, public, and so-called ontological objects.

Private, public, and ontological objects

A private object is a sheer datum (such as a perceived patch of yellow) taken purely as an uninterpreted item in the knower's own inner experience; a public object is one that the mind has projected into an objective conceptual frame of space and time shared in common with other minds, an object that the mind has constituted as a percept (such as the perceived Moon)—though it is still acknowledged to be in part mental (*e.g.,* its yellowness; or its visual size, which is larger when it is near the horizon); and an ontological object is the Kantian "thing-in-itself" (the Moon as it really is), which may as well consist of monads, of God's thoughts, of will, or of action, as of force and matter. Though Realists and Idealists both acknowledge that the knower transcends the private object, they make different assumptions about the relationship between the public and the ontological object: on the one hand, the Realist holds that the physical sphericity, yellowness, and hardness perceived (or perceivable) in the public object are in some degree actual properties of the ontological thing-in-itself; the Idealist, on the other hand, holds that the public object is merely a phenomenon, from which little can be inferred about the underlying *onta* (realities), least of all of their basic qualities, which are probably quite different from the roundness and hardness of the perceived object. The Idealist may, in fact, surmise that the nature of the *onta* is conveyed more faithfully in the fundamental mental tone of the public object—in the colours, feelings, and durations (which are of the nature of mind)—than in its specific material properties. (A third contender, that philosopher known as the metaphysical solipsist, would hold to the viewpoint that the ontological object does not exist at all.)

Similarly, if a particular thing regarded in its particularity (such as the Moon) is distinguished from a universal—*i.e.,* an entity comprising the essence of the thing (moonness)—that which it shares with all the other things of the same species or genus (as with the moons of Jupiter)—then a form of Realism can be defined as asserting the independent reality of universals, which it may even exalt above that of particulars.

Contrast with other views

Accordingly, Realism may be variously opposed to the tenets of other philosophical positions. As opposed to Nominalism, which denies that essences (or the specific and generic natures of things) have any reality at all (except as names), and conceptualism, which grants such universals reality only as concepts within the mind, Realism allows to the specific or generic nature of the thing a distinct existence in reality outside the mind. Against Idealism (see below), it asserts that the existence of sense objects (such as the perceived Moon) and that of their qualities is external to thought. In opposition to phenomenalism and sensationalism, which regard objects as comprising nothing more than private volleys or families of disconnected sense fragments, Realism grounds objects in real unified and enduring substances. Unlike conventionalism, a philosophy of science that regards scientific laws and theories as freely chosen constructs that are simply devised by the scientist for the purpose of describing reality, Realism holds that laws and theories have determined and real counterparts in things.

The term Realism first appeared early in the 19th century, though the adjective Realist dates from the late 16th century. These terms have been applied, however—often retroactively—to various systems that have arisen throughout history.

In its broadest scope, the term Realism has application in a number of distinct areas. In literature, art, and aesthetics, in law, and in philosophy, it emphasizes real existence or relation to it. The present article is concerned solely with Realism in the philosophical area.

Philosophical senses of Realism. Even within philosophy Realism has a wide range of applications. Though a definitely modern term, Realism is freely used today for tenets of the Greek and medieval epochs, as well as for the modern period.

Basic kinds of Realism. Among philosophical Realisms, two fundamentally different kinds can be distinguished: the Realism of natures and the Realism of things. In the Realism of natures, that which is viewed as having an existence external to the mind is an entity that, in some sense, is set apart in the world of things—an entity that is variously understood as the Form or Idea in which a thing participates, such as "manness" or "bedness" (Platonic Realism), as the essence or *to ti ēn einai,* "the 'what it is' of a thing" (Aristotelian Realism), or as its nature, either absolute, specific, or generic (medieval Realism or the Realism of universals), or, finally, as laws or theoretical models abstracted from scientific observations. In the Realism of things, on the other hand, that which is viewed as having an existence external to the mind is the total, concrete, and individual object of experience, which the Realist regards as retaining its chief properties at all times, even when left unseen. This Realism, too, can be variously conceived: the externality of the world, for example, can be regarded as simply and obviously given (commonsense Realism); the object itself, though external, can be viewed as the sole entity standing before the mind and grasped by it (neo-Realism); or the object can be conceived as, in some sense, duplicated, so that the mind directly encounters only a counterpart of the external object and not the object itself (critical Realism), a counterpart which was sometimes regarded as a representation of it (representational Realism).

Distinctions among the Realisms. As previously noted, the term Realism has been applied retroactively to the transcendence of the Platonic Forms or Ideas, to the extent that for Plato the natures of things have, in the ideas of them, an existence more real than that of sensible, individual things. Yet, from its emphasis on ideal as opposed to concrete existence, this Platonic doctrine would be classed as an Idealism instead of a Realism. In the parallel issue in Aristotelianism, the stand that the universals, or specific and generic natures, exist only in the mind but are nonetheless grounded in the real forms of things has been called a moderate Realism. Aristotle himself, however, vigorously denied that the universals have any substantiality (*Metaphysics,* Z: 13–14; 1038ᵇ8–1039ᵇ19), which clearly suggests that, for him, the universals have no existence independently of cognition; this tends, in this first context, to invalidate the designation Realism for the Aristotelian doctrine.

Realism of natures

Correspondingly, Realism is used to describe medieval views that allowed species and genera some kind of distinct existence outside of their conception by the mind. There it meant not only that individual men and individual animals and so on exist outside cognition but also that the specific nature of man and the generic nature of animal and the like have an existence of their own in the outside world. For Realism, objects "fall into" such categories as humanness, mountainness, and so on naturally. For its opponents, however, this is not always the case: thus, in terms of a modern illustration, graniteness—that which all the granite rocks share in common—does not exist except as an artificial category set up by the mind (conceptualism) because it merges by imperceptible gradations into diorite or felsite as its mineral composition and texture gradually change.

Yet in actual fact the various medieval doctrines do not fit neatly under these divisions. In the philosophies of several medieval Scholastics, for instance, both the particular thing and the universal are distinguished in one way or another from a third entity, the specific or generic nature taken absolutely in itself. This so-called absolute nature was given a "being of its own" by Avicenna, the early 11th-century Persian philosopher and physician, and—in the 13th century—by Henry of Ghent, an eclectic Christian Scholastic, and by the voluntarist John Duns Scotus, an important medieval Franciscan Scholastic, who gave the absolute nature a reality that was distinct in form from

the individual thing, but unitively contained in it. Thomas Aquinas gave it no being at all. Though these views reflect radically different metaphysical settings, they all variously bar the natures from real existence when separated in any way from the individual.

In its conventional applications to Greek and medieval thought, accordingly, Realism turns out to be an elusive and even confusing notion. It seems to be an inept way of emphasizing difficulties that are significantly present in the philosophies of these epochs, which require understanding and solution. But the granting of extramental existence to the generic and specific natures has raised more difficulties than it has solved.

Realism of things

All of these ancient and medieval doctrines—whether Realistic, conceptualistic, or nominalistic—accept the external existence of individual sensible things. From this viewpoint, they would all be Realisms in the second main sense of the term, that of the Realism of ordinary things, which is the sense in which Realism is predominantly employed in the modern era. Here it means the epistemological (or theory-of-knowledge) view that things taken as individual wholes have an existence that is outside of human cognition.

HISTORY OF WESTERN REALISM

In these and other related ways, modern writers have seen the philosophic attitude called Realism continually surfacing in the stream of Western thought, suggesting that it is a perennial feature.

Ancient Realism. In pre-Socratic thought, even in Parmenides (late 6th century), known for reducing reality to the One, the relevant reality of the objects of cognition was everywhere assumed. In Plato (5th and 4th centuries BC), the separate and more excellent existence of the natures, or Forms, was strongly asserted at times, though quite often the immanence of the form in individuals was just as surely implied without any satisfactory reconciliation of the contradiction. With Philo of Alexandria, a Hellenistic Jewish philosopher in the 1st century AD, the existence of the Platonic Forms was located within the mind of God, a view also found in the early 5th century in Augustine of Hippo (*De diversis quaestionibus*, "On Diverse Questions"). In the medieval Augustinian tradition, for instance in the writings of Anselm of Canterbury, the influence of this interpretation persisted. In the early 6th century, on the other hand, Boethius, perhaps the intellectual founder of the Middle Ages, in transmitting Aristotelian logic to the West, presented the universal notions with a strong cast of Platonic Realism, while acknowledging that the Aristotelian view was different. Among medieval thinkers in the early 12th century, such as William of Champeaux, the Parisian logician and theologian, the Platonizing tradition of Boethius was dominant, though it was brought under fire by such men as Roscelin, founder of nominalism, who saw universals as mere words. With the stormy controversialist Peter Abelard, who was the foremost dialectician of his time, the Boethian Realism was attacked. But according to Abelard, more than mere words were required to justify universality; in his view, universals were concepts signifying real things and had their ultimate basis in the divine ideas, as in the Augustinian tradition.

Scholastics, Cartesians, and Platonists

Medieval Realism. The approach of the early medieval thinkers to the problem of universals was made from the side of logic. But it soon involved theological issues, which, when added to the much deeper study of its metaphysical backgrounds made by Boethius, led the scholars to place the relevant questions in a different setting. The natures or common essences of things came to be scrutinized from a threefold viewpoint: as existent in sensible things, as existent in the mind, and as absolutely existent in themselves. This subjected the problem to metaphysical investigation. In that setting, Aquinas allowed neither being nor unity to be attributed to the nature taken absolutely. Duns Scotus, however, accorded it a lesser unity than that of the individual and gave it a kind of being proportionate to this real specific unity, but which required unitive containment of the nature by the individual. In these different ways the nature, so taken, provided the ground for the universal that existed only in the mind. Views incorporating this feature have been called moderate Realisms, though the designation is open to the same objection as it is in its application to Aristotle (see above, *Distinctions among the Realisms*). In later Scholastic tradition the currents became badly confused, and unending controversy raged on the various kinds of universals and the respective status of each type.

Moderate Realism

Modern Realism. In the familiar formula *cogito ergo sum* ("I think; therefore, I am") proffered by the first notable modern philosopher, René Descartes, methodical thinking was rooted in thought itself, thus raising the problem of how any material world outside of thought could be reached philosophically. In Descartes and a half century later in the British Empiricist John Locke, an external origin for sensations was accepted, though without any thoroughly philosophical justification. Rather, the denial of an external world was regarded as too absurd to be countenanced. In this perspective Locke's philosophy displayed a commonsense Realism. According to one of Locke's contemporaries, the Cartesian Nicolas Malebranche (known for his claim that God's will is the true cause of motion), religious faith guaranteed the external world. The Cambridge Platonists, a sober group of 17th-century moral and religious Rationalists, in a similar atmosphere of faith and with a Cartesian understanding of sensation, acquiesced in the external existence of sensible things while, against a Neoplatonic background, they accorded a respectively greater reality to the objects of intellectual cognition. For Berkeley, an early 18th-century Empiricist and Idealist, the scriptural guarantee was lacking because matter was nowhere mentioned in the revealed descriptions of the sensible universe; accordingly, in his view, no sensible world outside cognition was left. But in David Hume, whose teachings marked the climax of the Empiricist movement, even the cognitive subject, or soul, vanished.

Facing the impossibility of a genuine philosophical justification for arguing to an external world from the starting point of mind or idea, Claude Buffier, an early 18th-century French Jesuit, and, shortly later, the Scottish Realists leaned explicitly on common sense as the motive for accepting the world's external existence. The most prominent exponent of this school was Thomas Reid, an opponent of paradox and skepticism. And John Witherspoon, who was called from Scotland to the presidency of Princeton University, held that "the impression itself implies and supposes something external that communicates it, and cannot be separated from that supposition." Consequently, the attempt of the viewpoint of Berkeleian immaterialism "to unsettle the principles of common sense by metaphysical reasoning" could, in his view, never produce conviction.

Views based on common sense

20th-century Realism. Around the turn of the 20th century, a strong revolt against Kantian subjectivism and the dominant Idealisms appeared in such thinkers as William James, a psychologist and Pragmatist; Bertrand Russell, perhaps the most influential logician and philosopher of his time; and G.E. Moore, a meticulous pioneering Analyst. Thus it was that very early in the century philosophers came to use Realism, as opposed to Idealism, for their own ways of thinking. In 1904, James signalled the resurrection of natural Realism. In 1910, W.P. Montague of Columbia University and Ralph Barton Perry of Harvard University and several others signed an article entitled "The Program and First Platform of Six Realists," and followed it with a cooperative volume, *The New Realism* (1912). New Realism, or neo-Realism, in defending the independence of known things, explained that in cognition "the content of knowledge, that which lies in or before the mind when knowledge takes place, is numerically identical with the thing known." To other Realists this epistemological monism, as Perry called his theory of knowledge, failed to extricate itself from the egocentric predicament (*i.e.*, from the incapacity of the mind to transcend its private experience) that they all professed to see in the logic of Idealism. Nor could it give a satisfactory explanation of the mind's proneness to error, or even of cognition itself as being significantly different from the things known. Another

type of Realism was advanced against neo-Realism in a similarly cooperative volume entitled *Essays in Critical Realism* (1920), by the naturalist George Santayana and several others. To the monism of the neo-Realists such writers opposed an epistemological dualism, in which the object in cognition and the object in reality are numerically two at the time of perception. They divided, however, into a majority group and a minority group on the status of the immediately given object. For the majority group this datum was not an existent but merely an essence; for the others it was an existent—a mental or psychic existent for some and a physical (brain) existent for others. Here agreement failed, and the cooperative effort of the critical Realists soon fell apart.

These and the ensuing discussions left a recognized distinction between the schools of representative Realism and direct Realism: for representative Realism the immediate confrontation of cognition occurred over against a mental representation of the external object; for direct Realism the confrontation was immediately with the thing existent outside of cognition. The critical Realists themselves, in claiming that the datum was not an object as such but only the means of perceiving it, disavowed any representationalism; but in others, who proposed that the sense-datum was the image directly apprehended and was markedly different from the physical object, representative Realism was definitely present. Representationalism may also be seen in the Realism of the Belgian Neoscholastic Désiré Mercier, who founded the school of Louvain, and in the physiological Idealisms of contemporary neurophysiologists. In essence, representationalism was the inferential procedure employed by Descartes and Locke for reaching the external world. Direct Realism, on the other hand—as defended by recent writers—acknowledged no intermediate object between cognition and the external thing perceived.

Naive and tacit Realisms

Within the ambit of contemporary discussions, naive Realism was the label for any unquestioning belief that things in reality correspond exactly to human cognition of them. Expressly meant as a prephilosophical attitude, naive Realism can hardly be included under philosophical procedures. Yet its appropriateness to the man in the street has also been widely challenged; for the ordinary man is keenly interested in distinguishing critically between reality and figments of cognition and is continually doing so in ordinary life. He does not proceed, however, as did the aforementioned Realisms: he does not first regard the object in terms of its status in cognition and then explore its relation to reality. But to come under the notion as introduced by the Realists, naive Realism must be explained in terms of the cognitional relation—*e.g.,* as one of the "three typical theories of the knowledge relation." It is, accordingly, a philosophical category, though historians and controversialists shun the listing of recognized philosophers under such a title.

Further, a number of philosophies that neither bore the name of Realism nor defined reality in terms of its relation to cognition are frequently regarded as Realisms today. Aristotelianism, for example, explained the reality of things through their substantiality, Thomism through their existence in themselves, Scotism through the metaphysical priority of a nature possessed in common, and contemporary linguistic philosophy, as in John Austin, an important mid-20th-century Oxford Analyst, through a completely ostensive view of language; yet all have been seen as Realisms. The process philosophies of the Pragmatist John Dewey and of Alfred North Whitehead, an influential cosmologist and metaphysician, and—still more controversially—the philosophy of Charles Sanders Peirce, an individualistic American logician and Pragmatist, may also be taken as Realisms, even though they did not stress the basic relation to cognition; for these thinkers agreed that things as a fact do have, or may have, existence outside cognition, even though this existence was not reached from cognition nor defined through its relation to cognition. With them the cognitional relation was only an inessential afterthought. Serious interest in explaining as Realism the traditional tenets of pre-Cartesian philosophies may be seen in the writings of many contemporaries. Yet for Aristotle and Aquinas, what was meant by the reality of sensible things is already established metaphysically, through their substantiality or ontal (real) existence, before they are compared with cognition; hence to bring in the further notion of Realism for this purpose seems meaningless. The notion is therefore extraneous to the philosophical procedures of thinkers who locate the starting point of their philosophy in some actuality of the real thing itself; for relation to cognition does not play an operative role in their basic procedures. Only by means of entirely extrinsic bonds can they be grouped with the genuine Realisms.

Meta-physical Realisms

MAJOR ISSUES AND EVALUATION OF REALISM

From the foregoing survey of the historical development of Realistic thought, the major issues upon which Realism focusses attention stand out clearly. For both speculative and practical reasons, men wish to distinguish sharply between what they call reality and what they recognize or suspect to be merely products of their own cognition. Accordingly, the ancient Platonic concentration on specific and generic natures and, in Aristotle, the essential role played by the universal in reasoning led to a close scrutiny of the way in which these natures exist. Undoubtedly, they exist in human thought. For the Realistically inclined thinker, however, their crucial role tends to demand counterparts if human thinking is to bear on what really exists. Still more drastic are the post-Cartesian philosophies in which the existence of external things themselves does not enter human cognition in direct confrontation. Finally, the mathematical and scientific constructs, which have been so fruitful in man's struggle for mastery over nature, seem to require for the Realistically minded thinker some counterpart in the things themselves in order to provide an adequate philosophical explanation of their success.

Problems and current status

When the issues are faced in the foregoing manner, some lines of a procedure common to the various explicit Realisms emerge upon which it may be possible to base an evaluation. Universals, sensations and perceptions, scientific formulas and laws are all found to be existent in cognition. From that sure starting point, attempts are made to show that objects either corresponding to them or identical with them exist outside the mind. That pattern seems to be the general procedure followed in any way of thinking that has spontaneously given rise to the notion of an epistemological Realism and that can, with historical and philosophical significance, be labelled such. As is likewise apparent from the foregoing survey, this way of thinking follows a dubious procedure: Realism is not primarily a doctrine of the existence of things but rather a doctrine of cognition. In Realism, cognition is regarded as the object most present to itself; *i.e.,* a man knows his own thought processes more intimately than anything else. But the genuine Realist seems unwittingly to take the material thing as his model in conceiving cognition. No external object can be more present to a material thing than that material thing itself. If cognition is conceived after this analogy, it will be what is most present to itself, and will have to be the starting point from which the Realist reasons. This starting point seems to offer no exit. The objects reached from it can be only internal products or occurrences in the mind, for it offers nothing more basic from which to reason than the cognition itself. Any philosophically genuine Realism seems, in consequence, prone to failure in its basic objective.

Accordingly, Realism, in the senses responsible for the epistemological use of the term, has long since ceased to inspire vigorous debate. In a Platonic tradition that continues in modern thought, however, Realism in respect to the natures of things is by no means dead. In regard to the Cartesian problem of the world's external existence, representative Realism seems to have shared the fate of the sense-datum and to be quite inoperative outside of the wake of neurophysiological writings. But attempts at direct Realism are still made. In the scientific field the opposition to conventionalism and to retaining a merely instrumental status for laws or theories has remained a lively issue. Moreover, modern means, such as the electron

microscope (which shows molecules in real existence) and the hope of being able to see atoms foretoken a greater correspondence of scientific constructs with the structure of reality than had previously been demonstrated. But this verification process consists in a comparison of reality with thought rather than in any attempt to reach reality from cognition alone. (J.O./L.H.St.)

Scholasticism

From the time of the Renaissance until at least the beginning of the 19th century, the term Scholasticism, not unlike the name Middle Ages, was used as an expression of blame and contempt. The medieval period was widely viewed as an insignificant intermezzo between Greco-Roman antiquity and modern times, and Scholasticism was normally taken to describe a philosophy busied with sterile subtleties, written in bad Latin, and above all subservient to the theology of popery. Even the German Idealist Hegel, in his *Vorlesungen über die Geschichte der Philosophie* (1833–36; *Lectures on the History of Philosophy,* 1892–96), declared that he would "put on seven-league boots" in order to skip over the thousand years between the 6th and 17th centuries and, having at last arrived at Descartes, said that now he could "cry land like the sailor." In those same first decades of the 19th century, on the other hand, the Romanticists swung the pendulum sharply to the opposite side, to an indiscriminate overestimation of everything medieval.

Today, scholars seem better able to confront the medieval epoch, as well as Scholasticism—*i.e.,* its philosophy (and theology)—without prejudgments. One reason for this state of affairs is the voluminous research which has been devoted to this era and which has revealed its true nature, not only as a respectable continuation of the genuinely philosophical tradition but also as a period of exemplary personalities quite able to stand comparison with any of the great philosophers of antiquity or of modern times.

NATURE AND SIGNIFICANCE

Scholasticism is so much a many-sided phenomenon that, in spite of intensive research, scholars still differ considerably in their definition of the term and in the emphases that they place on individual aspects of the phenomenon. Some historians, seeming almost to capitulate to the complexity of the subject, confine themselves to the general point that Scholasticism can only be defined denotatively as that kind of philosophy that during the European Middle Ages was taught in the Christian schools. The question of its connotation, however, remains, viz., What kind of philosophy was it?

The answer that Scholasticism was "school" philosophy and, in fact, "Christian" school philosophy can be understood only by examining the historical exigencies that created the need for schools. The search thus leads the inquirer back to the transition from antiquity to the Middle Ages—a point which, according to Hegel, was marked by the symbolic date AD 529, when a decree of the Christian emperor Justinian closed the Platonic Academy in Athens and sealed "the downfall of the physical establishments of pagan philosophy." In that same year, however, still another event occurred, which points much less to the past than to the coming age and, especially, to the rise of Scholasticism, viz., the foundation of Monte Cassino, the first Benedictine abbey, above one of the highways of the great folk migrations. This highly symbolic fact not only suggests the initial shift of the scene of the intellectual life from places like the Platonic Academy to the cloisters of Christian monasteries, but it marks even more a change in the dramatis personae. New nations were about to overrun the Roman Empire and its Hellenistic culture with long-range effects: when, centuries later, for example, one of the great Scholastics, Thomas Aquinas, was born, though he was rightly a southern Italian, his mother was of Norman stock, and his Sicilian birthplace was under central European (Hohenstaufen) control.

Barbarians' need for school learning

It was a decisive and astonishing fact that the so-called barbarian peoples who penetrated from the north into the ancient world often became Christians and set out to master the body of tradition that they found, including the rich harvest of patristic theology as well as the philosophical ideas of the Greeks and the political wisdom of the Romans. This learning could be accomplished only in the conquered empire's language (*i.e.,* in Latin), which therefore had to be learned first. In fact, the incorporation of both a foreign vocabulary and a different mode of thinking and the assimilation of a tremendous amount of predeveloped thought was the chief problem that confronted medieval philosophy at its beginnings. And it is only in the light of this fact that one of the decisive traits of medieval Scholasticism becomes understandable: Scholasticism above all was an unprecedented process of learning, literally a vast "scholastic" enterprise that continued for several centuries. Since the existing material had to be ordered and made accessible to learning and teaching, the very prosaic labour and "schoolwork" of organizing, sorting, and classifying materials inevitably acquired an unprecedented importance. Consequently, the writings of medieval Scholasticism quite naturally lack the magic of personal immediacy, for schoolbooks leave little room for originality. It is therefore misleading, though understandable, that certain polemicists have wrongly characterized Scholasticism as involving no more than the use of special didactic methods or a narrow adherence to traditional teachings.

First of all, if the major historical task of that epoch was really to learn, to acquire, and to preserve the riches of tradition, a certain degree of "scholasticity" was not only inevitable but essential. It is not at all certain that today's historians would have direct intellectual access to Plato, Aristotle, and Augustine had the Scholastics not done their patient spadework. Besides, the progress from the stage of mere collection of given sentences and their interpretation (*expositio, catena, lectio*), to the systematic discussion of texts and problems (*quaestio, disputatio*), and finally to the grand attempts to give a comprehensive view of the whole of attainable truth (*Summa*) was necessarily at the same time a clear progression toward intellectual autonomy and independence, which in order to culminate, as it did in the 13th century, in the great works of Scholasticism's Golden Age, required in addition the powers of genius, of men like Albertus Magnus and Thomas Aquinas.

Desirability of Scholasticism

On the other hand, the moment had to come when the prevalent preoccupation with existing knowledge would give way to new questions, which demanded consideration and answers that could emerge only from direct experience. By the later Middle Ages, procedures for exploiting and discussing antecedent stocks of insight had been largely institutionalized, and it was an obvious temptation to perpetuate the dominion of those procedures—which could lead only to total sterility. It is widely agreed that this is almost exactly what did happen in the 14th century in what is called the "decline" and disintegration of Scholasticism.

HISTORY AND ISSUES

Roots of Scholasticism. From the beginning of medieval Scholasticism the natural aim of all philosophical endeavour to achieve the "whole of attainable truth" was clearly meant to include also the teachings of Christian faith, an inclusion which, in the very concept of Scholasticism, was perhaps its most characteristic and distinguishing element. Although the idea of including faith was expressed already by Augustine and the early Church Fathers, the principle was explicitly formulated by the pivotal, early 6th-century scholar Boethius. Born in Rome and educated in Athens, Boethius was one of the great mediators and translators, living on the narrow no-man's-land that divided the epochs. His famous book, *The Consolation of Philosophy,* was written while he, indicted for treachery and imprisoned by King Theodoric the Goth, awaited his own execution. It is true that the book is said to be, aside from the Bible, one of the most translated, most commented upon, and most printed books in world history; and that Boethius made (unfinished) plans to translate and to comment upon, as he said, "every book of Aristotle and all the dialogues of Plato." But the epithet that he won as "one of the founders of Scholasticism" refers to

Boethius and Cassiodorus

quite another side of his work. Strictly speaking, it refers to the last sentence of a very short tractate on the Holy Trinity, which reads, "As far as you are able, join faith to reason"—an injunction which in fact was to become, for centuries, the formal foundation of Scholasticism. Instead of "faith," such concepts as revelation, authority, or tradition could be (and, indeed, have been) cited; and "reason," though unambiguously meant to designate the natural powers of human cognition, could also be granted (and, in fact, has been granted) very different meanings. In any case, the connection between faith and reason postulated in this principle was from the beginning and by its very nature a highly explosive compound.

Boethius himself already carried out his program in a rather extraordinary way: though his *Opuscula sacra* ("Sacred Works") dealt almost exclusively with theological subjects, there was not a single Bible quotation in them: logic and analysis was all.

Though called the "first Scholastic," Boethius was at the same time destined to be for almost a millennium the last layman in the field of European philosophy. His friend Cassiodorus, author of the *Institutiones,* an unoriginal catalog of definitions and subdivisions, which (in spite of their dryness) became a source book and mine of information for the following centuries, who, like Boethius, occupied a position of high influence at the court of Theodoric and was also deeply concerned with the preservation of the intellectual heritage, decided in his later years to quit his political career and to live with his enormous library in a monastery. This fact again is highly characteristic of the development of medieval Scholasticism: intellectual life needs not only teachers and students and not only a stock of knowledge to be handed down; there is needed a certain guaranteed free area within human society as well, a kind of sheltered enclosure, within which the concern for "nothing but truth" can exist and unfold. The Platonic Academy, as well as (for a limited time) the court of Theodoric, had been enclosures of this kind; but in the politically unsettled epoch to come "no plant would thrive except one that germinated and grew in the cloister."

The principle of the conjunction of faith and reason, which Boethius had proclaimed, and the way in which he himself carried it out were both based on a profound and explicit confidence in man's natural intellectual capacity—a confidence that could possibly lead one day to the rationalistic conviction that there cannot be anything that exceeds the power of human reason to comprehend, not even the mysteries of divine revelation. To be sure, the great thinkers of Scholasticism, in spite of their emphatic affirmation of faith and reason, consistently rejected any such rationalistic claim. But it must nonetheless be admitted that Scholasticism on the whole, and by virture of its basic approach, contained within itself the danger of an overestimation of rationality, which recurrently emerged throughout its history.

The Pseudo-Dionysius

On the other hand, there had been built in, from the beginning, a corrective and warning, which in fact kept the internal peril of Rationalism within bounds, viz., the corrective exercised by the "negative theology" of the so-called Pseudo-Dionysius, around whose writings revolved some of the strangest events in the history of Western culture. The true name of this protagonist is, in spite of intensive research, unknown. Probably it will remain forever an enigma why the author of several Greek writings (among them *On the Divine Names,* "On the Celestial Hierarchy," and *The Mystical Theology*) called himself "Dionysius the Presbyter" and, to say the least, suggested that he was actually Denis the Areopagite, a disciple of Paul the Apostle (Acts). In reality, almost all historians agree that Pseudo-Dionysius, as he came to be called, was probably a Syrian Neoplatonist, a contemporary of Boethius. Whatever the truth of the matter may be, his writings exerted an inestimable influence for more than 1,000 years by virtue of the some-what surreptitious quasi-canonical authority of their author, whose books were venerated, as has been said, "almost like the Bible itself." A 7th-century Greek theologian, Maximus the Confessor, wrote the first commentaries on these writings, which were followed over the centuries by a long succession of commentators, among them Albertus Magnus and Thomas Aquinas. The main fact is that the unparalleled influence of the Areopagite writings preserved in the Latin West an idea, which otherwise could have been repressed and lost (since it cannot easily be coordinated with rationality)—that of a negative theology or philosophy that could act as a counter-poise against Rationalism. It could be called an Eastern idea present and effective in the Occident. But after the Great Schism, which erected a wall between East and West that lasted for centuries, Denis the Areopagite, having become himself (through translations and commentaries) a Westerner "by adoption," was the only one among all of the important Greco-Byzantine thinkers who penetrated into the schools of Western Christendom. Thus negative theology was brought to medieval Scholasticism, as it were, through the back door.

The most important book of Denis, which dealt with the names that can be applied to God, exemplified his negative theology. It maintained first of all the decidedly biblical thesis that no appropriate name can be given to God at all unless he himself reveals it. But then Denis showed that even the revealed names, since they must be comprehensible to man's finite understanding, cannot possibly reach or express the nature of God; and that in consequence, every affirmative statement about God requires at once the corrective of the coordinate negation. The theologian cannot even call God "real" or "being," because he derives these concepts from the things to which God has given reality; and the Creator cannot possibly be of the same nature as that which he has created. Thus, *The Mystical Theology* concluded by finally relativizing also the negations, because God surpasses anything that man may possibly say of him, whether it be affirmative or negative.

Scholasticism certainly could have learned all of this also from Augustine, who repeatedly warned that "Whatever you understand cannot be God." But probably an authority of even greater weight than Augustine was needed to counteract a reason that was tending to overrate its own powers; and this authority was attributed, although falsely, to the works of Denis the Areopagite. This impact could, of course, not be restricted to the idea of God; it necessarily concerned and changed man's whole conception of the world and of existence. The influence of Denis is reflected in the noteworthy fact that Thomas Aquinas, for instance, not only employed more than 1,700 quotations from Denis the Areopagite but also appealed almost regularly to his work whenever he spoke, as he often did (and in astonishingly strong terms), of the inexhaustible mystery of being. Thomas Aquinas, however, who also wrote a remarkable commentary on Denis' book *On the Divine Names,* is mentioned here only as an example, albeit a most telling example.

At the very end of the medieval era of Scholasticism, the Areopagite emerged once more in the work of a 15th-century cardinal, Nicholas of Cusa, also known as a mathematician and advocate of experimental knowledge, in whose library there are preserved several translations of the Areopagite writings—replete, moreover, with marginal notes in the Cardinal's handwriting. But even without this concrete evidence, it would be quite plain that Cusanus' doctrine of "knowing nonknowing" is closely linked to the Areopagite's conviction that all of reality is unfathomable.

Erigena's translation

The translation into Latin of the *Corpus Areopagiticum,* which was made in the 9th century—*i.e.,* some 400 years after the death of its author—by John Scotus Erigena, is itself worthy of mention, especially because the translator was one of the most remarkable figures of early medieval philosophy. After generations of brave and efficient collectors, organizers, and schoolmasters had come and gone, Erigena, in his *De divisione natura* ("On the Division of Nature"), developed the Dionysian Neoplatonism on his own and tried to construct a systematic conception of the universe, a more or less pantheistic world view, which (as Gilson says) for a moment offered the Latin West the opportunity—or the temptation—to choose the way of the East once and for all. The church, though not until centuries later, condemned the book, apparently convinced that any counterpoise to its own position can become dangerous in itself.

Early Scholastic period. If there was any philosophical-theological thinker of importance during the Middle Ages who remained untouched by the spirit of the Areopagite, it was the 11th-century Benedictine Anselm of Canterbury, a highly cultivated Franco-Italian theologian who for years was prior and abbot of the abbey Le Bec in Normandy and then became, somewhat violently, the archbishop of Canterbury. In Anselm's entire work there is not a single quotation from Denis; not even the name is mentioned. Consequently, Anselm's thinking, thus freed from the corrective embodied in the Areopagite's negative theology, displayed a practically unlimited confidence in the power of human reason to illuminate even the mysteries of Christian faith; he thus frequently approached a kind of Rationalism, which did not shrink from the attempt to demonstrate, on compelling rational grounds, that salvation (for example) through God incarnate was philosophically necessary. To be sure, a theologian such as Anselm certainly would never have subscribed to the extreme thesis that nothing exists that is beyond the power of human reason to comprehend. the two famous phrases, coined by him and expressing again, in a grandiose formulation, the principle of Boethius, "faith seeking to be understood" and "I believe in order to understand," clearly proclaim his faith in the mysteries of revelation as comprising the very basis of all reasoning. Nevertheless, in the case of Anselm, the very peculiar conjunction of faith and reason was accomplished not so much through any clear intellectual coordination as through the religious energy and saintliness of an unusual personality. It was accomplished, so to speak, rather as an act of violence, which could not possibly last. The conjunction was bound to break up, with the emphasis falling either on some kind of Rationalism or on a hazardous irrationalization of faith.

Anselm's belief in human reason

That this split did actually happen can be read to some extent in the fate of the "Anselmic argument," which Kant, 700 years later, was to reject as the "ontological proof of God"—connecting it, however, not with the name of Anselm but with that of Descartes, the earliest modern philosopher. It is, in fact, significant that Descartes, in his proof of the existence of God, imagined that he was saying the same thing as Anselm, and that, on the other hand, Anselm would scarcely have recognized his own argument had he encountered it in the context of Descartes's *Discours de la méthode* (1637; *Discourse on Method,* 1950), which claims to be "pure" philosophy based upon an explicit severance from the concept of God held by faith. But given Anselm's merely theoretical starting point, that severance was not only to be expected; it was almost inevitable.

But, also within the framework of medieval Scholasticism, a dispute was always brewing between the dialecticians, who emphasized or overemphasized reason, and those who stressed the supranational purity of faith. Berengar of Tours, an 11th-century logician, metaphysician, and theologian, who was fond of surprising formulations, maintained the preeminence of thinking over any authority holding, in particular, that the real Presence of Christ in the Eucharist was logically impossible. His contemporary the Italian hermit-monk and cardinal Peter Damian, however—who was apparently the first to use the ill-famed characterization of philosophy as the "handmaid of theology"—replied that, if God's omnipotence acts against the principle of contradiction, then so much the worse for the science of logic. Quite analogous to the foregoing controversy, though pitched on a much higher intellectual level, was the bitter fight that broke out almost one century later between a Cistercian reformer, Bernard of Clairvaux, and a logician and theologian, Peter Abelard. Bernard, a vigorous and ambivalent personality, was in the first place a man of religious practice and mystical contemplation, who, at the end of his dramatic life, characterized his odyssey as that of *anima quaerens Verbum,* "a soul in search of the Word." Although he by no means rejected philosophy on principle, he looked with deep suspicion upon the primarily logical approach to theology espoused by Abelard. "This man," said Bernard, "presumes to be able to comprehend by human reason the entirety of God."

Bernard and Abelard

Logic was at that time, as a matter of fact, the main battleground of all Scholastic disputations. "Of all philosophy, logic most appealed to me," said Abelard, who by "logic" understood primarily a discipline not unlike certain present-day approaches, the "critical analysis of thought on the basis of linguistic expression." From this viewpoint (of linguistic logic), Abelard also discussed with penetrating sharpness the so-called "problem of universals," which asks, Is there an "outside" and objective reality standing, for example, not only for the name "Socrates" but also for such common names as "man," "caniness," and the like? Or do common concepts ("universals") possess only the reality of subjective thought or perhaps merely that of the sound of the word? As is well known, it has been asserted that this was the principal, or even the only, subject of concern in medieval Scholasticism—a charge that is misleading, although the problem did greatly occupy philosophers from the time of Boethius. Their main concern from the beginning was the whole of reality and existence.

The advance of medieval thought to a highly creative level was foreshadowed, in those very same years before Peter Abelard died, by Hugh of Saint-Victor (an Augustinian monk of German descent), when he wrote *De sacramentis Christianae fidei* ("On the Sacraments of the Christian Faith"), the first book in the Middle Ages that could rightly be called a *summa;* in its introduction, in fact, the term itself is used as meaning a comprehensive view of all that exists (*brevis quaedam summa omnium*). To be sure, its author stands wholly in the tradition of Augustine and the Areopagite; yet he is also the first medieval theologian who proclaims an explicit openness toward the natural world. Knowledge of reality is, in his understanding, the prerequisite for contemplation; each of the seven liberal arts aims "to restore God's image in us." "Learn everything," he urged; "later you will see that nothing is superfluous."

Hugh's *summa* and Lombard's *Sentences*

It was on this basic that the university—which was not the least of the achievements of medieval Scholasticism—was to take shape. And it was the University of Paris, in particular, that for some centuries was to be the most representative university of the West. Though there are usually a variety of reasons and causes for such a development, in this case the importance of the university—unlike that of Bologna and also of Oxford—lay mainly in the fact that it was founded in the most radical way upon those branches of knowledge that are "universal" by their very nature: upon theology and philosophy. It is, thus, remarkable, though not altogether surprising, that there seems to have existed not a single *summa* of the Middle Ages that did not, in some way or other, derive from the University of Paris.

Strangely enough, the classical theological-philosophical textbook used in the following centuries at the universities of the West was not the first *summa,* composed by Hugh of Saint-Victor, but was instead a work by Peter Lombard, a theologian who probably attended Abelard's lectures and who became *magister* at the cathedral school of Notre-Dame and, two decades later, bishop of Paris. Lombard's famous *Four Books of Sentences,* which, though written one or two decades later than Hugh's *summa,* belonged to an earlier historical species, contained about 1,000 texts from the works of Augustine, which comprise nearly four-fifths of the whole. Much more important than the book itself, however, were the nearly 250 commentaries on it, by which—into the 16th century—every master of theology had to begin his career as a teacher. In view of this wide usage, it is not astonishing that Lombard's book underwent some transformations, at the hands, for instance, of its most ingenious commentator, Thomas Aquinas, but also (and even more so) at the hands of Duns Scotus in his *Opus Oxoniense,* which, in spite of being a work of extremely personal cast, was outwardly framed as a commentary on the "Master of Sentences."

Maturity of Scholasticism. Clearly, the world view of Western Christendom, on the whole Augustinian and Platonic in inspiration and founded upon Lombard's "Augustine breviary," was beginning to be rounded out into a system and to be institutionalized in the universities.

The renascence of Aristotle

At the very moment of its consolidation, however, an upheaval was brewing that would shake this novel conception to its foundations: the main works of Aristotle, hitherto unknown in the West, were being translated into Latin—among them his *Metaphysics,* the *Physics,* the *Nichomachean Ethics,* and the books *On the Soul.* These writings were not merely an addition of something new to the existing stock; they involved an enormous challenge. Suddenly, a new, rounded, coherent view of the world was pitted against another more-or-less coherent traditional view; and because this challenge bore the name of Aristotle, it could not possibly be ignored, for Aristotle's books on logic, translated and equipped with commentaries by Boethius, had for centuries been accepted as one of the foundations of all culture. During the lifetime of Abelard the full challenge of the Aristotelian work had not yet been presented, though it had been developing quietly along several paths, some of which were indeed rather fantastic. For instance, most of the medieval Latin translations of Aristotle stem not from the original Greek but from earlier Arabic translations.

Within the Western Christendom of the 2nd millennium, a wholly new readiness to open the mind to the concrete reality of the world had arisen, a view of the universe and life that resembled the Aristotelian viewpoint. The tremendous eagerness with which this new philosophy was embraced was balanced, however, by a deep concern lest the continuity of tradition and the totality of truth be shattered by the violence of its assimilation. And this danger was enhanced by the fact that Aristotle's works did not come alone; they came, in fact, accompanied by the work of Arabic commentators and their heterodox interpretations.

The most influential Arabic commentators were an 11th-century polymath, Avicenna, a Persian by birth, and a 12th-century philosopher, Averroës, born in Spain. Avicenna, personal physician to sovereigns, but also a philosopher and theologian, read—according to his own account—Aristotle's *Metaphysics* 40 times without understanding it, until he learned the text by heart. F.C. Copleston has called him "the real creator of a Scholastic system in the Islāmic world." In the view of Averroës, who was not only a philosopher but also a jurist and a doctor, Aristotle's philosophy represented simply the perfection of human knowledge; and to the West, he himself was to become *the* commentator. A third great commentator was a 12th-century orthodox Jewish philosopher, Moses Maimonides, also born in Spain, who wrote his main works in Arabic. Maimonides was at the same time a vigorous adherent of the Aristotelian world view and was, thus, confronted by the same unending task that preoccupied the great teachers of medieval Christendom. At first sight it appears strange that none of these three thinkers had any appreciable influence within his own world (neither Islām nor Judaism knew of any such thing as a "discovery" of Aristotle), whereas on almost every page of the 13th-century Christian *summae* the names of Avicenna, Averroës, and Maimonides are found.

Albertus Magnus and Aristotelianism

The first theologian of the Middle Ages who boldly accepted the challenge of the new Aristotelianism was a 13th-century Dominican, Albertus Magnus, an encyclopedic scholar. Although he knew no Greek, he conceived a plan of making accessible to the Latin West the complete works of Aristotle, by way of commentaries and paraphrases; and, unlike Boethius, he did carry out this resolve. He also penetrated and commented upon the works of the Areopagite; he was likewise acquainted with those of the Arabs, especially Avicenna; and he knew Augustine. Nevertheless, he was in no wise primarily a man of bookish scholarship; his strongest point, in fact, was the direct observation of nature and experimentation. After having taught for some years at the University of Paris, he travelled, as a Dominican superior, through almost all of Europe. Not only was he continually asking questions of fishermen, hunters, beekeepers, and birdcatchers but he himself also bent his sight to the things of the visible world. But amidst the most palpable descriptions of bees, spiders, and apples, recorded in two voluminous books on plants and animals, Albertus formulated completely new, and even revolutionary, methodological principles: for instance, "There can be no philosophy about concrete things," or, "in such matters only experience can provide certainty."

With Albertus, the problem of the conjunction of faith and reason had suddenly become much more difficult, because reason itself had acquired a somewhat new meaning. "Reason" implied, in his view, not only the capacity for formally correct thinking, for finding adequate creatural analogies to the truths of revelation, but it implied, above all, the capacity to grasp the reality that man encounters. Henceforth, the Boethian principle of "joining faith with reason" would entail the never-ending task of bringing belief into a meaningful coordination with the incessantly multiplying stock of natural knowledge of man and the universe. Since Albertus' nature, however, was given more to conquest than to the establishment of order, the business of integrating all of these new and naturally divergent elements into a somewhat consistent intellectual structure waited for another man, for his pupil Thomas Aquinas.

Thomas Aquinas

To epitomize the intellectual task that Aquinas set for himself, the image of Odysseus' bow, which was so difficult to bend that an almost superhuman strength was needed, is fitting. As a young student at the University of Naples, he had met in the purest possible form both extremes, which, though they seemed inevitably to be pulling away from one another, it was nevertheless his life's task to join: one of these extremes was the dynamic, voluntary poverty movement whose key word was "the Bible"; and the second phenomenon was the Aristotelian writings and outlook, which at that time could have been encountered nowhere else in so intensive a form. And "Aristotle" meant to Thomas not so much an individual author as a specific world view, viz., the affirmation of natural reality as a whole, including man's body and his natural cognitive powers. To be sure, the resulting *Summa theologiae* (which Thomas himself chose to leave incomplete) was a magnificent intellectual structure; but it was never intended to be a closed system of definitive knowledge. Thomas could no longer possess the magnificent naiveté of Boethius, who had considered it possible to discuss the Trinitarian God without resorting to the Bible, nor could he share Anselm's conviction that Christian faith so completely concurred with natural reason that it could be proved on compelling rational grounds.

In the meanwhile, the poles of the controversy—the biblical impulses, on the one hand, and the philosophical and secular ones, on the other—had begun to move vigorously apart, and partisans moving in both directions found some encouragement in Thomas himself. But in his later years he realized that the essential compatibility as well as the relative autonomy of these polar positions and the necessity for their conjunction had to be clarified anew by going back to a deeper root of both; that is, to a more consistent understanding of the concepts of creation and createdness. At Paris, he had to defend his own idea of "a theologically based worldliness and a theology open to the world" not only against the secularistic "philosophism" of Siger of Brabant, a stormy member of the faculty of arts, and against an aggressive group of heterodox Aristotelians around him, but also (and even more) against the traditional (Augustinian) objection that by advocating the rights of all natural things Thomas would encroach upon the rights of God, and that, besides, the theologian needs to know only that part of creation that is pertinent to his theological subject. The latter idea was supported also by the Italian mystical theologian Bonaventura, who, in his earlier days as a colleague of Thomas at the university, had likewise been enamoured of Aristotle, but later, alarmed by the secularism that was growing in the midst of Christendom, became more mistrustful of the capacities of natural reason. Thomas answered this objection in somewhat the following way: The benefit that the theologian may derive from an investigation of natural reality cannot be determined in advance, but, in general, faith presupposes and therefore needs natural knowledge of the world; at times, an error concerning the creation leads men astray also from the truth of faith. This may sound like an optimistic Rationalism; but the corrective

of negative theology and philosophy was always present in the mind of Thomas, as well. Not only, as he argued in his treatise on God, does man not know what God is, but he does not know the essences of things either.

Late Scholastic period. Thomas did not succeed in bridging the faith–reason gulf. When he left Paris (1272) and after his death (1274), the gulf became much more radical; and on March 7, 1277, the Archbishop of Paris, in fact, formally condemned a list of sentences, some of them close to what Thomas himself had allegedly or really taught. This ecclesiastical act, questionable though it may have been in its methods and personal motivations, was not only understandable; it was unavoidable, since it was directed against what, after all, amounted in principle to an antitheological, rationalistic secularism. Quite another matter, however, were the factual effects of the edict, which were rather disastrous. Above all, two of the effects were pernicious: instead of free dispûtes among individuals, organized blocks (or "schools") now began to form; and the cooperative dialogue between theology and philosophy turned into mutual indifference or distrust. Nonetheless, the basic principle itself ("join faith with reason") had not yet been explicitly repudiated. This was to happen in the next generation.

Duns Scotus and William of Ockham

The negative element, as formulated in the theology of the Areopagite, proved to be insufficient as a corrective to counter the overemphasis of reason, for reason seemed to imply the idea of necessity; Anselm's asserted "compelling grounds" for revealed truths, for example, were akin to such a necessitarianism. A second corrective was therefore demanded and this took the name of "freedom"—which indeed was the battle cry of an important Franciscan, Duns Scotus, known as the "subtle doctor," who lived at the turn of the 14th century. Scotus used "freedom" primarily with reference to God; consequently, since redemption, grace, and salvation as well as all of creation were the work of God's groundless, absolute freedom, there could be no "necessary reasons," if indeed any reasons at all, for anything. It was therefore futile to attempt to coordinate faith with speculative reason. Clearly, Scotus' theological starting point made the conjunction of what man believes with what he knows every bit as difficult as it had been in Siger of Brabant's secularistic "philosophism." From both positions there was only one step to the doctrine of a "double truth"—a step that in fact was taken in the 14th century by the Nominalist William of Ockham, also a Franciscan, to whom singular facts alone are "real," and their coherence is not; this mere factuality, he held, can neither be calculated nor deduced, but only experienced; reason therefore means nothing but the power to encounter concrete reality. And upon such soil only a consistently "positive" theology could thrive. Any collaboration with speculative reason must be rejected as untheological. Faith is one thing and knowledge an altogether different matter; and a conjunction of the two is neither meaningfully possible nor even desirable. Inexorably, and justified by reasons on both sides, a divorce was taking place between faith and reason—to the connection of which the energies of almost a thousand years had been devoted. What was occurring was the demise of medieval Scholasticism.

ENDURING FEATURES

But not all of Scholasticism is specifically medieval and therefore definitively belonging to the dead past; there are perennial elements that are meant for every age, the present one included, three of which may be here distinguished. First, not only has Scholasticism held true to the normal historical rule that ideas, once thought and expressed, remain present and significant in the following time; but the medieval intellectual accomplishments have surpassed the rule and exerted, though more or less anonymously, a quite exceptional influence even on philosophers who consciously revolted against Scholasticism. New historical investigations clearly show that the classical modern philosophers Descartes, Locke, Spinoza, and Leibniz owe much to medieval ideas. Of Descartes, for instance, it has been said, contrary to the usual view, that he could quite well have been "included with the later Scholastics"; and even Charles Sanders Peirce, the originator of 20th-century American Pragmatism, refers not too rarely to Scholastic maxims. Secondly, there have been explicit attempts to go back to Scholastic thinkers and inspire a renascence of their basic ideas. Two chief movements of this kind were the Scholasticism of the Renaissance (called *Barockscholastik*) and the Neoscholasticism of the 19th and 20th centuries, both of which were primarily interested in the work of Thomas Aquinas.

Renaissance and Neoscholasticism

Renaissance Scholasticism received its first impulses from the Reformation. One of its leading figures, a Dominican, Cardinal Thomas de Vio (16th century), commonly known as Cajetan, had some famous disputations with Martin Luther. Cajetan's great commentary on Thomas Aquinas, published again in a late edition of the *Summa theologiae* (1888–1906), exerted for at least three centuries an enormous influence on the formation of Catholic theology. He was much more than a commentator, however; his original treatise on the "Analogy of names," for example, can even pass as a prelude to modern linguistic philosophy. The so-called Silver Age of Scholastic thought, which occurred in the 16th century, is represented by two Spaniards: Francisco de Vitoria of the first half and Francisco Suárez of the last half of the century were both deeply engaged in what has been called the "Counter-Reformation." Though likewise commentators on the works of Thomas Aquinas, the Renaissance Scholastics were much less concerned with looking back to the past than with the problems of their own epoch, such as those of international law, colonialism, resistance to an unjust government, and world community. Though Suárez was for more than a hundred years among the most esteemed authors, even in Protestant universities, Renaissance Scholasticism was eradicated by Enlightenment philosophy and German Idealism. This, in turn, gave rise in due time to the Neoscholasticism of the 19th century, one of the most effective promoters of which was a German Jesuit, Joseph Kleutgen, who published a voluminous scholarly apology of patristic and Scholastic theology and philosophy and was also responsible for the outline of the papal encyclical *Aeterni Patris* of Leo XIII (1879), which explicitly proclaimed the "instauration of Christian philosophy according to St. Thomas." The result, fed of course from many different sources, was that all over the world new centres of Scholastic research and higher learning (universities) arose—some more traditionalistic, some from the start engaged in the dialogue with modern philosophy and science, and some primarily devoted to historical studies and the preparation of critical editions of the great medieval Scholastics—and that a multitude of periodicals and systematic textbooks were produced. It is too early for a competent judgment on this enterprise to be made. Its immeasurable educational benefit for several generations of students, however, is as undeniable as the unique contributions of some Neoscholastic thinkers to current intellectual life. A weak point, on the other hand, seems to be a somewhat "unhistorical" approach to reality and existence. In any case, it is scarcely a matter of mere chance that, after World War II, the impact of Existentialism and Marxism caused a noticeable decline in Neoscholasticism and that the positions of "Scholastic" authors active in the 1970s were already beyond Neoscholasticism.

Perennial tenets of Scholasticism

The third and most important aspect of the enduring significance of the Scholastic movement implies the acceptance of the following fundamental tenets: that there exist truths that man knows, and also revealed truths of faith; that these two kinds of truth are not simply reducible to one another; that faith and theology do not, by means of symbols and sensuous images, merely say the same as what reason and science say more clearly by conceptual argumentation (Averroës, Hegel); that, on the other hand, reason is not a "prostitute" (Luther), but is man's natural capacity to grasp the real world; that since reality and truth, though essentially inexhaustible, are basically one, faith and reason cannot ultimately contradict one another. Those who hold these convictions appear quite unable to refrain from trying to coordinate what they know with what they believe. Any epoch that addresses itself to this interminable task can ill afford to ignore the demand-

ing and multiform paradigm of Scholasticism; but to the problems posed it will have to find its own answer.

(Jf.Pi.)

Skepticism

As a philosophical attitude, skepticism is the doubting of knowledge claims set forth in various areas. Skeptics have challenged the adequacy or reliability of these claims by asking what they are based upon or what they actually establish. They have raised the question whether such claims about the world are either indubitable or necessarily true, and they have challenged the alleged grounds of accepted assumptions. Practically everyone is skeptical about some knowledge claims; but the Skeptics have raised doubts about any knowledge beyond the contents of directly felt experience. The original Greek meaning of *skeptikos* was "an inquirer," someone who was unsatisfied and still looking for truth.

Skepticism in history

From ancient times onward Skeptics have developed arguments to undermine the contentions of dogmatic philosophers, scientists, and theologians. The Skeptical arguments and their employment against various forms of dogmatism have played an important role in shaping both the problems and the solutions offered in the course of Western philosophy. As ancient philosophy and science developed, doubts arose about basic accepted views of the world. In ancient times Skeptics challenged the claims of Platonism, Aristotelianism, and Stoicism, and in the Renaissance those of Scholasticism and Calvinism. After Descartes, Skeptics attacked Cartesianism and other theories justifying the "new science." Later, a Skeptical offensive was levelled against Kantianism and then against Hegelianism. Each Skeptical challenge led to new attempts to resolve the difficulties. Skepticism, especially since the Enlightenment, has come to mean disbelief—primarily religious disbelief—and the Skeptic has often been likened to the village atheist.

VARIOUS SENSES AND APPLICATIONS

Skepticism developed with regard to various disciplines in which men claimed to have knowledge. It was questioned, for example, whether one could gain any certain knowledge in metaphysics (the study of the nature and significance of being as such) or in the sciences. In ancient times a chief form was medical Skepticism, which questioned whether one could know with certainty either the causes or cures of diseases. In the area of ethics doubts were raised about accepting various mores and customs and about claiming any objective basis for making value distinctions. Skepticisms about religion have questioned the doctrines of different traditions. Certain philosophies, like those of Hume and Kant, have seemed to show that no knowledge can be gained beyond the world of experience and that one cannot discover the causes of phenomena. Any attempt to do so, as Kant argued, leads to antinomies, contradictory knowledge claims. A dominant form of Skepticism, the subject of this article, concerns knowledge in general, questioning whether anything actually can be known with complete or adequate certainty. This type is called epistemological Skepticism.

Epistemological Skepticism

Kinds of epistemological Skepticism can be distinguished in terms of the areas in which doubts are raised; that is, whether they be directed toward reason, toward the senses, or toward knowledge of things-in-themselves. They can also be distinguished in terms of the motivation of the Skeptic—whether he is challenging views for ideological reasons or for pragmatic or practical ones to attain certain psychological goals. Among the chief ideological motives have been religious or antireligious concerns. Some Skeptics have challenged knowledge claims so that religious ones could be substituted—on faith. Others have challenged religious knowledge claims in order to overthrow some orthodoxy. Kinds of Skepticism also can be distinguished in terms of how restricted or how thoroughgoing they are—whether they apply only to certain areas and to certain kinds of knowledge claims or whether they are more general and universal.

ANCIENT SKEPTICISM

Historically, skeptical philosophical attitudes began to appear in pre-Socratic thought. In the 5th century BC, the Eleatic philosophers, known for reducing reality to a static One, questioned the reality of the sensory world, of change and plurality, and denied that reality could be described in the categories of ordinary experience. On the other hand, the Ephesian philosopher of change Heracleitus and his pupil Cratylus thought that the world was in such a state of flux that no permanent, unchangeable truth about it could be found; and Xenophanes, a wandering poet and philosopher, doubted whether man could distinguish true from false knowledge.

A more developed Skepticism appeared in some of Socrates' views and in a couple of the Sophists (see below). Socrates, in the early Platonic dialogues, was always questioning the knowledge claims of others; and in the *Apology,* he said that all that he really knew was that he knew nothing. Socrates' enemy, the Sophist Protagoras, contended that man is the measure of all things. This thesis was taken as a kind of skeptical relativism: no views are ultimately true, but each is merely one man's opinion. Another Sophist, Gorgias, advanced the skeptical-nihilist thesis that nothing exists; and if something did exist, it could not be known; and if it could be known, it could not be communicated.

The putative father of Greek Skepticism is Pyrrhon of Elis (*c.* 360–*c.* 272 BC), who tried to be a living Skeptic. He avoided committing himself to any views about what was actually going on and acted only according to appearances. In this way he sought happiness or at least mental peace.

Academic and Pyrrhonian schools

The first school of Skeptical philosophy developed in Plato's Academy (see above *Platonism*) in the 3rd century BC and was thus called "Academic" Skepticism. Starting from the skeptical side of Socrates, its leaders, Arcesilaus (316/315–*c.* 241 BC) and Carneades (214/213–129/128 BC), set forth a series of epistemological arguments to show that nothing could be known, challenging primarily the two foremost schools, those of the Stoics and Epicureans. They denied that any criteria could be found for distinguishing the true from the false; instead, only reasonable or probable standards could be established for knowledge. This limited or probabilistic Skepticism was the view of the Academy until the 1st century BC, when Cicero was a student there. His *Academica* and *De natura deorum* are the main sources for knowledge of this movement. (St. Augustine's *Contra academicos* is an answer to Cicero's views.)

The other major form of ancient Skepticism was Pyrrhonism, apparently developed by medical Skeptics in Alexandria. Beginning with Aenesidemus (1st century BC), this movement, named after Pyrrhon, criticized the Academic Skeptics because they claimed to know too much, namely, that nothing could be known and that some things are more probable than others. The Pyrrhonians advanced a series of tropes, or ways of opposing various kinds of knowledge claims, in order to bring about *epochē* (suspense of judgment). The Pyrrhonian attitude is preserved in the writings of one of its last leaders, Sextus Empiricus (2nd or 3rd century AD). In his *Outlines of Pyrrhonism* and *Adversus mathematicos,* Sextus presented the tropes developed by previous Pyrrhonists. The 10 tropes attributed to Aenesidemus showed the difficulties to be encountered in ascertaining the truth or reliability of judgments based on sense information, owing to the variability and differences of human and animal perceptions. Other arguments raised difficulties in determining whether there are any reliable criteria or standards—logical, rational, or otherwise—for judging whether anything is true or false. To settle any disagreement, a criterion seems to be required. Any purported criterion, however, would appear to be based on another criterion, thus requiring an infinite regress of criteria, or else it would be based upon itself, which would be circular. Sextus offered arguments to challenge any claims of dogmatic philosophers to know more than what is evident; and in so doing he presented in one form or another practically all of the skeptical arguments that have ever appeared in subsequent philosophy.

Sextus said that his arguments were aimed at leading people to a state of *ataraxia* (unperturbability). People who thought that they could know reality were constantly disturbed and frustrated. If they could be led to suspend judgment, however, they would find peace of mind. In this state of suspension they would neither affirm nor deny the possibility of knowledge but would remain peaceful, still waiting to see what might develop. The Pyrrhonist did not become inactive in this state of suspense but lived undogmatically according to appearances, customs, and natural inclinations.

MEDIEVAL SKEPTICISM

Pyrrhonism ended as a philosophical movement in the late Roman Empire, as religious concerns became paramount. In the Christian Middle Ages the main surviving form of Skepticism was the Academic, described in St. Augustine's *Contra academicos*. Augustine, before his conversion, had found Cicero's views attractive and had overcome them only through revelation. With faith, he could seek understanding. Augustine's account of Skepticism and his answer to it provided the basis for medieval discussions.

Fideistic views

In Islāmic Spain, where there was more contact with ancient learning, a form of antirational Skepticism developed among Muslim and Jewish theologians. Al-Ghazāli, an Arab theologian of the 11th and early 12th centuries, and his Jewish contemporary Judah ha-Levi (*c.* 1075/*c.* 1085–*c.* 1141), who was a poet and physician as well as a philosopher, offered skeptical challenges (much like those later employed by the occasionalist Nicolas Malebranche and by David Hume) against the contemporary Aristotelians in order to lead people to accept religious truths in mystical faith. This kind of fideism also appears in the late Middle Ages in the German cardinal and philosopher Nicolaus of Cusa's advocacy of learned ignorance as the way to religious knowledge.

MODERN SKEPTICISM

Modern Skepticism emerged in the 16th century, not from medieval views but from the intellectual crises of the Renaissance and Reformation and from the rediscovery of the Skeptical classics. The voyages of exploration; the humanistic rediscovery of the learning of ancient Greece, Rome, and Palestine; and the new science—all combined to undermine confidence in man's accepted picture of the world. The religious controversy between the Protestants and Catholics raised fundamental epistemological issues about the bases and criteria of religious knowledge. At the same time the texts of Cicero and Sextus became available again. (Sextus' *Outlines of Pyrrhonism* [*Hypotyposeis*] was published in Latin in 1562, his *Adversus matematicos* in 1569, and the Greek texts of both in 1621.)

In the Reformation. The fundamental skeptical issues raised by the Reformation appeared in the debate between the outstanding humanist scholar Erasmus and Luther. Erasmus, using Academic skeptical materials, insisted that the issues in dispute could not be resolved, that one should therefore suspend judgment and remain with the church. Luther insisted, on the other hand, that true and certain religious knowledge could and must be gained through conscience. Erasmus' view developed into a Christian Skepticism, accepting traditional Christianity on faith after seeing that no adequate evidence existed. Luther's view, and later that of Calvin, proposed a new criterion—that of inner experience—while the Catholics of the Counter-Reformation employed Pyrrhonian and Academic arguments to undermine the criterion.

Christian Skepticism

Following after Erasmus, another humanist, Giovanni Pico della Mirandola II (nephew of the famous count of the same name) and H.C. Agrippa von Nettesheim, a stormy occult philosopher and physician, employed the skeptical arguments against Scholasticism, Renaissance Naturalism, and many other views to win people to the "true religion." The Catholic scholar Gentian Hervet, in the preface to his 1569 edition of Sextus, saw the Skeptical arguments as the definitive answer to Calvinism and the way to true Christianity.

In the 17th century. The new concern with Skepticism was given a general philosophical formulation by Michel de Montaigne and his cousin Francisco Sanches. Montaigne in *Apology for Raimond Sebond* and Sanches in *Quod nihil scitur*, both written in 1576, explored the human epistemological situation and showed that man's knowledge claims in all areas were extremely dubious. Montaigne recommended living according to nature and custom and accepting whatever God reveals, and Sanches advocated recognizing that nothing can be known and then trying to gain what limited information one can through empirical scientific means.

Montaigne's Skepticism was extremely influential in the early 17th century. His followers, Pierre Charron, J.-P. Camus, La Mothe Le Vayer, and others, further popularized his views. Various French Counter-Reformers used the arguments of Montaigne and Sextus to undermine Calvinism. Montaigne's Skepticism opposed all sorts of disciplines, including the new science, and was coupled with a fideism that many suspected to be insincere.

In the 1620s efforts to refute or mitigate this new Skepticism appeared. A Christian Epicurean, Pierre Gassendi, himself originally a Skeptic, and Marin Mersenne, one of the most influential figures in the intellectual revolution of the times, while retaining epistemological doubts about knowledge of reality yet recognized that science provided useful and important information about the world. The constructive Skepticisms of Gassendi and Mersenne, and later of members of the Royal Society of England like Bishop John Wilkins and Joseph Glanvill, developed the attitude of Sanches into a hypothetical, empirical interpretation of the new science.

René Descartes offered a fundamental refutation of the new Skepticism, contending that, by applying the skeptical method of doubting all beliefs that could possibly be false (due to suffering illusions or being misled by some power), one would discover a truth that is genuinely indubitable, viz., "I think, therefore I am" (*cogito ergo sum*), and that from this truth one could discover the criterion of true knowledge, viz., that whatever is clearly and distinctly conceived is true. Using this criterion, one could then establish: God's existence, that he is not a deceiver, that he guarantees our clear and distinct ideas, and that an external world exists that can be known through mathematical physics. Descartes, starting from Skepticism, claimed to have found a new basis for certitude and for knowledge of reality. Throughout the 17th century Skeptical critics—Mersenne, Gassendi, the reviver of Academic philosophy Simon Foucher, and Pierre-Daniel Huet, one of the most learned men of the age—sought to show that Descartes had not succeeded, and that, if he sincerely followed his skeptical method, his new system could only lead to complete Skepticism. They challenged whether the *cogito* proved anything, or whether it was indubitable; whether Descartes' method could be successfully applied, or whether it was certain; and whether any of the knowledge claims of Cartesianism were *really* true. Nicolas Malebranche, the developer of occasionalism, revised the Cartesian system to meet the Skeptical attacks only to find his efforts challenged by the new Skeptical criticisms of Foucher and by the contention of the Jansenist philosopher Antoine Arnauld that Malebranchism led to a most dangerous Pyrrhonism.

Cartesian Skepticism

Various English philosophers culminating in Locke tried to blunt the force of Skepticism by appealing to common sense and to the "reasonable" man's inability to doubt everything. They admitted that there might not be sufficient evidence to support the knowledge claims extending beyond immediate experience. But this did not actually require that everything be doubted; by using standards of common sense, an adequate basis for many beliefs could be found. Blaise Pascal, who presented the case for Skepticism most forcefully in his *Pensées*, still denied that there can be a complete Skepticism; for nature prevents it. Lacking rational answers to complete Skepticism, man's only recourse lies in turning to God for help in overcoming doubts.

The culmination of 17th-century Skepticism appears in the writings of Pierre Bayle, especially in his monumental *Dictionnaire historique et critique* (1697–1702). Bayle, a superb dialectician, challenged philosophical, sci-

Pierre Bayle

entific, and theological theories, both ancient and modern, showing that they all led to perplexities, paradoxes, and contradictions. He argued that the theories of Descartes, Spinoza, Leibniz, and Malebranche, when skeptically analyzed, cast in doubt all information about the world, even whether a world exists. Bayle skillfully employed Skeptical arguments about such things as sense information, human judgments, logical explanations, and the criteria of knowledge in order to undermine confidence in human intellectual activity in all areas. Bayle suggested that man should abandon rational activity and turn blindly to faith and revelation; he can therefore only follow his conscience without any criterion for determining true faith. Bayle showed that the interpretations of religious knowledge were so implausible that even the most heretical views, like Manichaeism, known for its cosmic dualism of good and evil, and Atheism, made more sense. As a result Bayle's work became "the arsenal of the Enlightenment," and he was regarded as a major enemy of religion.

In the 18th century. Most 18th-century thinkers gave up the quest for metaphysical knowledge after imbibing Bayle's arguments. George Berkeley, an Empiricist and Idealist, fought Skeptical doubts by identifying appearance and reality and offering a spiritualistic metaphysics. He was immediately seen as just another Skeptic since he was denying the world beyond experience.

David Hume

Bayle's chief 18th-century successor was David Hume. Combining empirical and skeptical arguments, Hume charged that neither inductive nor deductive evidence could establish the truth of any matter of fact. Knowledge could only consist of intuitively obvious matters or demonstrable relations of ideas but not of anything beyond experience; the mind can discover no necessary connections within experience nor any root causes of experience. Beliefs about the world are based not upon reason or evidence nor even upon appeal to the uniformity of nature but only on habit and custom. Beliefs cannot be justified. Belief that there is an external world, a self, a God is common; but there is no adequate evidence for it. Although it is natural to hold these convictions, they are inconsistent and epistemologically dubious. "Philosophy would render us entirely Pyrrhonian, were not Nature too strong for it." The beliefs that a man is forced to hold enable him to describe the world scientifically, but when he tries to justify them he is led to complete Skepticism. Before he goes mad with doubts, however, Nature brings him back to common sense, to unjustifiable beliefs. Hume's fideism was a natural rather than a religious one; it is only animal faith that provides relief from complete doubt. The religious context of Skepticism from Montaigne to Bayle had been removed, and man was left with only his natural beliefs, which might be meaningless or valueless.

The central themes in Hume's Skeptical analysis—the basis of induction and causality, knowledge of the external world and the self, proofs of the existence of God—became the key issues of later philosophy. Hume's contemporary Thomas Reid hoped to rebut Hume's Skepticism by exposing it as the logical conclusion of the basic assumptions of modern philosophy from Descartes onward. Such disastrous assumptions should be abandoned for commonsensical principles that have to be believed. As Hume and Kant saw, Reid had not answered Hume's Skepticism but had only sidestepped the issue by appealing to commonsensical living. This provided, however, neither a theoretical basis for beliefs nor a refutation of the arguments that questioned them.

Kant and his critics

Kant saw that Hume had posed a most fundamental challenge to all human knowledge claims. To answer him, it had to be shown not that knowledge is possible but *how* it is possible. Kant combined a Skepticism toward metaphysical knowledge with the contention that certain universal and necessary conditions are involved in having experience and describing it. In terms of these it is possible to have genuine knowledge about the forms of all possible experience, space and time, and about the categories in which all experience is described. Any effort to apply this beyond all possible experience, however, leads into contradictions and Skepticism. It is not possible to know about things-in-themselves nor about the causes of experience.

Though Kant thought that he had resolved the Skeptical problems, some of his contemporaries saw his philosophy as commencing a new Skeptical era. G.E. Schulze (or Schulze-Aenesidemus) a notable critic of Kantianism, insisted that, on Kant's theory, no one could know any objective truths about anything; he could only know the subjective necessity of his views. The Jewish critic Salomon Maimon contended that, though there are such things as a priori concepts, their application to experience is always problematical, and whether they apply can only be found through experience. Hence, the possibility of knowledge can never be established with certainty. Assured truth on the basis of concepts is possible only of human creations, like mathematical ideas, and it is questionable whether these have any objective truth. The thesis that human creativity is the basis of truth, however, was soon to be developed by Johann G. Fichte, a leading German Idealist, as a new way of transcending Skepticism.

Another Skeptical critic of Kant, J.G. Hamann, saw in Hume's and Kant's work a new basis for fideism. If knowledge of reality cannot be gained by rational means, then one must turn to faith. Based on Hume's efforts, Hamann advanced an antirational Skepticism in an effort to convince Kant to become a fideistic Christian. Hamann's kind of fideism was also developed in France by Catholic opponents of the French Revolution and liberalism—like Joseph de Maistre and H.-F.-R. de Lamennais.

Existentialism

In recent and contemporary philosophy. Irrational Skepticism was developed into Existentialism by Søren Kierkegaard in the 19th century. Using traditional Skeptical themes to attack Hegelianism and liberal Christianity, Kierkegaard stressed the need for faith. Only by an unjustified and unjustifiable "leap into faith" could certainty be found—which would then be entirely subjective rather than objective. Modern neo-orthodox and Existentialist theologians have argued that Skepticism highlights man's inability to find any ultimate truth except through faith and commitment. Nonreligious forms of this view have been developed by Existentialist writers like Albert Camus, combining the epistemological Skepticism of Kierkegaard with the religious and value Skepticism of Nietzsche. The rational and scientific examination of the world shows it to be unintelligible and absurd; and if God is dead, as Nietzsche proclaimed, then the world is ultimately meaningless. But it is necessary to struggle with it. It is thus through action and commitment that one finds whatever personal meaning one can, though it has no objective significance.

Other kinds of Skepticism appear in various forms of recent and contemporary philosophy. The English Idealist F.H. Bradley used classical Skeptical arguments in his *Appearance and Reality: A Metaphysical Essay* to contend that the world could not be understood empirically or materialistically; true knowledge could be reached only by transcending the world of appearance.

George Santayana, an American critical Realist, in *Scepticism and Animal Faith,* presented a naturalistic Skepticism. Any interpretation of immediate or intuited experience is open to question. To make life meaningful, however, men make interpretations by "animal faith," according to biological and social factors. The resulting beliefs, though unjustified and perhaps illusory, enable them to persevere and find the richness of life.

Positivism and Analysis

Types of Skepticism also appear in Logical Positivism (see below *Positivism and Logical Empiricism*) and various forms of linguistic philosophy (see below *Analytic and Linguistic Philosophy*). The attack on speculative metaphysics developed by the physicist and early Positivist Ernst Mach, by Bertrand Russell, and by Rudolf Carnap, a leader in the Vienna Circle, where Logical Positivism was nourished, incorporated a Skepticism about the possibility of gaining knowledge beyond experience or logical tautologies. Russell and the important philosopher of science Karl Popper have further stressed the unjustifiability of the principle of induction, and Popper has criticized theories of knowledge based upon empirical verification. A founder of linguistic analysis, Fritz Mauthner, has set forth a Skepticism in which any language is merely relative to its users and thus subjective. Every attempt to tell

what is true just leads one back to linguistic formulations, not to objective states of affairs. The result is a complete Skepticism about reality—a reality that cannot even be expressed except in terms of what he called godless mystical contemplation. Mauthner's linguistic Skepticism bears some affinities to the views expressed in Ludwig Wittgenstein's *Tractatus Logico-Philosophicus.*

CRITICISM AND EVALUATION

In Western thought Skepticism has raised basic epistemological issues. In view of the varieties of human experience, it has questioned whether it is possible to tell which are veridical. The variations that occur in different perceptions of what is presumed to be one object raise the question of which is the correct view. The occurrence of illusory experiences raises the question of whether it is really possible to distinguish illusions and dreams from reality. The criteria employed can be questioned and require justification. On what basis does one tell whether one has the right criteria? By other criteria? Then, are these correct? On what standards? The attempt to justify criteria seems either to lead to an infinite regress or to just stop arbitrarily. If an attempt is made to justify knowledge claims by starting with first principles, what are these based upon? Can it be established that these principles cannot possibly be false? If so, is the proof itself such that it cannot be questioned? If it is claimed that the principles are self-evident, can one be sure of this, sure that one is not deceived? And can one be sure that one can recognize and apply the principles correctly? Through such questioning, Skeptics have indicated the basic problems that an investigator would have to resolve before he could be certain of possessing knowledge; *i.e.,* information that could not possibly be false.

Criticisms

Critics have contended that Skepticism is both a logically and a humanly untenable view. Any attempt to formulate the position will be self-refuting since it will assert at least some knowledge claims about what is supposed to be dubious. Montaigne suggested that the Skeptics needed a nonassertive language, reflecting the claim of Sextus that the Skeptic does not make assertions but only chronicles his feelings. The strength of Skepticism lies not in whether it can be stated consistently but upon the effects of its arguments on dogmatic philosophers. As Hume said, Skepticism may be self-refuting, but in the process of refuting itself it undermines dogmatism. Skepticism, Sextus said, is like a purge that eliminates itself as well as everything else.

Critics have claimed that anyone who tried to be a complete Skeptic, denying or suspending all judgments about ordinary beliefs, would soon be driven insane. Even Hume thought that the complete Skeptic would have to starve to death and would walk into walls or out of windows. Hume, therefore, separated the doubting activity from natural practical activities in the world. Skeptical philosophizing went on in theory, while believing occurred in practice. Sextus and the contemporary Norwegian Skeptic Arne Naess have said, on the other hand, that Skepticism is a form of mental health. Instead of going mad, the Skeptic—without commitment to fixed positions—can function better than the dogmatist.

Some recent thinkers like A.J. Ayer and John Austin have contended that Skepticism is unnecessary. If knowledge is defined in terms of satisfying meaningful criteria, then knowledge is open to all. The Skeptics have raised false problems, because it is, as a matter of fact, possible to tell that some experiences are illusory since we have criteria for distinguishing them from actual events. We do resolve doubts and reach a state of knowledge through various verification procedures, after which doubt is meaningless. Naess, in his book *Scepticism,* has sought to show, however, that, on the standards offered by Ayer and Austin, one can still ask if knowledge claims may not turn out to be false and hence that Skepticism has still to be overcome.

Evaluation

Skepticism throughout history has played a dynamic role in forcing dogmatic philosophers to find better or stronger bases for their views and to find answers to the Skeptical attacks. It has forced a continued reexamination of previous knowledge claims and has stimulated creative thinkers to work out new theories to meet the Skeptical problems.

The history of philosophy can be seen, in part, as a struggle with Skepticism. The attacks of the Skeptics also have served as a check on rash speculation; the various forms of modern Skepticism have gradually eroded the metaphysical and theological bases of European thought. Most contemporary thinkers have been sufficiently affected by Skepticism to abandon the search for certain and indubitable foundations of human knowledge. Instead, they have sought ways of living with the unresolved Skeptical problems through various forms of naturalistic, scientific, or religious faiths. (R.H.P.)

Sophists

The Sophists were certain Greek lecturers, writers, and teachers in the 5th and 4th centuries BC, most of whom travelled about the Greek-speaking world giving instruction in a wide range of subjects in return for fees.

HISTORY OF THE NAME

The term sophist (Greek *sophistes*) had earlier applications. It is sometimes said to have meant originally simply "clever" or "skilled man," but the list of those to whom Greek authors applied the term in its earlier sense makes it probable that it was rather more restricted in meaning. Seers, diviners, and poets predominate, and the earliest Sophists probably were the "sages" in early Greek societies. This would explain the subsequent application of the term to the Seven Wise Men (7th–6th century BC), who typified the highest early practical wisdom, and to Pre-Socratic philosophers generally. When Protagoras, in one of Plato's dialogues (*Protagoras,* 317 a–b) is made to say that, unlike others, he is willing to call himself a Sophist, he is using the term in its new sense of "professional teacher," but he wishes also to claim continuity with earlier sages as a teacher of wisdom. Plato and Aristotle altered the meaning again, however, when they claimed that professional teachers such as Protagoras were not seeking the truth but only victory in debate and were prepared to use dishonest means to achieve it. This produced the sense "captious or fallacious reasoner or quibbler," which has remained dominant to the present day. Finally, under the Roman Empire the term was applied to professors of rhetoric, to orators, and to prose writers generally, all of whom are sometimes regarded as constituting what is now called the Second Sophistic movement (see below).

THE 5TH-CENTURY SOPHISTS

The Sophistic movement

The names survive of nearly 30 Sophists properly so called, of whom the most important were Protagoras, Gorgias, Antiphon, Prodicus, and Thrasymachus. Plato protested strongly that Socrates was in no sense a Sophist—he took no fees, and his devotion to the truth was beyond question. But from many points of view he is rightly regarded as a rather special member of the movement. The actual number of Sophists was clearly much larger than 30, and for about 70 years, until *c.* 380 BC, they were the sole source of higher education in the more advanced Greek cities. Thereafter, at least at Athens, they were largely replaced by the new philosophic schools, such as those of Plato and Isocrates. Plato's dialogue *Protagoras* describes something like a conference of Sophists at the house of Callias in Athens just before the Peloponnesian War. Antimoerus of Mende, described as one of the most distinguished of Protagoras' pupils, is there receiving professional instruction in order to become a Sophist (*Protagoras,* 315 a), and it is clear that this was already a normal way of entering the profession.

Most of the major Sophists were not Athenians, but they made Athens the centre for their activities, although travelling continuously. The importance of Athens was doubtless due in part to the greater freedom of speech prevailing there, in part to the patronage of wealthy men like Callias, and even to the positive encouragement of Pericles, who was said to have held long discussions with Sophists in his house. But primarily the Sophists congregated at Athens because they found there the greatest demand for what they had to offer, namely, instruction to young men, and the extent of this demand followed from the nature

of the city's political life. Athens was a democracy, and although its limits were such that Thucydides could say it was governed by one man, Pericles, it nonetheless gave opportunities for a successful political career to citizens of the most diverse backgrounds, provided they could impress their audiences sufficiently in the council and the assembly. After Pericles' death this avenue became the highroad to political success.

The Sophistic teaching

The Sophists taught men how to speak and what arguments to use in public debate. A Sophistic education was increasingly sought after both by members of the oldest families and by aspiring newcomers without family backing. The changing pattern of Athenian society made merely traditional attitudes in many cases no longer adequate. Criticizing such attitudes and replacing them by rational arguments held special attraction for the young, and it explains the violent distaste which they aroused in traditionalists. Plato thought that much of the Sophistic attack upon traditional values was unfair and unjustified. But even he learned at least one thing from the Sophists—if the older values were to be defended, it must be by reasoned argument, not by appeals to tradition and unreflecting faith.

Seen from this point of view, the Sophistic movement was a valuable function of Athenian democracy in the 5th century BC. It offered an education designed to facilitate and promote success in public life. All of the Sophists appear to have provided a training in rhetoric and in the art of speaking, and the Sophistic movement, responsible for large advances in rhetorical theory, contributed greatly to the development of style in oratory. In modern times the view occasionally has been advanced that this was the Sophists' only concern. But the range of topics dealt with by the major Sophists makes this unlikely, and even if success in this direction was their ultimate aim, the means they used were surely as much indirect as direct, for the pupils were instructed not merely in the art of speaking, but in grammar; in the nature of virtue (*aretē*) and the bases of morality; in the history of society and the arts; in poetry, music, and mathematics; and also in astronomy and the physical sciences. Naturally the balance and emphasis differed from Sophist to Sophist, and some offered wider curricula than others. But this was an individual matter, and attempts by earlier historians of philosophy to divide the Sophistic movement into periods in which the nature of the instruction was altered are now seen to fail for lack of evidence. The 5th-century Sophists inaugurated a method of higher education that in range and method anticipated the modern humanistic approach inaugurated or revived during the Renaissance.

NATURE OF SOPHISTIC THOUGHT

A question still discussed is whether the Sophists in general had any real regard for truth or whether they taught their pupils that truth was unimportant compared with success in argument. Plato's hostile judgment on both counts is still frequently repeated without question. The Platonic writings make frequent reference to what Plato calls "eristic" (Greek *eristikos,* "fond of wrangling") and "antilogic"; the two often have been incorrectly treated as identical. Eristic, for Plato, consists in arguments aimed at victory rather than at truth. Antilogic involves the assignment to any argument of a counterargument that negates it, with the implication that both argument and counterargument are equally true. Antilogic in this sense was especially associated with Protagoras; but Plato, no doubt correctly, attributes its use to other Sophists as well. He regards the use of antilogic as essentially eristic, whether it be used to silence an opponent by making his position seem self-contradictory, or whether it be used mechanically to negate any proposition put forward in debate. He concludes that the widespread use of antilogic is evidence that Sophists had no real regard for the truth, which must itself be free from antilogic.

But Plato himself believed, for much or possibly all of his life, that the phenomenal world was essentially antilogical inasmuch as no statement about it could be made possessing a greater degree of truth than the contradictory of that statement. For example, if a man is tall in relation to one object, he will be short in relation to another object. In so characterizing the phenomenal world, Plato certainly did not wish to be called eristic—he regarded the application of antilogic to the description of the phenomenal world as an essential preliminary to the search for the truth residing in the Platonic Forms, which are themselves free from antilogic.

Seen in this perspective, the Sophistic use of antilogic must be judged less harshly. To the extent that it was used irresponsibly to secure success in debate it was eristic, and the temptation so to use it must often have arisen. But where it was invoked in the sincere belief that antilogic elements were indeed involved, or where it was used for analyzing a complex situation in order to reveal its complexity, then antilogic was in no way inconsistent with devotion to truth. This raises the question to what extent the Sophists possessed any general view of the world or gave expression to any genuine philosophical views, whether original or derived. Ancient writers, influenced by Plato and Aristotle, seem to have excluded the Sophists, apart from Protagoras, for their schematized accounts of early Greek thinkers. Modern writers have frequently maintained that, whatever else they were, the Sophists were in no sense philosophers. Even those who acknowledge the philosophical interest of certain particular doctrines attributed to individual Sophists often tend to regard these as exceptions and claim that, inasmuch as the Sophists were not a school but only independent teachers and writers, as a class they were not philosophers. Two questions are involved: whether the Sophists held common intellectual doctrines and whether some or all of these could actually be termed philosophical.

Philosophy of the Sophists

Among moderns, Hegel was one of the first to reinsert the Sophists into the history of Greek philosophy. He did so within the framework of his own dialectic, in which every thesis invokes its own opposite, or antithesis; thus he treated the Sophists as representing the antithesis to the thesis of the group of philosophers known collectively as the Pre-Socratics. Pre-Socratics such as Thales, Heracleitus, and Parmenides sought the truth about the external world with a bold enthusiasm that produced a series of explanations, each claiming to be correct. None of these explanations of the physical world paid attention to the observer and each was driven to reject more and more of the phenomenal world itself as unreal. Finally, with the Eleatics, a 5th-century school at Elea in Italy that held that reality is a static one, of which Parmenides and Zeno are representatives, little or nothing of the phenomenal world was left as real. This trend in turn produced a growing distrust of the power of human beings to attain knowledge of the ultimate basis of natural phenomena. Philosophy had reached an impasse, and there was a danger of complete skepticism. Such an extreme position, according to Hegel's view, provoked the "antithesis" of the Sophistic movement, which rejected the "thesis" of the objectivists and concentrated attention upon man rather than upon nature. To Hegel, the Sophists were subjective Idealists, holding that reality is only minds and their contents, and so philosophy could move forward by turning its attention to the subjective element in knowing. Reflection upon the contrast between the thought of the Sophists and that of their predecessors produced the "syntheses" of Plato and Aristotle.

Whether any of the Sophists actually were subjective Idealists may be doubted. The conclusion depends in part on whether Protagoras held that phenomena had subjective existence only, or whether he thought that all things perceived had objective existence but were perceived differently according to the nature of the percipient and their relation to him—*i.e.,* whether he interpreted phenomena subjectively or relativistically. It is fairly clear, however, that the Sophists did concentrate very largely upon man and human society, upon questions of words in their relations to things, upon issues in the theory of knowledge, and upon the importance of the observer and the subjective element in reality and in the correct understanding of reality.

This emphasis helps to explain the philosophical hostility of Plato and Aristotle. Particularly in the eyes of Plato,

anyone who looks for the truth in phenomena alone, whether he interprets it subjectively or relativistically, cannot hope to find it there; and his persistence in turning away from the right direction virtually amounts to a rejection of philosophy and of the search for truth. Many a subsequent thinker for whom metaphysics, or the investigation of the deepest nature of reality, was the crowning achievement of philosophy has felt with Plato that the Sophists were so antimetaphysical that they have no claim to rank as philosophers. But in a period when, for many philosophers, metaphysics is no longer the most important part of philosophy and is even for some no part at all, there is growing appreciation of a number of problems and doctrines recurring in the discussions of the Sophists in the 5th and 4th centuries BC. In the 18th and early 19th centuries the Sophists were considered charlatans. Their intellectual honesty was impugned, and their doctrines were blamed for weakening the moral fibre of Greece. The charge was based on two contentions, both correct: first, that many of the Sophists attacked the traditionally accepted moral code; and second, that they explored and even commended alternative approaches to morality that would condone or allow behaviour of a kind inadmissible under the stricter traditional code.

Moral views of the Sophists

Much less weight has been attached to these charges since about the mid-19th century. First, many of the attacks on the traditional morality were in the name of a new morality that claimed to be of greater validity. Attacks upon particular doctrines often claimed that accepted views should be abandoned as morally defective. Furthermore, even when socially disfavoured action seemed to be commended, this was frequently done to introduce a principle necessary in any satisfactory moral theory. Thus when Thrasymachus in the first book of Plato's *Republic* argues that justice is unwarranted when it merely contributes to another's good and not to the good of the doer, Plato agrees. Finally, there is no evidence that any of the Sophists were personally immoral or that any of their pupils were induced to immoral actions by Sophistic teaching. The serious discussion of moral problems and the theory of morality tends to improve behaviour, not to corrupt it.

WRITINGS

In addition to their teaching, the Sophists wrote many books, the titles of which are preserved by writers such as Diogenes Laërtius, who probably derived them from library catalogues. It has usually been supposed that the writings themselves hardly survived beyond the period of Plato and Aristotle, but this view requires modification in the light of papyrus finds, admittedly few, that were copied from Sophistic writings in the Christian Era. It also has been possible to identify in the works of later writers certain imitations or summaries of 5th-century Sophistic writers, whose names are unknown. The most important of these are the discussion of law in the *Protrepticus*, or "Exhortation to Philosophy," by the 3rd-century-AD Syrian Neoplatonist Iamblichus, and the so-called *Dissoi logoi* found in the manuscripts of Sextus Empiricus (3rd century AD). This evidence suggests that while most later writers took their accounts of the Sophists from earlier writers, especially from Plato, the original writings did in many cases survive and were consulted.

PARTICULAR DOCTRINES

As part of his defense of the Sophists against the charge of immoral teachings, the English historian George Grote (1794–1871) maintained that they had nothing in common with each other except their profession, as paid teachers qualifying young men to think, speak, and act with credit to themselves as citizens. This denial of common doctrines cannot be sustained—the evidence is against it. While the Sophists were not a sect, with a set of obligatory beliefs or doctrines, they had a common interest in a whole series of questions to which they sought to apply solutions along certain clearly defined lines.

There are great difficulties, however, in the precise reconstruction of individual Sophistic doctrines. No complete writings survive from any of the Sophists to check the accounts found in Plato, and later writers were often, but not always, dependent upon what they found in Plato. Plato doubtless knew well the doctrines of individual Sophists; but he was writing for those to whom these doctrines were already well known, and he was always more interested in following the argument where it led than in providing precise statements of other people's views for the sake of posterity. Consequently, almost everything that is said about particular Sophistic doctrines is subject to controversy.

Theoretical issues. Relativism and skepticism have often been regarded as common features of the Sophistic movement as a whole. But it was early pointed out that only in Protagoras and Gorgias is there any suggestion of a radical skepticism about the possibility of knowledge; and even in their case Sextus Empiricus, in his discussion of skepticism, is probably right when he declares that neither was really a skeptic. Protagoras does seem to have restricted knowledge to sense experience, but he believed emphatically that whatever was perceived by the senses was certainly true. This led him to assert that the tangent does not touch the circle at a point only, but along a definite length of the circumference; clearly he was referring to human perception of drawn tangents and circles. Gorgias, who claimed that nothing exists, or if it does exist it cannot be known, or if it exists and is knowable it cannot be communicated to another, has often been accused of denying all reality and all knowledge. Yet he also seems to have appealed in his very discussion of these themes to the certainty of perceived facts about the physical world; *e.g.*, that chariots do not race across the sea. Others dismiss his whole thesis as a satire or joke against philosophers.

Relativism and skepticism

Probably neither view is correct. What Gorgias seems to have been attacking was not perceived reality nor one's power to perceive it but the attempt to assign existence or nonexistence (with the metaphysical implications of such an operation) to what we perceive around us. There is evidence that other Sophists (*e.g.*, Hippias) were interested in questions of this kind, and it is likely that they were all concerned to some degree with rejecting claims of any nonsensible existence, such as those of the Eleatics. The Sophists, in fact, were attempting to explain the phenomenal world without appealing to any principles outside of phenomena. They believed that this could be done by including the observer within the phenomenal world. Their refusal to go beyond phenomena was, for Plato, the great weakness in their thinking.

A second common generalization about the Sophists has been that they represent a revolt against science and the study of the physical world. The evidence is against this, inasmuch as for Hippias, Prodicus, Gorgias, and Protagoras there are records of a definite interest in questions of this kind. The truth is rather that they were in revolt against attempts to explain the physical world by appeals to principles that could not be perceived by the senses; and instead of framing new "objective" explanations, they attempted to explain things, where explanation was required, by introducing the perceiver as one element in the perceptual situation.

One of the most famous doctrines associated with the Sophistic movement was the opposition between nature and custom or convention in morals. It is probable that the antithesis did not originate in Sophistic circles but was rather earlier; but it was clearly very popular and figured largely in Sophistic discussions. The commonest form of the doctrine involved an appeal from conventional laws to supposedly higher laws based on nature. Sometimes these higher laws were invoked to remedy defects in actual laws and to impose more stringent obligations; but usually it was in order to free men from restrictions unjustifiably imposed by human laws that the appeal to nature was made. In its extreme form the appeal involved the throwing off of all restraints upon self-interest and the desires of the individual (*e.g.*, the doctrine of Callicles in Plato's *Gorgias* that might, if one possesses it, is actually right), and it was this, more than anything else, that gave support to charges against the Sophists of immoral teaching. On other occasions the terms of the antithesis were reversed and human laws were explicitly acclaimed as superior to

Nature and law

the laws of nature and as representing progress achieved by human endeavour. In all cases the laws of nature were regarded not as generalized descriptions of what actually happens in the natural world (and so not like the laws of physics to which no exceptions are possible) but rather as norms that people ought to follow but are free to ignore. Thus the appeal to nature tended to mean an appeal to the nature of man treated as a source for norms of conduct.

To Greeks this appeal was not very novel. It represented a conscious probing and exploration into an area wherein, according to their whole tradition of thought, lay the true source for norms of conduct. If Callicles in Plato's *Gorgias* represents a position actually held by a living Sophist when he advocates free rein for the passions, then it was easy for Plato to argue in reply that the nature of man, if it is to be fulfilled, requires organization and restraint in the license given to the desires of particular aspects of it; otherwise the interests of the whole will be frustrated. Both Plato and Aristotle, in basing so much of their ethics on the nature of man, are only following up the approach begun by the Sophists.

Religion and society

Humanistic issues. The Sophists have sometimes been characterized by their attacks on the traditional religious beliefs of the Greeks. It is true that more than one Sophist seems to have faced prosecution for impiety, as did Socrates also. Protagoras wrote "concerning the gods, I cannot know either that they exist or that they do not exist nor what they are like in form," and Prodicus offered a sociological account of the development of religion. Critias went further when he supposed that the gods were deliberately invented to inspire fear in the evildoer. It is thus probably correct to say that the tendency of much Sophistic thought was to reject the traditional doctrines about the gods. Indeed this follows almost inevitably if the supposition is correct that all the Sophists were attempting to explain the phenomenal world from within itself, while excluding all principles or entities not discernible in phenomena. But in their agnostic attitudes toward the Olympian deities the Sophists were probably at one with most of the Pre-Socratic philosophers of the 6th and 5th centuries and also with most thinking people living toward the end of the 5th century. It is thus probably misleading to regard them as revolutionary in their religious beliefs.

The importance the Sophists attached to man meant that they were extremely interested in the history and organization of human societies. Here again most is known about Protagoras, and there is a danger of treating his particular doctrines as typical of the Sophistic movement as a whole. In the 5th century, human history was very commonly seen in terms of a decline from an earlier golden age. Another view supposed that there were recurring cycles in human affairs according to which a progression from good to bad would give way to one from bad to good. The typical Sophistic attitude toward society rejected both of these views in favour of one that saw human history in terms of progress from savagery to civilization. In a famous myth Protagoras explained how man achieved civilized society first with the aid of arts and crafts and then by gaining a sense of respect and justice in the ordering of his affairs. The general thinking of most of the Sophists seems to have been along similar lines.

One of the most distinctive Sophistic tenets was that virtue can be taught, a position springing naturally from the Sophists' professional claim to be the teachers of young men. But the word virtue (*aretē*) implied both success in living and the qualities necessary for achieving such success, and the claim that *aretē* could be taught by the kind of teaching that the Sophists offered had far-ranging implications. It involved the rejection of the view that *aretē* came only by birth—for example, by being born a member of a noble family—and it involved also the rejection of the doctrine that *aretē* was a matter of the chance occurrence of specified qualities in particular individuals. *Aretē,* in the Sophists' view, was the result of known and controllable procedures, a contention of profound importance for the organization of society. Moreover, what can be taught has some relation to what can be known and understood. The belief that teaching of a high intellectual calibre could produce success both for the individual and for governments has had a profound influence upon the subsequent history of education. Once again, it is through the acceptance of this doctrine by Plato and Aristotle that the Sophistic position came to be part of subsequent humanist tradition.

THE SECOND SOPHISTIC MOVEMENT

It is a historical accident that the name "Sophist" came to be applied to the Second Sophistic movement. Greek literature underwent a period of eclipse during the 1st century BC and under the early Roman Empire. But Roman dominance did not prevent a growing interest in sophistic oratory in the Greek-speaking world during the 1st century AD. This oratory aimed merely at instructing or interesting an audience and had of necessity no political function. But it was based on elaborate rules and required a thorough knowledge of the poets and prose writers of antiquity. Training was provided by professional teachers of rhetoric who claimed the title of Sophists, just as the 5th-century Sophists had adopted a name already used by others.

The revival of the Greek spirit under Hadrian and other emperors in the 2nd century AD who were also admirers of Greek culture found expression in a fresh flowering of Greek prose following principles developed and applied by the professors of rhetoric in the 1st century AD. Hence a group of Greek prose writers in the 2nd century AD were regarded as constituting the Second Sophistic movement. This was a backward-looking movement that took as its models Athenian writers of the 5th and 4th centuries BC; hence the label "Atticists" (Greek *Attikos,* "Athenian") applied to some of its leading members. The limits of the movement were never clear. It is usually taken to include Polemon of Athens, Herodes Atticus, Aelius Aristides, Maximus of Tyre, and the group of Philostrati. Dio Chrysostom of Prusa is often included, although others would regard him as preparing the way for the main period. Other writers, like Lucian, Aelian, and Alciphron, were influenced by the movement even if not properly members of it; and the writers of prose romances, such as Longus and Heliodorus, and the historians Dio Cassius and Herodian are also associated with the general trend. By the 3rd century AD, however, its impulse was weakening, and it was shortly no longer distinguishable within the general stream of Greek literature. (G.B.K.)

Stoicism

Stoicism, a school of thought that flourished in Greek and Roman antiquity, was one of the loftiest and most sublime philosophies in the record of Western civilization. In urging participation in the affairs of man, Stoics have always believed that the goal of all inquiry is to provide man with a mode of conduct characterized by tranquillity of mind and certainty of moral worth.

NATURE AND SCOPE OF STOICISM

For the early Stoic philosopher, as for all the post-Aristotelian schools, knowledge and its pursuit are no longer held to be ends in themselves. The Hellenistic Age was a time of transition, and the Stoic philosopher was perhaps its most influential spokesman. A new culture was in the making. The heritage of an earlier period, with Athens as its intellectual leader, was to continue, but to undergo many changes. If, as with Socrates, to know is to know oneself, rationality as the sole means by which something outside of the self might be achieved may be said to be the hallmark of Stoic belief. As a Hellenistic philosophy, Stoicism presented an *ars vitae,* a way of accommodation for people to whom the human condition no longer appeared as the mirror of a universal, calm, and ordered existence. Reason alone could reveal the constancy of cosmic order and the originative source of unyielding value; thus, reason became the true model for human existence. To the Stoic, virtue is an inherent feature of the world, no less inexorable in relation to man than are the laws of nature.

Fundamental tenets

The Stoics believed that perception is the basis of true knowledge. In logic, their comprehensive presentation of the topic is derived from perception, yielding not only the

judgment that knowledge is possible but also that *certain* knowledge is possible, on the analogy of the incorrigibility of perceptual experience. To them, the world is composed of material things, with some few exceptions (*e.g.*, meaning), and the irreducible element in all things is right reason, which pervades the world as divine fire. Things, such as material, or corporeal, bodies, are governed by this reason or fate, in which virtue is inherent. The world in its awesome entirety is so ruled as to exhibit a grandeur of orderly arrangement that can only serve as a standard for mankind in the regulation and ordering of his life. Thus, the goal of man is to live according to nature, in agreement with the world design. Stoic moral theory is also based on the view that the world, as one great city, is a unity. Man, as a world citizen, has an obligation and loyalty to all things in that city. He must play an active role in world affairs, remembering that the world exemplifies virtue and right action. Thus, moral worth, duty, and justice are singularly Stoic emphases, together with a certain sternness of mind. For the moral man neither is merciful nor shows pity, because each suggests a deviation from duty and from the fated necessity that rules the world. Nonetheless—with its loftiness of spirit and its emphasis on man's essential worth—the themes of universal brotherhood and the benevolence of divine nature make Stoicism one of the most appealing of philosophies.

Rival schools in antiquity

Its chief competitors in antiquity were: (1) Epicureanism, with its doctrine of a life of withdrawal in contemplation and escape from worldly affairs and its belief that pleasure, as the absence of pain, is the goal of man; (2) Skepticism, which rejected certain knowledge in favour of local beliefs and customs, in the expectation that these guides would provide man with the quietude and serenity that the dogmatic philosopher (*e.g.*, the Stoic) could not hope to achieve; and (3) Christianity, with its hope of personal salvation provided by an appeal to faith as an immanent aid to human understanding and by the beneficent intervention of a merciful God.

Along with its rivals, Stoicism enabled the individual to better order his own life and to avoid the excesses of human nature that promote disquietude and anxiety. It was easily the most influential of the schools from the time of its founding through the first two centuries AD, and it continued to have a marked effect on later thought. During the late Roman and medieval periods, elements of Stoic moral theory were known and used in the formulation of Christian, Jewish, and Muslim theories of man and nature, of the state and society, and of law and sanctions—*e.g.*, in the works of Marcus Cicero, Roman statesman and orator; in Lactantius, often called the "Christian Cicero"; and in A.M.S. Boethius, a scholar transitional to the Middle Ages. In the Renaissance, Stoic political and moral theory became more popular to theorists of natural law and political authority and of educational reform—*e.g.*, in Hugo Grotius, a Dutch jurist and statesman, and in Philipp Melanchthon, a major Reformation scholar. More recently, Stoicism has become popular again for its insistence on the value of the individual and the place of value in a world of strife and uncertainty—*e.g.*, in Existentialism and in Neo-orthodox Protestant theology. Stoicism also plays an important role in reassessments of the history of logic—*e.g.*, in Jan Łukasiewicz, a Polish logician, and in William and Martha Kneale, mid-20th-century British logicians.

ANCIENT STOICISM

Historical and philosophical setting

Early Greek Stoicism. With the death of Aristotle (322 BC) and that of Alexander the Great (323 BC), the greatness of the life and thought of the Greek city-state (polis) ended. With Athens no longer the centre of worldly attraction, its claim to urbanity and cultural prominence passed on to other cities—to Rome, to Alexandria, and to Pergamum. The Greek polis gave way to larger political units; local rule was replaced by that of distant governors. The earlier distinction between Greek and barbarian was destroyed; provincial and tribal loyalties were broken apart, first by Alexander and then by Roman legions. The loss of freedom by subject peoples further encouraged a deterioration of the concept of the freeman and resulted in the rendering of obligation and service to a ruler whose moral force held little meaning. The earlier intimacy of order, cosmic and civic, was now replaced by social and political disorder; and traditional mores gave way to uncertain and transient values.

Stoicism had its beginnings in a changing world, in which earlier codes of conduct and ways of understanding proved no longer suitable. But it was also influenced by tenets of the older schools. The earliest Greek philosophers, the Milesians, had called attention to cosmic order and the beauty of nature. Later, the monist Parmenides of Elea stressed the power of reason and thought, whereas Heracleitus of Ephesus, precursor of the philosophy of becoming, had alluded to the constancy of change and the omnipresence of divine fire, which illumined all things. A deeper understanding of man himself came with Socrates, symbol of the philosophic man, who personified *sophia* and *sapientia* (Greek and Latin: "wisdom"). Of the several schools of philosophy stemming from Socrates, the Cynic and Megarian schools were influential in the early development of Stoic doctrine: the Cynics for their emphasis on the simple life, unadorned and free of emotional involvement; and the Megarians for their study of dialectic, logical form, and paradoxes.

The Early Stoa

Stoicism takes its name from the place where its founder, Zeno of Citium (Cyprus), customarily lectured—the Stoa Poikile (Painted Colonnade). Zeno, who flourished in the early 3rd century BC, showed in his own doctrines the influence of earlier Greek attitudes, particularly those mentioned above. He was apparently well versed in Platonic thought, for he had studied at Plato's Academy both with Xenocrates of Chalcedon and with Polemon of Athens, successive heads of the Academy. Zeno was responsible for the division of philosophy into three parts: logic, physics, and ethics. He also established the central Stoic doctrines in each part, so that later Stoics were to expand rather than to change radically the views of the founder. With some exceptions (in the field of logic), Zeno thus provided the following themes as the essential framework of Stoic philosophy: logic as an instrument and not as an end in itself; human happiness as a product of life according to nature; physical theory as providing the means by which right actions are to be determined; perception as the basis of certain knowledge; the wise man as the model of human excellence; Platonic Ideas—or the abstract Forms that things of the same genus share—as being unreal; true knowledge as always accompanied by assent; the fundamental substance of all existing things as being a divine fire, the universal principles of which are (1) passive (matter) and (2) active (reason inherent in matter); belief in a world conflagration and renewal; belief in the corporeality of all things; belief in the fated causality that necessarily binds all things; cosmopolitanism, or cultural outlook transcending narrower loyalties; and man's obligation, or duty, to choose only those acts that are in accord with nature, all other acts being a matter of indifference.

Cleanthes of Assos, who succeeded Zeno as head of the school, is best known for his *Hymn to Zeus*, which movingly describes Stoic reverence for the cosmic order and the power of universal reason and law. The third head of the school, Chrysippus of Soli, who lived to the end of the 3rd century, was perhaps the greatest and certainly the most productive of the early Stoics. He devoted his considerable energies to the almost complete development of the Zenonian themes in logic, physics, and ethics. In logic particularly, he defended against the Megarian logicians and the Skeptics such concepts as certain knowledge, comprehensive presentation, proposition and argument, truth and its criterion, and assent. His work in propositional logic, in which unanalyzed propositions joined by connectives are studied, made important contributions to the history of ancient logic and is of particular relevance to more recent developments in logic.

In physics, Chrysippus was responsible for the attempt to show that fate and free will are not mutually exclusive conceptual features of Stoic doctrine. He further distinguished between "whole" and "all," or "universe," arguing that the whole is the world, while the all is the external void together with the world. Zeno's view of the origin of man

as providentially generated by "fiery reason" out of matter was expanded by Chrysippus to include the concept of self-preservation, which governs all living things. Another earlier view (Zeno's), that of nature as a model for life, was amplified first by Cleanthes and then by Chrysippus. The Zenonian appeal to life "according to nature" had evidently been left vague, because to Cleanthes it seemed necessary to speak of life in accord with nature conceived as the world at large (the cosmos), whereas Chrysippus distinguished between world nature and human nature. Thus, to do good is to act in accord with both human and universal nature. Chrysippus also expanded the Stoic view that seminal reasons were the impetus for animate motion.

He established firmly that logic and (especially) physics are necessary and are means for the differentiation of goods and evils. Thus, a knowledge of physics (or theology) is required before an ethics can be formulated. Indeed, physics and logic find their value chiefly in this very purpose. Chrysippus covered almost every feature of Stoic doctrine and treated each so thoroughly that the essential features of the school were to change relatively little after his time.

The Middle Stoa

Later Roman Stoicism. The Middle Stoa, which flourished in the 2nd and early 1st centuries BC, was dominated chiefly by two men of Rhodes: Panaetius, its founder, and his disciple Poseidonius. Panaetius organized a Stoic school in Rome before returning to Athens, and Poseidonius was largely responsible for an emphasis on the religious features of the doctrine. Both were antagonistic to the ethical doctrines of Chrysippus, who, they believed, had strayed too far from the Platonic and Aristotelian roots of Stoicism. It may have been because of the considerable time that Panaetius and Poseidonius lived in Rome that the Stoa there turned so much of its emphasis to the moral and religious themes within the Stoic doctrine. Panaetius was highly regarded by Cicero, who used him as a model for his own work. Poseidonius, who had been a disciple of Panaetius in Athens, taught Cicero at his school at Rhodes and later went to Rome and remained there for a time with Cicero. If Poseidonius admired Plato and Aristotle, he was particularly interested—unlike most of his school—in the study of natural and providential phenomena. In presenting the Stoic system in the second book of *De natura deorum* (45 BC), Cicero most probably followed Poseidonius. Because his master, Panaetius, was chiefly concerned with concepts of duty and obligation, it was his studies that served as a model for the *De officiis* (44 BC) of Cicero. Hecaton, another of Panaetius' students and an active Stoic philosopher, also stressed similar ethical themes.

If Chrysippus is to be commended for his diligence in defending Stoic logic and epistemology against the Skepticism of the New Academy (3rd–2nd century BC), it was chiefly Panaetius and Poseidonius who were responsible for the widespread popularity of Stoicism in Rome. It was precisely their turning of doctrine to themes in moral philosophy and natural science that appealed to the intensely practical Romans. The times perhaps demanded such interests, and with them Stoicism was to become predominantly a philosophy for the individual, showing how—given the vicissitudes of life—one might be stoical. Law, world citizenship, nature, and the benevolent workings of Providence and the divine reason were the principal areas of interest of Stoicism at this time.

The Late Stoa

These tendencies toward practicality are also well illustrated in the later period of the school (in the first two centuries AD) in the writings of Lucius Seneca, a Roman statesman; of Epictetus, a slave freed by Nero; and of Marcus Aurelius, a Roman emperor. Both style and content in the *Libri morales* (Eng. trans., *Moral Essays*) and *Epistulae morales* (*Moral Letters*) of Seneca reinforce the new direction in Stoic thought. The *Encheiridion* (*Manual*) of Epictetus and the *Meditations* of Marcus Aurelius furthered the sublime and yet personal consolation of the Stoic message and increasingly showed the strength of its rivalry to the burgeoning power of the new Christianity. The mark of a guide, of the religious teacher, is preeminent in these writings. It is difficult to establish with any precision, however, the extent of Stoic influence by the time of the first half of the 2nd century AD. So popular had these ideas become that many specifically Stoic terms (viz., right reason, comprehension, assent, indifference, *logos*, natural law, and the notion of the wise man) commonly were used in debate and intellectual disputes.

Stoic elements in Pauline and patristic thought. There is much disagreement as to the measure of Stoic influence on the writings of St. Paul, the Apostle of Christ. At Tarsus, Paul certainly had opportunities for hearing Stoic lectures on philosophy. And it may be that his discussion of nature and the teaching of it (I Cor. 11:14) is Stoic in origin, for it has a parallel in the *Manual* of Epictetus 1.16, 10. Although not a Stoic technical term, *syneidēsis*, which Paul used as "conscience," was generally employed by Stoic philosophers. In I Cor. 13 and in the report of Paul's speech at Athens (Acts 17), there is much that is Hellenistic, more than a little tinged by Stoic elements—*e.g.*, the arguments concerning man's natural belief in God and the belief that man's existence is in God.

The assimilation of Stoic elements by the Church Fathers was generally better understood by the 4th century. Stoic influence can be seen, for example, in the relation between reason and the passions in the works of St. Ambrose, one of the great scholars of the church, and of Marcus Minucius Felix, a Christian Apologist. Each took a wealth of ideas from Stoic morality as Cicero had interpreted it in *De officiis*. In general, whereas the emerging Christian morality affirmed its originality, it also assimilated much of the pagan literature, the more congenial elements of which were essentially Stoic.

Stoic concepts in Tertullian and Origen

Earlier, in the 3rd century, Quintus Tertullian, often called the father of Latin Christian literature, seems to have been versed in Stoic philosophy; *e.g.*, in his theory of the agreement between the supernatural and the human soul, in his use of the Stoic tenet that from a truth there follow truths, and in his employment of the idea of universal consent. Even in his polemical writings, which reveal an unrelenting hostility to pagan philosophy, Tertullian showed a fundamental grasp and appreciation of such Stoic themes as the world *logos* and the relation of body to soul. This is well illustrated in his argument against the Stoics, particularly on their theme that God is a corporeal being and identified with reason as inherent in matter—also to be found in his polemics against Marcion, father of a heretical Christian sect, and against Hermogenes of Tarsus, author of an important digest of rhetoric. Yet in his doctrine of the Word, he appealed directly to Zeno and Cleanthes of the Early Stoa. Another important polemic against the Stoics is found in the treatise *Contra Celsum*, by Origen, the most influential Greek theologian of the 3rd century, in which he argued at some length against Stoic doctrines linking God to matter.

Also, St. Cyprian, bishop of Carthage in the 3rd century, revealed the currency of Stoic views; *e.g.*, in his *Ad Demetrianum*, a denunciation of an enemy to Christianity, in which Cyprian castigates the ill treatment of slaves, who, no less than their masters, are formed of the same matter and endowed with the same soul and live according to the same law. The beliefs in the brotherhood of man and in the world as a great city, commonly found in early Christian literature, were current Stoic themes. The Christian attitude appears in what St. Paul said of Baptism: "You are all sons of God through Faith. For as many of you as were baptized into Christ have put on Christ" (Gal. 3:26–27).

STOICISM IN MEDIEVAL AND MODERN PHILOSOPHY

Stoic concepts in Boethius and Lactantius

Stoic undercurrents in medieval thought. During the period when Christian institutions and doctrines were developing (AD 230–1450), Stoicism continued to play a popular role. The *De consolatione philosophiae* (524) of Boethius (died AD 524/525) was widely known and appreciated as a discourse on the mysterious questions of the nature of good and evil, of fortune, chance, or freedom, and of divine foreknowledge. If the plan of Boethius was to serve as an interpreter of Plato and Aristotle, he succeeded only in working through some logical theories of Aristotle, together with several commentaries on those theories. In the *Consolatione*, however, the themes are quite different;

in the fifth book, for example, he attempted to resolve the apparent difficulty of reconciling human freedom with the divine foreknowledge, a problem that among Stoic thinkers—though by no means uniquely among them—had been in general currency for a long time. This work of emancipation from worldly travail through the glories of reason and philosophy, which included Stoic doctrines as found in the writings of Cicero and Seneca, was much more influential for later medieval thought than that of Lactantius, of the late 3rd to early 4th century, who was largely concerned with the writing of a history of religion—a summary statement of Christian doctrine and life from earliest times. Lactantius also wrote a not unimportant work, *De ira Dei* (313), on the possibility of anger in God. It poses a problem of how to deal with the essentially Greek, or philosophic, view that God cannot feel anger because he is not subject to passions and that *apatheia* ("apathy," or "imperturbableness") is not merely the mark of the wise man but is also a divine attribute. This view, which had been most thoroughly developed among Stoic thinkers and particularly by Epictetus, raised a peculiarly Christian problem, the concern of the power of God to reward the righteous and punish the transgressor; thus, it challenged the very idea of Providence. Other manifestations of anthropopathism, the attributing of human feelings to God, had also been charged against the early Christian religionists; and the writers of the time—Lactantius and Tertullian among them—took great pains to refute the largely Stoic formulations of these charges. Although the refutations took the form—in St. Augustine, for example—of denying that the wrath of God is a perturbation of the soul and of holding that it is rather a judgment, the concept of the divine essence excludes all passions. Within the monastic tradition, there remained more than a residue of concern over apathy as a divine attribute and as a model for the truly religious.

Stoicism in legal and political theory

Other significant Stoic influences appeared in medieval discussions of the popular origin of political authority, *res publica* and *res populi,* and on the distinctions made in law between *jus naturale, jus gentium, jus civile*—doctrines of Stoic origin—found in 3rd-century Roman juridical texts gathered together by Isidore of Seville (died AD 636), a Spanish encyclopaedist and theologian. The Stoic belief—as against Aristotle—that men are by nature equal was an integral part of the knowledge that certain rules of law are universally recognized, laws that all people might naturally follow. In this way, the Romans—whose genius lay in organization and in law—fostered the conception of natural, or common, law, which reason was supposed to make evident to all men. Thus, in the second half of the 11th century, the Stoic texts of Cicero and Seneca became important doctrinal sources for the initial discussions of social and political philosophy. These early theories of law, of the natural equality of men, and of the rights of prince and populace were to become the basis for 13th-century systems of social and political privilege and obligation.

In the 12th century, John of Salisbury, an English critical scholar, produced, in his *Policraticus* (1159), the first complete attempt at a philosophy of the state since Classical times. Stoic doctrines of natural law, society, state, and Providence were important elements in his effort to construct a social philosophy on ethical and metaphysical principles. The impact of these doctrines and the lengthy history of their use in the earlier Middle Ages can also be found in the views of Thomas Aquinas on the philosophy of the state and of man.

Renascence of Stoicism in modern times. If the influence of Stoic doctrines during the Middle Ages was largely restricted to the resolution of problems of social and political significance, it remained for the Renaissance, in its passion for the rediscovery of Greek and Roman antiquity, to provide a basis for the rebirth of Stoic views in logic, epistemology, and metaphysics, as well as the documentation of the more familiar Stoic doctrines in ethics and politics. Late in the 16th century, Justus Lipsius, a Flemish scholar and Latin Humanist, was responsible for the first restatement of Stoicism as a defensible and thoroughgoing (Christian) philosophy of man. His treatises *De constantia* (1584) and *Politicorum sive civilis doctrinae libri V* (1589) were widely known in many editions and translations. His defense of Stoic doctrine in *Manuductio ad Stoicam Philosophiam* (1604) and *Physiologia Stoicorum* (1604) provided the basis for the considerable Stoic influence during the Renaissance. Around the turn of the 17th century, Guillaume du Vair, a French lawyer and Christian philosopher, made Stoic moral philosophy popular, while Pierre Charron, a French theologian and Skeptic, utilized Stoic themes in *De la sagesse* (1601; Eng. trans., *Of Wisdome,* 1608), as did the Skeptic Michel de Montaigne in his *Essais* (1580; Eng. trans. 1603). Through the work of Lipsius, Stoic doctrines were to influence the thought of Francis Bacon, a precursor of modern philosophy of science, and, later, the *De l'esprit des lois* (1748; Eng. trans., *The Spirit of Laws,* 1750), by the political theorist Charles-Louis, baron de Montesquieu. In the continuing and relentless war against the Aristotelianism of the later Middle Ages, the doctrines of Stoicism influenced many prominent figures of the Renaissance and Reformation periods.

Pietro Pomponazzi, an Aristotelian of early 16th-century Italy, in defending an anti-Scholastic Aristotelianism against the Averroists, who viewed the world as a strictly necessitarian and fated order, adopted the Stoic view of Providence and human liberty. The 15th-century Humanist Leonardo Bruni absorbed Stoic views on reason, fate, and free will. Pantheism, the view that God and nature are unitary in the sense that God is an impersonal being, and naturalism, the view that nothing is supernatural, both of which identify God with the cosmos and ascribe to it a life process of which the world soul is the principle, were widely held Renaissance notions. Such a pantheistic naturalism was advocated—though from diverse standpoints—by Francesco Patrizi, a versatile Platonist, and by Giordano Bruno, defender of an infinite cosmos; and in both authors the inspiration and source were fundamentally Stoic. In the development of a philosophy of public law based upon a study of man, Stoic elements are found in the *Utopia* (1516; Eng. trans. 1551), by Thomas More, and the *De Jure Belli ac Pacis* (1625; Eng. trans. 1682), by Hugo Grotius. This latter work is one of the most famous Renaissance treatises on the theory of natural and social rights.

The foremost Swiss reformer of the early 16th century, Huldrych Zwingli, who regarded justification by subjective belief as the foundation of the new Christianity, utilized Stoic views on the autonomy of the will, on the absolute predestination of the good and evil man, and on moral determinism.

Another Stoic influence of considerable importance in the tradition of Christian Humanism was the view that all religions have a common basis of truths concerning God—a universal Deism. Among those who favoured such a view were Zwingli and Desiderius Erasmus, the great Renaissance Humanist and scholar. More and Grotius also laid special stress on this view, and its influence was felt in the moral, social, and even the artistic life of the 16th century. Later, Herbert of Cherbury, often called the father of Deism, further developed the idea of religious peace and the reduction of opposing religious views to common elements. This view became one of the most popular ideas of the 17th century.

Philipp Melanchthon also cultivated Humanism and the philosophy of antiquity as a basis for a reborn Christianity. Although Aristotle was his chief inspiration, Melanchthon made telling use of the Stoic theory of knowledge, with its notions of innate principles and the natural light of reason, which teach man the great truths of metaphysical and moral order. Stoicism thus became the basis for the natural-law theory, which holds that the state is of immediately divine origin and independent of the church—a Protestant view opposed by Catholic writers.

Stoic concepts in Lipsius and certain Skeptics

Stoic concepts in Cartesian and Spinozistic thought

The Cartesian revolution in thought in the 17th century brought forward several Stoic notions: that morality consists of obedience to the law of reason, which God has deposited within man; that ethics presupposes a knowledge of nature, because man must learn to know his place in the world, for only then may he act rightly; that self-examination is the foundation of ethics; and that the innateness and commonality of truths bespeak the view that

only thoughts and the will belong properly to man, for the body is a part of the material world. Such views were particularly developed by René Descartes, often hailed as the father of modern philosophy, in his dualism of mind (or soul) and body.

Benedict de Spinoza, a freethinking Jewish Rationalist, made similar use of Stoic views on the nature of man and the world. That aspect of Spinoza's thought that is debatably labelled pantheist is essentially Stoic in character. Together with the Cartesians, Spinoza insisted upon the importance of internal and right reason as the sole means by which to attain to indubitable truths and to the possibility of human freedom.

Blaise Pascal, a French scientist and religious writer, also was sympathetic to Cartesian conceptions of man and nature. Though he turned his back on philosophy, his religious thought retained the Cartesian and Stoic insistence on the independence of human reason, holding that man is fundamentally a thinking being, innately capable of making right decisions. There is an important and crucial difference and conflict between Pascal's views and those of Spinoza and the Cartesians: for Pascal, though the use of (the Stoic) right reason might result in proofs and demonstrations that lead to the God of truth, it would never lead to the God of love, the one true God. Thus, the Stoic exaltation of reason to an entity in its own right—indeed, to a divine entity—as exemplified among the Cartesians and in the thought of Spinoza, was rejected by Pascal in the Jansenist Christianity that he finally adopted—a rejection that, because it also repudiated free will, distinguishes Pascal from those who held Stoic as well as alternative conceptions of human freedom and responsibility.

Stoic concepts in Christianity and modern philosophical thought

Christianity in general, in spite of striking contrasts with Stoicism, has found elements within it that parallel its own position. As the Stoic, for example, feels safe and protected in the rational care of some immanent Providence, so the Christian senses that a transcendent though incarnate and loving God is looking after him. And in general, Stoicism has played a great part throughout the ages in the theological formulation of Christian thought as well as in the actual realization of the Christian ideals.

Contemporary philosophy has borrowed from Stoicism, at least in part, its conviction that man must be conceived as being closely and essentially connected with the whole universe. And contemporary Humanism still contains some obviously Stoic elements, such as its belief in the solidarity of all peoples based upon their common nature, and in the primacy of reason. It is perhaps just because Stoicism has never become a full-fledged philosophic system that, many centuries after the dissolution of the Stoic school, fundamental themes of its philosophy have emerged again and again, and many have become incorporated into modern thinking. (Ja.L.S.)

MODERN SCHOOLS

Analytic and Linguistic philosophy

The methods that have dominated British philosophy for most of the 20th century and American philosophy since somewhat more recently have been called Linguistic and Analytic because language and the analysis of the concepts expressed by language have been a central concern. Though Australia and the Scandinavian countries have also contributed to this movement, it has won a very limited following elsewhere. Although there is a unity of outlook in the tradition, individual philosophers and movements within it have differed, often radically, about the goals and methodology of philosophy. A leading figure prior to the mid-20th century, Ludwig Wittgenstein, an Austrian-born Cambridge philosopher, for example, may have been unique in the history of philosophy in having engaged in two periods of profoundly influential philosophical productivity of which the later work was in large part a renunciation and a sustained argument against the earlier. Yet both the early Wittgenstein (represented by his *Tractatus Logico-Philosophicus,* 1922) and the later (represented by the *Philosophical Investigations,* 1953) are central examples of Analytic philosophy.

Different approaches to the study of language

Moreover, the aims assigned to the philosophical study of language have often been different. Some philosophers, among them Bertrand Russell and the early Wittgenstein, have thought that the underlying structure of language mirrors that of the world—that from an analysis of language a philosopher can grasp important truths about reality. This so-called picture theory of language, though influential, is generally repudiated by current Analytic philosophers. Another important dispute concerns whether everyday language is defective, vague, misleading, and even, at times, contradictory. Some Analytic philosophers have thus proposed the construction of an "ideal" language: precise, free of ambiguity, and clear in structure. The general model for such a language has been symbolic logic, the growth of which in the 20th century has played a central role in Analytic philosophy. An ideal language, it was thought, would resolve many traditional philosophical disputes that have arisen from the misleading structure of natural languages. At the other pole, some philosophers have thought that many philosophic problems have come from paying too little attention to what men say in everyday language about various situations.

Despite such disagreements, Analytic philosophers have much in common. Most of them, for example, have concentrated on particular philosophical problems, such as that of induction, or have examined specific concepts, such as those of memory or of personal identity, without attempting to construct any grand metaphysical schemes—an attitude that has roots as ancient as those of the Socratic method exemplified in Plato's dialogues. Almost invariably Plato began with specific questions such as "What is knowledge?" or "What is justice?" and pursued them in a way that can be viewed, without undue strain, as philosophical analysis in the modern sense.

Ideally, a philosophical analysis illuminates some important concept and helps to answer philosophical questions involving the concept. A famous example of such analysis is contained in Bertrand Russell's theory of definite descriptions. In a simple subject–predicate statement such as "Socrates is wise," he said, there seems to be something referred to (Socrates) and something said about it (that he is wise). If, instead of a proper name, however, a "definite description" is substituted, as in the statement "The president of the United States is wise," there is apparently still something referred to and something said about it. But a problem arises when nothing fits the description, as in the statement "The present king of France is wise." Though there is apparently nothing for the statement to be about, one nevertheless understands what it says. Consequently, a pre-World War I philosopher, Alexius Meinong, celebrated for his *Gegenstandstheorie* ("theory of objects"), felt forced by such examples to distinguish between things that have real existence and things that have some other sort of existence; for such statements could not be understood unless they were about something.

Theory of definite descriptions

In Russell's view, philosophers such as Meinong were misled by surface grammatical form into thinking that such statements are simple subject–predicate statements. In reality they are complex; in fact, an analysis of the foregoing example shows that the definite description, "the present king of France," is not an independent unit in the statement at all. Upon analysis, the statement is a complex conjunction of statements: (1) "There is a present king of France"; (2) "There is at most one present king of France"; and (3) "If anyone is a present king of France, he is wise." But, more importantly, each of the three components is a general statement and is not about anything or anyone in particular. There is no phrase in the complete analysis equivalent to "the present king of France," which shows that the phrase is not an expression, like a proper name, that refers to something as the thing

that the whole statement talks about. There is no need, therefore, to make Meinong's distinction between things that have real existence and things that have some other kind of existence.

GENERAL VIEWPOINT OF ANALYTIC PHILOSOPHY

Nature, role, and method of analysis. Analytic philosophy is concerned with the close and careful examination of concepts.

Status of philosophy in the Empiricist tradition. In spirit and style Analytic philosophy has strong ties with the Empiricist tradition, which stresses the data received through the senses and which, except for brief periods, has characterized British philosophy for some centuries, distinguishing it from the more Rationalistic trends of continental European philosophy. It is not surprising, therefore, that Analytic philosophy should find its home mainly in the Anglo-Saxon countries. In fact, the beginning of modern Analytic philosophy is generally dated from the time when two of its major figures, Bertrand Russell and G.E. Moore, both Cambridge philosophers, rebelled against an anti-Empiricist Idealism that had temporarily captured the English philosophical scene. The most famous British Empiricists—John Locke, George Berkeley, David Hume, and John Stuart Mill—had many interests, doctrines, and methods in common with contemporary Analytic philosophers. Although many of their particular doctrines are favorite targets of attack by Analytic philosophers today, one feels that this is more the result of a common interest in certain problems than any difference in general philosophical outlook.

Most Empiricists, though admitting that the senses fail to yield the certainty requisite for knowledge, hold nonetheless that it is only through observation and experimentation that justified beliefs about the world can be gained; *i.e.*, a priori reasoning from self-evident premises cannot reveal how the world is. This view has resulted in a sharp dichotomy among the sciences: between the physical sciences, which ultimately must verify their theories by observation, and the deductive or a priori sciences—*e.g.*, mathematics and logic—the method of which is the deduction of theorems from given axioms. Thus, the deductive sciences cannot give justified beliefs, much less knowledge, of the world. This consequence was one of the cornerstones of two important movements within Analytic philosophy, logical atomism and Logical Positivism (see below). In the Positivist's view, for example, the theorems of mathematics are merely the result of working out the consequences of the conventions that have been adopted for the use of its symbols.

Physical versus a priori sciences

The question then arises whether philosophy itself is to be assimilated to empirical or to a priori sciences. Early Empiricists assimilated philosophy to the Empirical sciences. They were less self-reflective about its methods than contemporary Analytic philosophers are. Being preoccupied with epistemology (theory of knowledge) and the philosophy of mind, and holding that fundamental facts can be learned about these subjects from individual introspection, they took their work to be a kind of introspective psychology. Analytic philosophers in the 20th century, on the other hand, have been less inclined to appeal ultimately to direct introspection. Moreover, the development of rigorous methods in formal logic seemed to promise help in solving philosophical problems—and logic is as a priori as a science can be. It seemed, then, that philosophy must be classed with mathematics and logic.

Function and methodology of philosophy

Conceptual, linguistic, and scientific analysis. The question remained, however, what philosophy's function and methodology are. For a great many Analytic philosophers who do philosophy in the minute and meticulous manner of G.E. Moore and, in particular, for those who have made Oxford the centre of Analytic philosophy (see below, *Recent trends in England: Oxford philosophers*), its business is the analysis of concepts. For them, philosophy is an a priori discipline because the philosopher in some sense already possesses the concept in which he is interested and needs no observations in order to analyze it.

Philosophy can be seen either as conceptual or as linguistic analysis. In the analysis of the concept of seeing, for example, the philosopher is not expressing purely linguistic concerns—with, say, the English verb "to see"—though an investigation of what can be said using that verb may be relevant to his conclusions. For a concept is independent of any particular languages; a concept is something that all languages, insofar as they are capable of expressing the concept, have in common. Thus, philosophers who stress that it is concepts that they analyze attempt to rebut the charge that their problems and solutions are merely verbal.

In contrast, other Analytic philosophers have been concerned with how expressions are used in a particular, nontechnical, everyday language. Thus, the term ordinary language philosophy has been applied by critics as a term of opprobrium to such philosophers. An influential study, *The Concept of Mind* (1949), by Gilbert Ryle, a prominent Oxford Analyst, is an example of a work that some critics took to depend in large part on a trivial appeal to how English speakers talk; but many of Ryle's arguments could equally well have been given by Analytic philosophers who would look upon the term ordinary language with horror.

The problem of perception illustrates how Analytic philosophers who do conceptual analysis think of the goal of philosophy as both different from and complementary to science. Physiologists, psychologists, and physicists—through experiments, observations, and testable theories—have also contributed to man's understanding of perception. There is in the sciences, however, a strong tendency to advance beyond earlier positions, which seems to be absent from philosophy. In philosophy, for example, the account of perception given by such 20th-century Analytic philosophers as G.E. Moore and the Positivist A.J. Ayer has a close connection with that of Locke in the 17th century.

Philosophy as compared with science

The difference between philosophy and science is that, whereas the scientist investigates an actual occurrence, such as seeing, the philosopher investigates a concept that he already possesses quite independently of what he might discover through the occurrence. Whereas the scientist begins by supposing that he can recognize examples of seeing and is already exercising the concept, the philosopher wants to know what is involved in seeing in the sense of what conditions one can use to classify cases as examples of seeing. He may want to know, for example, whether certain conditions are necessary or sufficient. In testing the philosophical theory that, for an observer to see an object, the object must cause a visual experience in him (the causal theory of perception), one does not set up a scientific experiment. It would be of no use to set up situations in which various physical objects are not causing any visual experiences in order to see whether they still can be seen. For if the theory is correct, no such experimental situation will be an instance of seeing; and if it is wrong, merely describing a hypothetical situation would suffice. The question is one about how situations are classified, and for that purpose hypothetical situations are as good as real ones.

Therapeutic function of analysis. For some philosophers in the Analytic tradition, especially those influenced by Wittgenstein, the analysis of concepts has therapeutic value beyond the intrinsic enjoyment of doing it. Even scientists and laymen in their philosophical moments generate problems by not understanding the proper analyses of the concepts that they employ. They are then tempted to formulate theories to explain these difficulties, when instead they should be sorting out the roles of the concepts, which would show them that there was no problem to begin with. Thus, the failure to see how psychological concepts—sensations, emotions, and desires—are employed has led philosophers to such problems as how one can know what is going on in another's mind or how desires and emotions can produce physical changes in the body, and vice versa. Analysis of the concepts involved would, in this way of looking at philosophy, "dissolve" rather than solve the problems, for philosophers would come to see that their formulations of the problem rest on mistakes about the concepts involved.

Dissolving of problems by analysis

This way of looking at philosophy has often been criticized as making it merely a clearing up of the confusions

of other philosophers and therefore a sterile enterprise. The confusions, however, need not be only those of other philosophers. Scientists, for example, can also generate philosophical theories that affect how they design their experiments, which may, thus, be subjects for philosophical therapeutics. Behaviorism in psychology—which views emotions, desires, and attitudes as being dispositions to behave in certain ways—seems to be a philosophical theory and perhaps to be based on a confusion about the analysis of psychological concepts. Yet Behaviorism has influenced psychologists in their approach to the science. Thus, in this view, philosophy can have a therapeutic value beyond the sphere of philosophical games.

Philosophy, in spite of its abstractness, has traditionally been concerned with human needs, and the therapeutic model may even fulfill this ideal. Laymen, as well as philosophers, for example, are bothered by the thought that their actions are determined not by themselves but by prior conditions. This is a problem that, if the therapeutic view is correct, rests on the misunderstanding of such concepts as causation, responsibility, and action, which need clarification.

Formal versus ordinary language. The role of language as a central concern of Analytic philosophers is the dimension most involved in disputes about the methodology employed. Philosophers outside the Analytic movement tend to think that its preoccupation with language is a departure from philosophy as classically conceived. Yet Plato and Aristotle, medieval philosophers, the Empiricists—and, in fact, most of the philosophers whose works have been considered important—have found it essential to talk about language. There are serious differences, however, about what role language should play. One such difference concerns the importance of formal languages (in the sense employed in symbolic logic) for philosophical problems.

Development of mathematical logic. Since the time of Aristotle, logic has been allied to philosophy. Until the late 19th century, however, logic was largely confined to formulating elaborate rules for one fairly simple form of argument—the syllogism; and there was a lack of systematic development of the subject along lines that had been taken in mathematics since early times.

Axiomatic method and use of symbols

Almost from the beginning, mathematicians had rigorously exploited two important techniques: (1) the use of the axiomatic method (as in Euclid's geometry) in developing the subject; and (2) the use of schematic letters or variables for stating general truths in the subject (thus, one can write "$A + B = B + A$," in which any names or numbers whatsoever can be substituted for A and B, and the result will still be true).

It is surprising that logicians through the ages failed to grasp the power of the use of schematic letters. When they finally began to employ these and other mathematical techniques, they made great contributions to man's understanding of the subject.

Among the developments that occurred in the 19th century, primarily through the work of mathematicians, those of the Englishman George Boole, creator of Boolean algebra, and of Georg Cantor, the Russian-born creator of set theory, are especially important inasmuch as they gave promise of bringing logic and mathematics closer together. The one figure who was both a mathematician and a philosopher and so might be credited with the marriage of logic as a philosophical subject with the techniques of mathematics was Gottlob Frege (died 1925), of the University of Jena in Germany. Historically, Frege, whose works are now appreciated in their own right, was important principally for his influence on Bertrand Russell, whose monumental work, *Principia Mathematica* (1910–13), written in collaboration with Alfred North Whitehead, together with Russell's earlier *Principles of Mathematics* (1903), awakened philosophers to the fact that the use of mathematical techniques in logic might prove to be of great importance for philosophy. Its symbolism had the advantage of being closely connected with ordinary language, whereas its rules can be precisely formulated. Moreover, work in symbolic logic has produced many distinctions and techniques that can be applied to ordinary language.

Divergence of ordinary language from formal logic. Ordinary language, however, seems to differ from the artificial language of symbolic logic in more respects than its lack of precisely stated rules. On the surface, it often appears to violate the rules of symbolic logic. In the English statement "If this is gold [symbolized by p], then this will dissolve in aqua regia [symbolized by q]," for example, which in symbolic logic is expressed in a form known as the material conditional, $p \supset q$ (in which $\supset$ means "If . . . then . . . "), one of the rules is that the statement is true whenever "This is gold" is false. In ordinary language, on the contrary, one would not count the statement as true merely on formal logical grounds but only if there were some real connection in the world of chemical reactions between being gold and dissolving in aqua regia—a connection that plays no role in symbolic logic.

Divergences between logic and language

Among Analytic philosophers the existence of many such apparent divergences between symbolic logic and ordinary language has generated attitudes ranging from complete mistrust of symbolic logic as relevant to nonartificial languages to the position that ordinary language is not a proper vehicle for the rigorous statement of scientific truths.

Interpretations of the relation of logic to language. Symbolic logic has been viewed by many Analytic philosophers as providing the framework for an ideal or perfect language. This statement can be taken in two ways:

1. Russell and the early Wittgenstein thought of logic as revealing, in a precise fashion, the real structure of any language. Any seeming departure from this structure in ordinary language must therefore be attributed to the fact that its surface grammar fails to reveal its real structure and is apt to be misleading. As a corollary, philosophers who have held this view have often explained philosophical problems as arising from being taken in by the surface features of the language. Because of the similarity of sentences such as "Tigers bite" and "Tigers exist," for example, the verb "to exist" may seem to function, as other verbs do, to predicate something of the subject. It may seem, then, that existence is a property of tigers just as their biting is. In symbolic logic, however, the symbolic equivalent of the two sentences would be quite different; existence would not be represented by a symbol for a predicate but by what is called the existential quantifier, $(\exists x)$, which means "There exists at least one x such that . . . ."

2. The other sense in which symbolic logic has been seen as the framework of an ideal language is exemplified in the work of Rudolf Carnap, a 20th-century semanticist, who was concerned with what the best language—especially the best for the purposes of science—is.

One distinctive feature of the formal language of *Principia Mathematica* is that it becomes, when interpreted, a language of true-or-false statements. In ordinary language, on the contrary, one is not restricted to statements of truths; in it one can also issue commands, ask questions, make promises, express beliefs, give permission, and assert necessities and possibilities. Consequently, many philosophers have developed nonstandard logics that incorporate the nonassertoric features of language. Thus, various systems of logic have been formulated and studied (see LOGIC).

Nonassertoric uses of language

On the other side of the coin, many philosophers—most notably the later Wittgenstein and those influenced by him—have thought that attempting to put language into the straitjacket of a formal system is to falsify the way that language works. Language performs a multitude of tasks, and even among expressions that seem to be alike in the way they function—those sentences, for example, that one might think are used simply for expressing facts—examination of their actual use reveals many differences: differences, for instance, in what is counted as showing them to be true or false and in their relationships to other parts of language. Formal systems, according to this view, at best oversimplify and at worst can lead to philosophical problems generated by supposing that all language operates strictly according to a simple set of rules. Accordingly, far from settling philosophical disputes by getting underneath the misleading exterior of ordinary language, formal systems add their own share of confusion.

EARLY HISTORY OF ANALYTIC PHILOSOPHY

Reaction against Idealism. During the last decades of the 19th century, English philosophy was dominated by an absolute Idealism that stemmed from the German philosopher G.W.F. Hegel. For English philosophy this represented a break in an almost solid tradition of Empiricism. The seeds of modern Analytic philosophy were sown when two of the most important figures in its history, Bertrand Russell and G.E. Moore, broke with Idealism at the turn of the 20th century.

Ideal-ism and science

Absolute Idealism was avowedly metaphysical in the sense that its adherents thought of themselves as describing, in a way not open to scientists, certain very fundamental truths about the world. Indeed, what pass for truths in the sciences, were, in their view, not really truths at all; for the scientist must, perforce, treat the world as composed of distinct objects and can only describe and state the relationships supposedly holding among them. But the Idealists held that to talk about reality as if it were a multiplicity of objects is to falsify it; in the end only the whole, the absolute, has reality.

In their conclusions and, most importantly, in their methodology, the Idealists were decidedly not on the side of commonsense intuition. Thus, a Cambridge philosopher, J.M.E. McTaggart, argued that the concept of time is inconsistent and cannot therefore be exemplified in reality. British Empiricism, on the other hand, had always thought of common sense as an ally and science as the model of the way in which to find out about the world. Even when their views might seem out of step with common sense, the Empiricists were generally concerned to reconcile the two.

One can hardly claim that Analytic philosophers have universally thought of themselves as on the side of common sense and much less that metaphysical conclusions (on the ultimate nature of reality) are absent from their writings. But there is in the history of the Analytic movement a strong antimetaphysical strain, and its exponents have generally assumed that the methods of science and of everyday life are the authentic ways of finding out the truth.

Founding fathers: Moore and Russell. The first break from the Idealist view that the physical world is really only a world of appearances occurred when Moore, in a paper, "The Nature of Judgment" (1899), argued for a theory of truth that implies that the physical world has the independent existence that, apart from philosophical theories, it is naively supposed to have. Though the theory was soon abandoned, it did represent a return to common sense.

Compari-son of the two founders of Analytic philosophy

The influences on Russell and Moore—and thus their methods of dealing with problems—soon diverged, and their different approaches became the roots of two broadly different methodologies in the Analytic tradition.

Russell was a major influence on those who approached philosophical problems armed with the technical equipment of formal logic, who saw the physical sciences as the only means of gaining knowledge of the world, and who regarded philosophy—if a science at all—as a deductive and a priori enterprise on a par with mathematics. Russell's contributions to this side of the Analytic tradition have been important and, in great part, lasting.

Moore, on the other hand, never found much need to employ technical tools nor to turn philosophy into a science. His dominant themes were (1) the defense of commonsensical views about the nature of the world against esoteric, skeptical, or grandly metaphysical views and (2) the conviction that the right way to approach philosophical puzzles is to ask exactly what the question is that generated the puzzle before trying to solve it. Philosophical problems, he thought, are often intractable because philosophers have not stopped to formulate precisely what is at issue.

G.E. Moore. Because of these two themes, Moore enlisted much more sympathy among Analytic philosophers from the 1930s and onward who were followers of Wittgenstein's later writings, of Gilbert Ryle's postwar *The Concept of Mind,* and of John Austin's work (see below *Oxford philosophers*). These philosophers, like Moore, saw little hope in advanced formal logic as a means of solving traditional philosophical problems and believed that philosophical skepticism about the existence of an independent external world or of other minds—or, in general, about what men label as common sense—must be wrong. The followers of Wittgenstein also shared with Moore the belief that it is often more important to look at the questions that philosophers pose than at their proposed answers. Thus, unlike Russell, who was important for his solutions in formal logic and ideal models of language, it was more the spirit of Moore's conception of philosophy than its lasting contributions that makes him a seminal influence.

The Idealists were given to arguing for what, in Moore's eyes, were outrageous positions. Thus, in his essay "A Defence of Common Sense" (1925), as in others, his defense was not only against such Idealist doctrines as the unreality of time but also against any of the forms of skepticism—about the existence of other minds or of a material world—that philosophers have espoused. The skeptic, he pointed out, usually has some argument for his conclusion. Instead of examining such arguments, however, Moore pitted against the skeptic's premises quite everyday beliefs, such as, for example, that he had breakfast that morning (thus time cannot be unreal) or that he does in fact have a pencil in his hand (thus there must be a material world). His challenge to the skeptic is to show that the premises of the skeptic's argument are more certain than the everyday beliefs that form Moore's premises.

Moore's attack on Idealism

Although some commentators have seen Moore as an early practitioner of the appeal to "ordinary language," his appeal was really not to what it is proper to say but rather to the beliefs of common sense. His rejection of anything that offends against common sense, however, was influential not only in the release that it afforded from the metaphysical excesses of absolute Idealism but also in its impact on the continuing attitudes of most Analytic philosophers—even though they may have given it a linguistic turn.

Moore was also important for his vision of the proper business of philosophy—analysis. He was puzzled about what is the proper analysis of "X sees Y," in which Y designates a physical object (*e.g.,* a pencil). There must be a special sense of "see," in which one does not see the pencil but only part of its surface. And finally—and most importantly—there is also a sense in which what is directly perceived is not even the surface of the pencil but, rather, what Moore called "sense data" and which earlier Empiricists had called "visual sensations" or "sense impressions." Moore's problem was to discern the relationships among these various elements in perception and, in particular, to discover how a person can be justified, as Moore fully believed he is, in his claims to see physical objects when what he immediately perceives are really sense data. The idea that sense impressions form the immediate objects of perception has played a large role in Analytic philosophy, showing once again its Empiricist roots. It later became an important source of division, however, among the Logical Positivists (see below). Most post-World War II Oxford philosophers, however, together with those closely influenced by Wittgenstein's later work, have found sense data to be as unpalatable and unwarranted as Moore had found McTaggert's doctrine of the unreality of time to be.

Moore's notion of "sense data"

Bertrand Russell. One of the recurring themes in philosophy is the idea that the subject needs to be given a new methodology. Among Empiricists this has often meant making it more scientific. From an early date, Russell enunciated this viewpoint (which was not shared by Moore), finding in the techniques of symbolic logic a measure of reassurance that philosophy might be put on a new basis. Russell did not see the philosopher, however, as merely a logician. Symbolic logic might provide the framework for a perfect language, but the content of that language is something else. The job of the philosopher is—for Russell, as it was for Moore—analysis. But the purpose is somewhat different. In most of Russell's work, analysis has the task of uncovering the necessary assumptions—especially about the kinds of things that exist—for a description of the world as it is. For the most part this description is the one that science gives and is there-

fore realistic. Thus, Russell's use of analysis was openly metaphysical.

The question then arises of how philosophical analysis, which is concerned with how men talk about the world, can presume to give any answers about how the world is. The search for an answer begins with the above-mentioned theory of descriptions—a theory that seems to be closely tied to linguistic concerns. It will be recalled that Russell considered that such definite descriptions as "the author of 'On Denoting' " are not really expressions used to refer to things in the world but that, instead, they make the statements in which they occur into quite general propositions about the world, to the effect that one and only one thing of a certain sort exists and that it has a certain property. Because there must be some way, however, of directly speaking of the things in the world, Russell turned his attention to proper names. The name Aristotle, for example, does not seem to carry any descriptive content. But Russell argues, on the contrary, that ordinary names are really concealed definite descriptions ("Aristotle" may simply mean "The student of Plato who taught Alexander, wrote the *Metaphysics*, etc."). If a name had no descriptive content, one could not sensibly ask about the existence of its bearer, for one could then not understand what is expressed by a statement involving it. If "Bosco" were a name in this sense (without any descriptive content), then merely to understand the statement that Bosco exists or the statement that Bosco does not exist presupposes that one already knows what the name Bosco refers to. But then there cannot be any genuine question about Bosco's existence, for just to understand the question one must know the thing to which the name refers. Ordinary proper names, however—Russell, Homer, Aristotle, and Santa Claus—as Russell pointed out, are such that it makes sense to question the existence of their bearers. Thus, ordinary names must be concealed descriptions and cannot be the means of directly referring to the particular things in the world.

Russell's use of the theory of descriptions

Names in the strict logical sense, then, are very rare; Russell, in fact, suggests that in English the only possible candidates are the demonstrative pronouns, this and that. Yet, if men are ever to talk about the actual things in the world directly, there must be the possibility of such demonstrative expressions underlying their language—in their private thoughts about the world if not in their public language.

To this point, Russell had concluded that things in the world can be talked about only through the medium of a special kind of name; in particular, one about which no question can arise whether it names something or not. At this point there was a transition from questions about the nature of language to results about the nature of the world. Russell asked what sort of thing it is that can be named in the strict logical sense, that can be known and talked about, and that can tell a man something about the world. The important restriction is that no question can arise about whether it exists or not. Ordinary physical objects and other people seem not to fit this requirement.

The question of what can be known and talked about

In his search for something whose existence cannot be questioned, Russell hit upon present experience and, in particular, upon sense data: one can question whether he is really seeing some physical object—whether, for example, there is a desk before him—but a person cannot question that he has had visual impressions or sense data; thus, what a man can name in the strict logical sense and what things he can actually talk about turn out to be the elements of his present experience. Russell therefore made a distinction between what can be known by acquaintance and what can be known only by description; *i.e.*, between those things the existence of which cannot be doubted and those about which, at least theoretically, doubt can be raised. What is novel about Russell's conclusion is that it was arrived at from a fairly technical analysis of language: to be directly acquainted with something is to be in a position to give it a name in the strict logical sense, and to know something only by description is to know only that something uniquely fits the description.

Russell was not constant in his view about physical objects. At one point he thought that the observer must infer their existence as the best hypothesis to explain his experience. Later he argued that they could be taken as logical constructions out of sense data.

Logical atomism: Russell and the early Wittgenstein. The next important development in Analytic philosophy was initiated when Russell published a series of articles entitled "Philosophy of Logical Atomism" (1918–19), in which he acknowledged a debt to Wittgenstein, who had studied with Russell before the war. Wittgenstein's own work, the *Tractatus Logico-Philosophicus* (1922), which can also justly be said to present a logical atomism, turned out to be not only tremendously influential on developments in Analytic philosophy but also such a deep and difficult text that it has generated a growing body of scholarly interpretation.

Russell's choice of the words logical atomism to describe this viewpoint was, in fact, particularly apt. By using the word logical Russell meant to sustain the position, described earlier, that through analysis—particularly with the aid of the ideal structure provided by symbolic logic—the fundamental truths about how any language functions can be revealed and that this disclosure, in turn, would show the fundamental structure of that which the language is used to describe. And by using the word atomism Russell highlighted the particulate nature of the results that his analyses and those of Wittgenstein seemed to yield.

On the linguistic level, the atoms in question are atomic propositions, the simplest statements that it is possible to make about the world; and on the level of what language talks about, the atoms are the simplest atomic facts, those expressible by atomic propositions. More complex propositions, called molecular propositions, can then be built up out of atomic propositions via logical connectives—such as "either . . . or . . . ," "both . . . and . . . ," and "not . . ."—the truth-value of the molecular proposition being in each case a function of the truth values of its component atomic propositions.

Simple atomic and molecular propositions

Language, then, must break down, upon analysis, into ultimate elements that cannot be analyzed into any other component propositions; and, insofar as language mirrors reality, the world must then be composed of facts that are utterly simple. Atomic propositions are composed, however, of strings of names understood, as Russell had explained it, in the strict logical sense; and atomic facts are composed of simple objects, the things that could be thus named.

The details of the Russell–Wittgenstein view have fascinated philosophers by the way in which they not only formed a coherent view but also seemed to follow inexorably from the central assumptions. There are close connections between this period, which was perhaps the most metaphysical in contemporary Analytic philosophy, and traditional Empiricism. The breakdown of language and the world into atomic elements had been one of the prominent features in the classical Empiricists, John Locke, George Berkeley, and David Hume. There was also a view of the connection between language and the world—adumbrated in Russell but fully evident in the *Tractatus*—which has been important and influential, viz., the picture theory, which holds that the structure of language mirrors that of the world. Analysis is important because ordinary language does not show immediately, for example, that it is founded on the atomic–molecular proposition model. Another theme is that the deductive sciences—mathematics and logic—are based solely on the way that language operates and cannot reveal any truths about the world, not even about a world of entities called numbers. Finally, logical atomism, in Wittgenstein's thought as opposed to Russell's, was at one and the same time metaphysical—in the sense of conveying via pure reasoning something about how the world is—and antimetaphysical. Wittgenstein's *Tractatus* is unique in the history of Empiricism in its acceptance of the fact that it is itself a metaphysic and that part of its metaphysics is that metaphysics is impossible: the *Tractatus* says of itself that what it says cannot be coherently said. Only empirical science can tell a man anything about the world as it is. Yet the *Tractatus* apparently tells him, for example, about the relationship between language and the

Logical atomism as an antimetaphysical metaphysic

facts of the world. For Wittgenstein, the solution of this seeming paradox lies in his distinction between what can be said and what can only be shown. There are certain things that can somehow be seen to be so—in particular, the ways in which language is connected with the world. The *Tractatus* could not straightforwardly tell its readers about these matters—metaphysics cannot be a body of facts expressible in any language—but the attempt to say these things, done in the right way, can show them what it cannot coherently express.

Logical Positivism: Carnap and Schlick. Wittgenstein's *Tractatus* was both a landmark in the history of contemporary Analytic philosophy and perhaps its most aberrant example. It not only contained the most highly sophisticated metaphysics but also was an important influence on the most antimetaphysical of the positions taken by Analytic philosophers, viz., that of Logical Positivism, which was mainly developed by a group of philosophers, scientists, and logicians who were centred in Vienna and came to be known as the Vienna Circle. Among these, Rudolf Carnap and Moritz Schlick have perhaps had the most influence on Anglo-American philosophy, although it was an English philosopher, A.J. Ayer—whose *Language, Truth and Logic* (1936) is still the most widely read work of the movement in America and England—who introduced the ideas of Logical Positivism to English philosophy. Its main tenets have struck sympathetic chords in the Analytic philosophers and are still important today, even if in repudiation.

Above all else, Logical Positivism was antimetaphysical; nothing can be learned about the world, it held, except through the methods of the empirical sciences. The Positivists sought a method for showing both (1) when a theory that seemed to be about the world was really metaphysical and (2) that such a theory was, in fact, meaningless, and this they found in the principle of verification. In its positive form, the principle said that the meaning of any statement that is really about the world is given by the methods employed for verifying its truth or falsity—the only allowable methods being, ultimately, those of observation and experiment. In its negative form, the principle said that no statement could both be a statement about the world and have no method of verification attached to it. Its negative form was the weapon used against metaphysics and for the vindication of science as the only possible source of knowledge about the world. The principle would, thus, class as meaningless many philosophical and religious theories that purport to say something about the world but provide no way of testing the truth of the statements; for example, in religion it would render suspect the statement that God exists, which, being metaphysical, would be, strictly speaking, meaningless.

Principle of verification

The principle of verification ran almost immediately into difficulties, most of which were first raised by the Positivists themselves. The attempt to work out these difficulties belongs to a more detailed study of the movement (see below *Positivism and Logical Empiricism*). It is sufficient to note here that these problems were sufficient to make most subsequent Analytic philosophers wary of appealing directly to the principle. It has, however, influenced philosophical work in more subtle ways.

With the principle of verification in hand, the Positivists thought that they could show a great many theories to be nonsense. There were several areas of discourse, however, which failed the test of the principle but which were simply impossible to rule out as concealed nonsense. Foremost among these disciplines were mathematics and ethics. Mathematics (and logic) could hardly be written off as nonsense. Yet their theorems are not verifiable by observation and experiment; they are known, in fact, by pure a priori reasoning alone. The answer seemed to be provided in Wittgenstein's *Tractatus,* which held that the propositions of mathematics and logic are, in Kantian terms, analytic; *i.e.,* true—like the statement "All bachelors are unmarried"—in virtue of the conventions that lie behind the use of the symbols involved.

Emotive theory of ethics

About ethics or, more precisely, about any statements involving value judgments, the Positivist view was different, yet still of lasting importance. In this view, value judgments are not, like mathematical truths, necessary adjuncts to science. But they cannot be put off as nonsense; nor, obviously, are they true by definition or linguistic convention. The usual view of the Positivists, called emotivism, is that what look like statements of fact (*e.g.,* that one should not tell lies) are really expressions of one's feelings toward a certain action; thus, value judgments are not really true or false. The Positivist's position was that neither mathematical nor ethical statements could be dismissed, as were metaphysical propositions. Both had then to be exempted from the principle of verification; and this was done by arguing that their statements are not really about the world: mathematical truths are conventions, and ethical statements are merely expressions of feelings. The divorce of ethics from science, once again, reflects an old Empiricist theme, to be seen, for example, in David Hume's dictum that from matters of fact one cannot derive a conclusion about what ought to be nor vice versa.

LATER HISTORY OF THE MOVEMENT

"Philosophical Investigations": the later Wittgenstein. A crucial turn that initiated developments that were destined to have a lasting and profound effect on much of contemporary Analytic philosophy occurred in 1929, when Wittgenstein, after some years in Austria during which he was not philosophically very active, returned to England and established his residence at Cambridge. There, the direction of his thought soon shifted radically away from his *Tractatus,* and his views became in many ways diametrically opposed to those of logical atomism. Because he published none of the materials of this period, his influence on other English philosophers—and ultimately on those in all of the countries associated with Analytic philosophy—spread by way of his students and those who heard him in the small groups to whom he spoke at Cambridge. His style, too, changed from the semi-rigorous and formally organized propositions of the *Tractatus* to sets of loosely connected paragraphs and remarks in which the ideas are often conveyed more by suggestion and example than discursively. The result has been that one of the major splits within the ranks of Analytic philosophy is that between those who derive their methods from the later Wittgenstein and those who have followed the *Tractatus.*

Although Wittgenstein's thoughts ranged over almost the entire field of philosophy, from the philosophy of mathematics to ethics and aesthetics, their impact has been felt most, perhaps, where it has concerned the nature of language and the relationship between the mental and the physical.

Language and following rules. In logical atomism, as shown above, language was conceived as having a certain necessary and fairly simple underlying structure that it was the job of philosophy to expose. Wittgenstein began to tear away at this assumption.Language, he now thought, is like an instrument that can be used for an indefinite number of purposes. Hence, any effort to codify how it must operate by giving some small set of rules would be like supposing that there is some rigid necessity that a screwdriver (for instance) can be used only to drive screws and forgetting that screwdrivers are also, quite successfully, used to open jars and to jimmy windows. Language is a human institution that is not bound by an outside set of rules—only by what men consider to be correct and incorrect. And that, in turn, is not really a matter for a priori theories to consider.

Language as a multipurpose conventional instrument

The notion of a rule and what it means to follow a rule was especially prominent in his writings. Several concerns made this point of particular interest to Wittgenstein. In mathematics and logic, emphasis was being placed on the rules for manipulating the symbolism. As has been seen, symbolic logic has also been a model for the underlying structure of language. If this fact is coupled with the fact that Russell and the Wittgenstein of the *Tractatus* saw language as reflecting these rules and with the general Empiricist tradition that explains how language operates by each person following internal rules and standards for the use of his words, the picture of the system that Wittgenstein thought mistaken then emerges, and it becomes clear why he placed the notion of a rule so centrally.

Post-factum status of rules

Natural languages, however, are significantly different in that one does not first learn the rules and then use the language; indeed, prior to learning the language, one would not know what to do with rules. Mathematics and logic are, in this sense, bad models for language because they aim at setting out before hand the rules and principles that are subsequently to be used. They encourage the belief that language must have a rigid structure and that, without rules, no language would be possible. The "rules" that one might plausibly discern in the language that one speaks are not, as rules, already there, in a ghostly way, guiding what one says; they are either generalizations from the finite data of what is counted as correct or incorrect, or they are rules that, as Wittgenstein metaphorically expressed it, one puts away in the archives—one adopts the rule but only after the fact.

Following a rule, however, was a concept that Wittgenstein saw as wrongly analyzed in many classical views about language. Thus, he cast irrevocable doubt on the prevalent theory—typified best, perhaps, in John Locke's *Essay Concerning Human Understanding* (1690)—that to use an expression meaningfully is to have in one's mind a standard or a rule for applying it correctly. Against this theme, Wittgenstein's point was that a rule by itself is dead—it is like a ruler in the hands of someone who has never learned to use it, a mere stick of wood. Rules cannot compel nor even guide a person unless he knows how to use them; and the same is true about mental images, which have often been thought to provide the standard for using linguistic expressions. But if rules themselves do not give life to words but require a similar explanation for what gives them life, then there is a useless regress and no (philosophical) explanatory value in the whole apparatus of internal rules and standards.

Relation between mental and physical events. In some respects, Wittgenstein made some significant breaks with the Empiricist tradition—in his views about language and the explanation of the rigour of the deductive sciences. His treatment of the relationship between mental events and physical events also represents an important departure. Empiricists generally have started from the important assumption that what a person is immediately acquainted with is his own sensations, ideas, and volitions, and that these are mental and not physical; and, most importantly, that the things he knows immediately are essentially private and inaccessible to others. For both Moore and Russell there then arose the problem of how, in view of the privacy stressed by the sense-datum theory, the world of physical objects could be known. Wittgenstein's attack on this viewpoint, which has come to be known as "the private language" argument, has become well known, partly because it was in this area that Wittgenstein presented what could most easily be picked out as a more or less formal argument—one that could then be analyzed and criticized in an analytic manner. Even in this case, however, his style of writing was such that his precise formulation of the argument has become a main source of controversy. Wittgenstein argued that the notion of an utterly private experience would imply: (1) that what goes on in the mental life of a person could be talked about only in a language that that person alone whose mental life it was could understand; (2) that such a private language would be no language at all (this has been the main source of controversy); and (3) that the widely held doctrine that there are absolutely private mental events cannot be intelligibly stated, because to do so would be to suppose that one can publically say something about what the doctrine itself says cannot be mentioned in a language accessible to more than one person.

Arguments against private language

The fact that Wittgenstein's argument against private language depends essentially on the question, "What is it to follow a rule?" illustrates a common characteristic of his writings, viz., that themes developed in one area of philosophy continually emerge in apparently quite divorced areas. His extraordinary ability to see a common source of difficulty in philosophical problems that seem to be unrelated helps to explain his style of writing, which seems at first sight to be a somewhat chaotic arrangement of ideas.

Analytic philosophy has also been attracted to a behaviouristic view of mental phenomena that holds that such apparently private events as the feeling of fear are not only not really private but also that they can be identified with publicly observable patterns of behaviour. The disposition toward empirical science, with observation as its foundation, united with the observation that the evidence men have of what goes on in the mental lives of other people must come from what they see of their behaviour, has often warred against the other inclination of Empiricism to regard the starting point of all knowledge of the world, for each person, as being essentially private sense experience. Wittgenstein has had tremendous influence, however, in suggesting that these two extremes are not the only alternatives. Yet attempts to state how Wittgenstein could deny the privacy of experience without espousing some form of behaviourism have not been very successful. Sympathetic interpreters have taken up the notion of "criteria," used, but not developed in any detail, by Wittgenstein. For mental states such as fear, outward behaviour (*e.g.,* running away, blanching, or cringing) does not constitute what it is to be in that state, as behaviourism would have it, but neither is it merely evidence of some completely private event. The problem has been to characterize the relation between behaviour and mental states so that the two are neither identical nor evidence one for the other, while still acknowledging that a knowledge of the person's characteristic behaviour is essential to understanding the notion of a certain mental state.

Problem of behaviourism

Recent trends in England. Those philosophers who might fairly be labelled "Wittgensteinians," who follow the methods that Wittgenstein employed in his later period, should be distinguished from those who have been influenced more indirectly by the general trends and philosophical atmosphere that arose in large part from Wittgenstein's work.

Wittgensteinians. Close students of his ideas have tended to work chiefly on particular concepts that lie at the core of traditional philosophical problems. As an example of such an investigation, a monograph entitled *Intention* (1957), by G.E.M. Anscombe, an editor of Wittgenstein's posthumous works, may be cited as an extended study of what it is for a person to intend to do something and of what the relationship is between his intention and the actions that he performs. This work has occupied a central place in a growing literature about human actions, which in turn has influenced views about the nature of psychology, of the social sciences, and of ethics. And, as an extension of this British influence into the United States, one of Wittgenstein's students, Norman Malcolm of Cornell University, has investigated such concepts as knowledge, certainty, memory, and dreaming. As these topics suggest, Wittgensteinians have tended to concentrate on Wittgenstein's ideas about the nature of mental concepts and to work in the area of philosophical psychology. Typically, they begin with classical philosophical theories and attack them by arguing that they employ some key concept, such as that of knowledge, in a manner incongruous with the way in which the concept would actually be employed in various situations. Their works thus abound with descriptions of hypothetical, though usually homely, situations and with questions of the form, "What would a person say if . . .?" or "Would one call this a case of *X*?" In doing so, they are following out Wittgenstein's advice that, instead of trying to capture the essence of a concept by an abstract analysis, the philosopher should look at how it is employed in a variety of situations.

Concentration on philosophical psychology

Oxford philosophers. After World War II, Oxford University was the centre of extraordinary philosophical activity; and, although Wittgenstein's general outlook on philosophy—his turning away, for example, from the notion of formal methods in philosophical analysis—was an important ingredient, many of the Oxford philosophers could not be called Wittgensteinians in the strict sense. The method employed by many of these philosophers has often been characterized—especially by critics—as an "appeal to ordinary language," and they were thus identified as belonging to the school of "ordinary language" philosophy. Exactly what this form of argument is supposed to be and what exemplifies it in the writings of these

Gilbert Ryle's *The Concept of Mind*

philosophers has been by no means clear. Gilbert Ryle, Moore's successor as editor of a leading journal, *Mind*—and especially in his *The Concept of Mind*—was among the most prominent of those analysts who were regarded as using ordinary language as a philosophical tool. Ryle, like Wittgenstein, pointed out the mistake of regarding the mind as what he called "a ghost in a machine"—to defeat the radical dualism of mind and body that has characterized much of philosophical thinking—by investigating how people employ a variety of concepts, such as memory, perception, and imagination, that designate "mental" properties. He tried to show that, when philosophers carry out such investigations, they find that, roughly speaking, it is the way people act and behave that leads to attributing these properties to them, and that there is no involvement of anything internally private. He also attempted to show how philosophers have come to dualistic conclusions—usually from having a wrong model in terms of which to interpret human activities. A dualistic model may be constructed, for example, by wrongly supposing that an intelligently behaving person must be continually utilizing knowledge of facts—knowledge that something is the case. Ryle contended, on the contrary, that much intelligent behaviour is a matter of knowing how to do something and that, once this fact is acknowledged, there is no temptation to explain the behaviour by looking for a private internal knowledge of facts. Though Ryle's objectives were similar to those of Wittgenstein, his results have often seemed more behavioristic than Wittgenstein's.

It is true that Ryle did ask, in pursuit of his method, some fairly detailed questions about when a person would say, for example, that someone had been imagining something; but it is by no means clear that he was appealing to ordinary language in the sense that his was an investigation into how, say, speakers of English use certain expressions. In any case, the charge, often voiced by critics, that this style of philosophizing trivializes and perverts philosophy from its traditional function would probably also have to be levelled against Aristotle, who frequently appealed to "what we would say."

The work of John Austin

A powerful philosophical figure among postwar Oxford philosophers was John Austin, who was White's professor of moral philosophy until his death, in 1960. Austin felt that many philosophical theories derive their plausibility from overlooking distinctions—often very fine—between different uses of expressions, and he also thought that philosophers too frequently think that any one of a number of expressions will do just as well for their purposes. (Thus, ignoring the difference between an illusion and a delusion, for example, gives credence to the view that what one immediately perceives are not physical objects, but sense data.) Austin's work was, in many respects, much closer to the ideal of philosophy as comprising the analysis of concepts than was that of Ryle or Wittgenstein. He was also much more concerned with the nature of language itself and with general theories of how it functions. This novel approach, as exemplified in *How to Do Things with Words* (1962), set a trend that has been followed out in a growing literature in the philosophy of language. Austin took the total speech act as the starting point of analysis, which allowed him to make distinctions based not only upon words and their place in a language but also upon such points as the speaker's intentions in making the utterance and its expected effect on the audience. There was also in Austin's approach something of the program of Russell and the early Wittgenstein for laying bare the fundamental structure of language.

Recent trends in the United States. Although the Oxford philosophers and the posthumous publication of Wittgenstein's writings have produced a revolution in Anglo-American philosophy, the branch of Analytic philosophy that emphasized formal analyses by means of modern logic has by no means been dormant. Since the appearance of *Principia Mathematica,* striking new findings have emerged in logic, many of which, though requiring for their understanding a high level of mathematical sophistication, are nevertheless important for philosophy.

W.V.O. Quine's use of symbolic logic

Among those philosophers for whom symbolic logic occupies a central position, W.V.O. Quine, Pierce professor of philosophy at Harvard University, has been especially important. Symbolic logic represented for him, as it did for many earlier Analytic philosophers, the framework for the language of science. There were two important themes in his work, however, that represent significant departures from, say, the positions of the logical atomists and the Logical Positivists. In the first place, Quine rejected the distinction between those statements in which their truth or falsity depends upon the meaning of the terms involved and those in which their truth or falsity is a matter of empirical and observable fact—a distinction that had played an essential role in Logical Positivism and was thought by most Empiricists to be the basis for a division between the deductive and the empirical sciences. Quine, in "Two Dogmas of Empiricism" (1951) and subsequent writings, argued that the sort of distinction intended by philosophers is impossible to draw. In the course of his argument, a similar doubt was cast upon concepts traditional not only to philosophy but also to linguistics—in particular, the concept of synonymy or sameness of meaning. Quine's attack has been a threat not only to some long-held doctrines of the Analytic tradition but also to its conception of the nature of philosophy, which has generally depended upon contrasting it with the empirical sciences.

The second important departure of Quine's philosophy has been his attempt to show that science can be successfully conducted without what he calls "intentional entities." In contrast to "extensional," used above as an essential feature of standard symbolic logic, intentional entities include many of the common items that Analytic philosophers often assume that they can talk about without difficulty, such as the meanings of expressions, propositions, or the property of certain statements (such as those of mathematics) of being necessarily true. Quine's program—as exemplified by *Word and Object* (1960)—is intended in part to show that science can say everything that it needs to say without using concepts that cannot be expressed in the extensional language of standard logic. Quine's work, though by no means widely accepted, has made Analytic philosophers at least wary of uncritically accepting certain of their standard distinctions.

Since the mid-20th century, there has been an interaction between the science of linguistics and Analytic philosophy. This did not occur before because Analytic philosophers had almost always considered their study of language to be a priori and unconcerned with empirical facts about particular languages. Recently, however, a book by Noam Chomsky, a U.S. generative grammarian, entitled *Syntactic Structures* (1957), has produced a theory of grammar that not only has profoundly affected the course of linguistics but also bears striking resemblances to philosophical analysis. At first, some Analytic philosophers saw in Chomsky's theory a technique that could be applied to philosophy. It was subsequently considered, however, that, whereas the possibility of looking at grammar in Chomsky's way had contributed valuable concepts for philosophers, the possibility that it would become a methodology for Analytic philosophy had receded. The interchange between linguists and philosophers, however, has continued.

Noam Chomsky's generative grammar

Analytic philosophy today. It is not possible to forecast in any detail the future trends of Analytic philosophy in Anglo-American and Scandinavian countries. It seems relatively certain, however, that the two conceptions of the subject that stem from Moore and Russell will both continue.

Analytic philosophers, mainly influenced by Oxford philosophy, and those for whom symbolic logic is a touchstone analyze many of the same problems and benefit from each other's work. Analysis in the more rigorous sense that Russell's theory of definite descriptions represents is more frequently an aim, despite the doubts of Wittgenstein and many of the Oxford philosophers. The general idea that the only ultimate explanations of the world are the scientific ones and the usual corollary that philosophy is in the service of science—which was a central idea for Russell, for the Logical Positivists, and (in recent times) for Quine—has apparently lost nothing of its vigour. The opposing tendencies, noted above, among Empiricists in

general, and present also in Analytic philosophy, toward behaviourism or Materialism, on the one hand, and toward an Idealism of a phenomenalistic sort (such as that of the Irish bishop George Berkeley), on the other, are not present in the same form—mainly because of the sustained criticisms of Wittgenstein, of his followers, and of the Oxford philosophers. The battleground has shifted to a more subtle level. A substantial number of Analytic philosophers who are styled Materialists or physicalists have proposed a novel technique for reducing mental events and states to physical states. They avoid the well-exposed difficulties of older attempts in which it was held that, when one apparently talks about a separate realm of the mind—speaking of such things as thoughts, emotions, and sensations—the proper analysis of its meaning would be in terms of physical properties and events (usually observable behaviour). The novel idea, on the contrary, is that there is, in fact, an identity between so-called mental events and certain physical events, particularly those occurring in the brain, an identity that it is eventually the task of science to specify—in a way modelled after that in which science discovered that lightning is identical with an electrical discharge. The opposition against this new brand of scientific Materialism does not set up against it a view of the mind as a separate realm coexisting with the physical nor as an essentially private collection of nonphysical events and objects. Rather, the issue has been joined on the question whether the language (or perhaps the concepts) of the psychological and the physical are such as to allow for a scientifically discovered identity between items of the one and items of the other. That there still remains a division among Analytic philosophers concerning the problem of the mental and the physical (though in much altered form) shows both the continuity of the movement and the changes that have occurred. (K.S.D.)

The New Materialism

Empiricism

In its most general sense, Empiricism is the name applied to the attitude that beliefs are to be accepted and acted upon only if they first have been confirmed by actual experience—a definition that accords with the derivation of the name from the Greek word *empeiria*, "experience."

More specifically, Empiricism comprises a pair of closely related, but still distinct, philosophical doctrines—one pertaining to concepts and the other to propositions:

Empiricism of concepts and of beliefs

1. The first of these doctrines, a theory of meaning, holds that words (*e.g.*, the word substance) can be understood or the concepts requisite for any articulate thought possessed only if they are connected by their users with things that they have experienced or could experience (*e.g.*, pieces of wood, or the gases in a gasoline engine).

2. The second doctrine, a philosophical theory of knowledge, views beliefs, or at least some vital classes of beliefs (*e.g.*, that Jane is kind), as depending ultimately and necessarily on experience for justification (Jane is seen performing acts of kindness).

It is not obvious, however, that either of these two doctrines strictly implies the other. Several recognized Empiricists have admitted that there are a priori propositions but have denied that there are a priori concepts. The reverse disconnection between the two forms of Empiricism, however, has no obvious exponents, since there are hardly any philosophers who totally deny a priori propositions and certainly none who would at the same time accept a priori concepts.

Stressing experience, Empiricism is thus opposed to the claims of authority, intuition, imaginative conjecture, and abstract, theoretical, or systematic reasoning as sources of reliable belief. Its most fundamental antithesis is with the latter (*i.e.*, with Rationalism, also called intellectualism or apriorism). A Rationalist theory of meaning asserts that there are concepts not derived from or correlated with experienceable features of the world, such as "cause," "identity," or "perfect circle," and that these concepts are a priori (Latin: "from the former") in the traditional sense of being part of the mind's innate or natural equipment—as opposed to being a posteriori (Latin: "from the latter"), or grounded in the experience of facts. On the other hand, a Rationalist theory of knowledge holds that there are beliefs that are a priori (*i.e.*, that depend for their justification upon thought alone), such as the belief that everything must have a sufficient reason or that a process cannot exist by itself but must occur within some substance. Such beliefs can arise either from intellectual intuition, the direct apprehension of self-evident truth, or from purely deductive reasoning.

VARIOUS MEANINGS OF EMPIRICISM

Broader senses. In both everyday attitudes and philosophical theories, the experiences referred to are principally those arising from stimulation of the sense organs, in particular those of sight and touch. Most philosophical Empiricists, however, have maintained that sensation is not the only provider of experience, admitting as empirical the awareness of mental states in introspection or reflection, such as feelings of pain or of fear, often metaphorically described as present to the "inner sense." It is a controversial question whether still further types of experience, such as moral, aesthetic, or religious experience, ought to be acknowledged as empirical.

Two other viewpoints related to but not the same as Empiricism are the Pragmatism of the American philosopher and psychologist William James, an aspect of which was Radical Empiricism, and Logical Positivism, also called Logical Empiricism (see below *Positivism and Logical Empiricism*). Though these philosophies are, indeed, empirical, each has a distinctive focus that warrants its treatment as a separate movement. Pragmatism stresses the involvement of ideas in practical experience and action, whereas Logical Empiricism is more concerned with scientific experience.

Popular and medical Empiricism

When describing an everyday attitude, the word Empiricism sometimes conveys an unfavourable implication of ignorance of or indifference to relevant theory. Thus, to call a doctor an "Empiric" has been to call him a quack—a usage traceable to a sect of medical men who were opposed to the elaborate medical, and in some views metaphysical, theories of Galen, a prominent Greek physician of the 2nd century AD, theories which dominated medicine until the 17th century. The medical Empiricists opposed to Galen preferred to rely on treatments of observed clinical effectiveness, without inquiring into the mechanisms sought by therapeutic theory. But "Empiricism," detached from this medical association, may also be used, more favourably, to describe a hard-headed refusal to be swayed by anything but the facts that the thinker has observed for himself, a blunt resistance to received opinion or precarious chains of abstract reasoning.

Stricter senses. As a more strictly defined movement, Empiricism reflects certain fundamental distinctions and occurs in varying degrees.

Fundamental distinctions. If the blurring of the distinction between concepts and propositions has confused discussions of Empiricism, another influence at least equally vexing is that which, embodied in the traditional terminology of the debate, contrasts the empirical not with the a priori but with the innate. Since logical problems are easily confused with psychological problems, it is difficult to disentangle the question of the causal origin of man's concepts and beliefs from that of their meaning and justification.

Innate versus a priori

A concept, such as "Five," is said to be innate if a man's possession of it is causally independent of his experience—*e.g.*, of groupings of five objects. A concept, such as "Ought," is a priori, on the other hand, if the logical conditions of its application (or those of the word expressing it) do not include any reference to experienceable states of affairs (thus "Ought" cannot be defined in terms of facts alone). Similarly, a belief is innate if its acceptance is causally independent of the believer's experience; and it is a priori if its justification is logically independent of experience. Propositions could be innate without being a priori: for example, the baby's empirical belief that its mother's breast will nourish it.

Some colour is lent to the confusion of the a priori and the innate by the fact that most empirical concepts are actually acquired through the technique of ostensive defi-

nition, in which a concept (such as "long") is conveyed by introducing instances of it (several long pencils) into the experience of the learner. But it is not in virtue of their mode of acquisition that concepts are empirical; it is the way in which they are applied, once they are possessed, that qualifies them as such. Even if a man were born with an instinctive capacity to use the word blue and never had to learn how to use it, it would still be an empirical concept if the occasion for its use were always his sense experience of a blue object. Furthermore, the fact that men learn, postnatally, to use the word cause does not prove that it expresses an empirical concept if its application to something always implies more than the sense which experience of that kind of thing presents to the mind.

Another supposedly identical but in fact more or less irrelevant property of concepts and beliefs is that of the universality of their possession or acceptance—that a priori or innate concepts must be the common possession of all men and that such beliefs must be accepted by everyone. There may be, in fact, some basis for inferring universality from innateness, since many innate characteristics such as the fear of loud noises appear to be common to the whole species.

Formal versus categorial

Two main kinds of concept have been held to be a priori and thus nonempirical. First, there are certain formal concepts of logic and of mathematics that reflect the basic structure of discourse: "not," "and," "or," "if," "all," "some," "existence," "unity," "number," "successor," "infinity," Secondly, there are categorial concepts, such as "substance," "cause," "mind," and "God," so called after the "categories of thought" as listed by Aristotle and Kant, which the mind imposes upon the given data of experience.

A very large variety of different types of proposition has been held to be a priori. Few would deny this status to such definitional truisms or obvious tautologies as "all hairless heads are bald" or "a rose is a rose." There are also the truths of logic, of mathematics, and of metaphysics—whether transcendent, such as the existence of God or things-in-themselves (lying behind appearances), or immanent and thus discernible within reality, such as the principles (presupposed by much natural science) of conservation, causality, and sufficient reason. Some have held that the basic principles of ethics or the causal laws of nature are a priori. Empiricism maintains, however, that some of these are a priori.

Degrees of Empiricism. Empiricism, whether concerned with meaning or knowledge, can be held with varying degrees of strength. On this basis, absolute, substantive, and partial Empiricisms can be distinguished.

Absolute Empiricists hold that there are no a priori concepts, either formal or categorial, and no a priori propositions. Absolute Empiricism about knowledge is less common than that about concepts, and nearly all philosophers admit that at least obvious tautologies and definitional truisms are a priori; but many would add that these represent a degenerate case.

Substantive Empiricism

A more moderate form of Empiricism is that of the substantive Empiricists, who are unconvinced by attempts that have been made to interpret formal concepts empirically and who therefore concede that formal concepts are a priori but deny that categorial concepts, such as "substance," "cause," and "God," are a priori. In this view, formal concepts would be no longer semantical, pertaining to the relation of words to things; they would be, instead, merely descriptive or purely syntactical, pertaining to the relations between ideas. On this basis "God," for example, would not be an entity alongside other entities but a device for arranging a man's factual beliefs about the world; the concept "God" would thus play a structural and not an informative role.

The parallel point of view about knowledge assumes that the truth of logical and mathematical propositions is determined, as is that of definitional truisms, by the relationships between meanings that are established prior to experience. The truth often espoused by moralists, for example, that one is truly obliged to rescue a man from drowning only if it is possible to do so is a matter of meanings and not of facts about the world. On this view, all propositions that, in contrast to the foregoing example, are in any way substantially informative about the world are empirical. Even if there are a priori propositions, they are all formal or verbal or conceptual in nature, and their necessary truth derives simply from the meanings that man has attached to the words he uses. A priori knowledge is useful because it makes explicit the hidden implications of substantive, factual assertions. But a priori propositions do not themselves express genuinely new knowledge about the world; they are factually empty, "true in all possible worlds." Thus "All bachelors are unmarried" merely gives explicit recognition to the commitment to describe as unmarried anyone who has been described as a bachelor; but it does not add anything new.

Substantive Empiricism about knowledge, which is fundamental to most contemporary Analytical philosophy, regards all a priori propositions as being more-or-less concealed tautologies, or, in Kantian terms, as being analytic. If a person's "duty" is thus defined as that which he should always do, the statement "A person should always do his duty" then becomes "A person should always do what he should always do." Deductive reasoning is conceived accordingly as a way of bringing this concealed tautological status to light. That such extrication is nearly always required means that a priori knowledge is far from trivial.

For the substantive Empiricist, truisms and the propositions of logic and mathematics exhaust the domain of the a priori. Science, on the other hand—from the fundamental assumptions about the structure of the universe to the singular items of evidence used to confirm its theories—is regarded as empirical throughout. The propositions of ethics and those of metaphysics, which deals with the nature of Being as such (for example, "Only that which is not subject to change is real"), are either disguised tautologies or empirical statements or only pseudo-propositions; *i.e.*, combinations of words that, despite their grammatical respectability, cannot be taken as true or false assertions at all.

Partial Empiricism

The least thoroughgoing type of Empiricism here distinguished, ranking third in degree, can be termed partial Empiricism. Many philosophers hold that other concepts besides the formal are a priori and that there are substantially informative propositions about the world that are nevertheless not empirical. The theses of transcendental, or Kantian, metaphysics and those of theology, the general scientific principles of conservation and causality, the basic principles of morality, and the causal laws of nature have all been held to be at once synthetic and a priori—substantial and yet establishable by reasoning alone without recourse to experience. In all versions of this view, however, there remain a great many straightforwardly empirical concepts and propositions: ordinary singular propositions about matters of fact and the concepts that figure in them are held to fall in the empirical domain.

HISTORY OF EMPIRICISM

In ancient philosophy. So-called common sense is inarticulately empiricist; and philosophy, in seeking to correct it, has to start from a rationalistic position. Philosophical Empiricism is thus always critical, a resistance to the pretensions of a more speculative philosophy. The ground was prepared for Plato, the greatest of Rationalist philosophers, by three earlier bodies of thought: the Ionian cosmologies of the 6th century BC—so-called from their concentration along the western coast of Asia Minor—with their distinction between sensible appearance and a reality accessible only to pure reason; the philosophy of Parmenides (early 5th century BC), the important early monist, in which purely rational argument is used to prove that the world is really an unchanging unity; and Pythagoreanism (see above), which, holding that the world is really made of numbers, took mathematics to be the repository of ultimate truth.

The first Empiricists in Western philosophy were the Sophists, who rejected such Rationalist speculation about the world as a whole and took man and society to be the proper objects of philosophical inquiry. Invoking skeptical arguments to undermine the claims of pure reason, they posed a challenge that invited the reaction that comprised

Plato's philosophy (see below, *Criticism and evaluation*).

Plato and to a lesser extent Aristotle were both Rationalists. But Aristotle's successors in the ancient Greek schools of Stoicism and Epicureanism advanced an explicitly Empiricist account of the formation of man's concepts or ideas. For the Stoics the human mind is at birth a clean slate, which comes to be stocked with ideas by the sensory impingement of the material world upon it. Yet they also held that there are some ideas or beliefs, the "common notions," present to the minds of all men; and these soon came to be conceived in a nonempirical way. The Empiricism of the Epicureans, however, was more pronounced and consistent. For them man's concepts are memory images, the mental residues of previous sense experience; and knowledge is as empirical as the ideas of which it is composed.

In medieval philosophy. Most medieval philosophers after St. Augustine (see below, *Criticism and evaluation*) took an Empiricist position, at least about concepts, even if they recognized much substantial but nonempirical knowledge. The standard formulation of this age was: "There is nothing in the intellect that was not previously in the senses." Thus St. Thomas Aquinas (1225–74) altogether rejected innate ideas. Both soul and body participate in perception, and all of man's ideas are abstracted by the intellect from what is given to the senses. Man's ideas of unseen things, like God and angels, are derived by analogy from the seen.

The 13th-century scientist Roger Bacon emphasized empirical knowledge of the natural world and anticipated the polymath Renaissance philosopher of science Francis Bacon (1561–1626) in preferring observation to deductive reasoning as a source of knowledge. The Empiricism of the 14th-century Franciscan Nominalist William of Ockham was more systematic. All knowledge of what exists in nature, he held, comes from the senses, though there is, to be sure, "abstractive knowledge" of necessary truths; but this is hypothetical and does not imply the existence of anything. His more extreme followers extended his line of reasoning toward a radical Empiricism, in which causation is not a rationally intelligible connection but merely an observed regular sequence.

The Empiricism of William of Ockham

In modern philosophy. In the earlier and unsystematically speculative phases of Renaissance philosophy, the claims of Aristotelian logic to yield substantial knowledge were attacked by several 16th-century logicians, and, in the same century, the role of observation was stressed. One mildly skeptical Christian thinker, Pierre Gassendi (1592–1655), advanced a deliberate revival of the empirical doctrines of Epicurus. But the most important defender of Empiricism was Francis Bacon, who, though he did not deny the existence of a priori knowledge, claimed that, in effect, the only knowledge that is worth having (as contributing to the relief of man's estate) is empirically based knowledge of the natural world, which should be pursued by the systematic, indeed almost mechanical, arrangement of the findings of observation and is best undertaken in the cooperative and impersonal style of modern scientific research. Bacon was, indeed, the first to formulate the principles of scientific induction.

A Materialist and Nominalist, Thomas Hobbes (1588–1679), combined an extreme Empiricism about concepts, which he saw as the outcome of material impacts on the bodily senses, with an extreme Rationalism about knowledge, of which, like Plato, he took geometry to be the paradigm. For him all genuine knowledge is a priori, a matter of rigorous deduction from definitions. The senses provide ideas; but all knowledge comes from "reckoning," from deductive calculations carried out on the names that the thinker has assigned to them. True knowledge is thus not merely a priori but also analytic. Yet it all concerns material and sensible existences: everything that exists is a body.

Locke's Empiricism

The most elaborate and influential presentation of Empiricism was made by John Locke (1632–1704), an early Enlightenment philosopher, in the first two books of his *Essay Concerning Human Understanding* (1690). All knowledge, he held, comes from sensation or from reflection, by which he meant the introspective awareness of the workings of man's own mind. Locke confused the two issues of the nature of concepts and the justification of beliefs. His Book I, though titled "Innate Ideas," is largely devoted to refuting innate knowledge. And even so, he later admitted that much substantial knowledge—in particular, that of mathematics and morals—is a priori. He argued that infants know nothing; that if men are said to know innately what they are capable of coming to know, then all knowledge is, trivially, innate; and that no beliefs whatever are universally accepted. Locke was more consistent about the empirical character of all man's concepts and displayed in detail the ways in which simple ideas can be combined to form complex ideas of what has not in fact been experienced. One group of dubiously empirical concepts—those of unity, existence, and number—he took to be derived both from sensation and from reflection. But he allowed one a priori concept—that of substance—which the mind adds, seemingly from its own resources, to its conception of any regularly associated group of perceptible qualities.

Bishop George Berkeley (1685–1753), a theistic Idealist and opponent of Materialism, applied Locke's Empiricism about concepts to refute Locke's account of man's knowledge of the external world. He drew and embraced the inevitable conclusion that material things are simply collections of perceived ideas, a position that ultimately leads to phenomenalism; *i.e.,* to the view that reality is nothing but sensations. He accounted for the continuity and orderliness of the world by supposing that its reality is upheld in the perceptions of an unsleeping God. The theory of spiritual substance involved in Berkeley's position seems to be vulnerable, however, to most of the same objections as those that he posed against Locke.

The Scottish Skeptical philosopher David Hume (1711–76) fully elaborated Locke's Empiricism and used it reductively to argue that there can be no more to man's concepts of body, mind, and causal connection than what occurs in the experiences that he has of them. For Hume all necessary truth is formal or conceptual, determined by the relations of identity and exclusion that hold between ideas.

Voltaire imported Locke's philosophy into France; and its Empiricism, in a very stark form, is the basis of sensationalism, in which all of the constituents of human mental life are analyzed in terms of sensations alone.

A genuinely original and clarifying attempt to resolve the controversy between Empiricists and their opponents was made in the critical philosophy of Immanuel Kant (1724–1804), drawing upon Leibniz and Hume. With the dictum that, although all knowledge begins with experience it does not all arise from experience, he established a clear distinction between the innate and the a priori. He held that there are a priori concepts, or categories—substance and cause being the most important—and also substantial or synthetic a priori truths. Although not derived from experience, the latter apply to experience. A priori concepts and propositions do not relate to a reality that transcends experience; they reflect, instead, the mind's way of organizing the amorphous mass of sense impressions that flow in upon it.

Lockean Empiricism prevailed in 19th-century England until the turn to Hegel occurred in the last quarter of the century. To be sure, the Scottish philosophers who followed Hume but avoided his Skeptical conclusions insisted that man does have substantial a priori knowledge. But the philosophy of John Stuart Mill (1806–73), logician, economist, and Utilitarian moralist, is thoroughly Empiricist. He held that all knowledge worth having, including mathematics, is empirical. The apparent necessity of mathematics, according to Mill, is the result of the unique massiveness of its empirical confirmation. All real knowledge for Mill is inductive and empirical; and deduction is sterile. On similar lines, the philosopher of evolution Herbert Spencer (1820–1903) offered another explanation of the apparent necessity of some of man's beliefs: they are the well-attested empirical beliefs of his ancestors from whom he has inherited them, an evolutionary revival of the doctrine of innateness. Two important mathematicians and pioneers in the philosophy of modern physics,

J.S. Mill and Positivism

W.K. Clifford (1845–79) and Karl Pearson (1857–1936), defended radically Empiricist philosophies of science, anticipating the Logical Empiricism of the 20th century.

In contemporary philosophy. The most influential Empiricist of the 20th century was the great British philosopher and logician Bertrand Russell (1872–1970), who at first was Lockean in his theory of knowledge—admitting both synthetic a priori knowledge and concepts of unobservable entities. Ludwig Wittgenstein (1889–1951), the influential pioneer of the school of Linguistic Analysis, convinced Russell that the truths of logic and mathematics are analytic; and Russell then came to believe, with Hume, that the task of philosophy is to analyze all concepts in terms of what can be directly present to the senses. In this spirit, he tried to show that even the concepts of formal logic are ultimately empirical though the experience that supplies them may be introspective instead of sensory.

Doctrines developed through the collaboration of Russell and Wittgenstein yielded the Logical Positivism of the German philosopher Rudolf Carnap (1891–1970) and of the Vienna Circle, a discussion group in which that philosophy was worked out. The Empiricism of Logical Positivism is especially evident in its restatement of the fundamental thesis of Hume's philosophy in a form known as "the verification principle," which recognizes as meaningful and synthetic only those sentences that are in principle verifiable by reference to sense experience.

CRITICISM AND EVALUATION

The critique of Plato and Aristotle

The earliest expressions of Empiricism in ancient Greek philosophy were those of the Sophists. In reaction to them Plato presented the Rationalistic view that man has only "opinion" about changing, perceptible, existing things in space and time; that "knowledge" can be had only of timeless, necessary truths; and that the objects of knowledge—the unchanging and imperceptible forms or universals (such as "Bed," "Man," etc.)—are the truly real. The circles and triangles of geometrical "knowledge," in this view, are quite different in their perfect exactness from the approximately circular and triangular things present to man's senses. In the *Phaedo,* Plato expounded a theory of literally innate ideas; man, for example, has a conception of exact "equality," which, since it could not have been supplied by the senses, must have been acquired by the soul before it was embodied.

Aristotle agreed with Plato that knowledge is of the universal but held that such universal forms should not be conceived as "separated" from the matter embodying them. This belief does not make Aristotle an Empiricist, although he was certainly a less extreme Rationalist than Plato. Aristotle took the Rationalist view that every science or body of knowledge must resemble Euclidean geometry in consisting of deductions from first principles that are self-evidently and necessarily true and that, although the senses acquaint man with the sensible forms of things, he cannot have knowledge of them unless reason is brought into play to acquaint him with their intelligible forms as well.

The Stoic view of "common notions," or beliefs that are held by all men—a Rationalistic element in an otherwise Empirical school of thought—was expanded during the early medieval period by the Christian Platonist St. Augustine (AD 354–430), a thoroughgoing Rationalist. The Stoic common notions, Augustine held, are truths that God has implanted in the human mind through direct illumination.

Although the early modern expression of Empiricism in the 17th century by Francis Bacon heralded the scientific age, its influence was lessened by his failure to appreciate the revolutionary use of mathematics that comprised the genius of Galileo's new physics and, even more fundamentally, by his underestimation of the need for imaginative conjecture in the formation of scientific hypotheses to restrict the overwhelming number of facts that would otherwise have to be handled. In contrast to Bacon's view, the philosopher and mathematician René Descartes (1596–1650), his contemporary and one of the principal founders of modern thought, developed a comprehensive Rationalism that was more immediately influential. For Descartes all clear and distinct ideas, and in particular those of philosophy and of geometrical physics, are innate; sense experience is at most the agency that elicits ideas already present in the mind. In principle, all knowledge is a priori and demonstrable by pure reasoning, but in practice, because of man's finite intellect, it is necessary to rely on experience to confirm propositions for which rational proof is beyond reach. In England innate ideas and knowledge were defended by Edward Lord Herbert of Cherbury (1582–1648), precursor of Deism, and by a school of Puritan Humanists known as the Cambridge Platonists. The case for innate ideas, however, is hard to establish: there can be in the nature of the case little actual evidence that one can possess concepts before having had some relevant experience.

In the second half of the 17th century, the Empiricist views of John Locke were similarly controverted by a systematic thinker and man of affairs, G.W. Leibniz, who examined Locke's views in minute detail in his book *Nouveaux essais sur l'entendement humain* (1704, published 1765; Eng. trans., *New Essays,* 1916), arguing that ideas can be virtually innate in a less trivial sense than Locke allowed. Interpreting Locke's notion of reflection as reasoning rather than as introspection, Leibniz supposed that Locke was more of a Rationalist than he really was.

In contemporary philosophy, there are thinkers who, though broadly sympathetic to Positivism, have voiced reservations about its more specifically Empiricist elements. One important philosopher of science, Karl Popper, has rejected the inductivism that views the growth of empirical knowledge as the result of a mechanical routine of generalization. To him it is falsifiability by experience that makes a statement empirical. An influential American philosopher and logician, W.V. Quine, has been critical of the Logical Empiricists' frequent recourse to the concept of meaning and has rejected the sharp distinction they make between analytic and factual truths, on which most of contemporary Empiricism rests. For Quine, both human concepts and beliefs are the joint outcome of experience and conventional decision, and he denies that the role of the two factors can be readily distinguished as Empiricists assert.

The theory of knowledge has been the central discipline in philosophy since the 17th century, and its most basic issue is that between Empiricism and Rationalism, an issue that is still being actively debated. On the one hand, the idea that science rests on substantial but nonempirical presuppositions has been put in question by the fact that in some areas it seems to get along without them: without conservation in cosmology, without causality in quantum physics. On the other hand, the traditional theory of the innate powers of the mind has been reanimated by the considerations underlying the theory of language offered by Noam Chomsky, a generative grammarian, who holds that the learning of language is too rapid and too universal to be attributed entirely to an empirical process of conditioning. The basic strength of Empiricism consists in its recognition that human concepts and beliefs apply to a world outside oneself, and that it is by way of the senses that this world acts upon the individual. The question, however, of just how much the mind itself contributes to the task of processing its sensory input is one that no simple argument can answer. (A.M.Q.)

Existentialism

The various philosophies (dating from about 1930) that have been referred to by the term Existentialism have in common an interpretation of human existence in the world that stresses its concreteness and its problematic character.

NATURE OF EXISTENTIALIST THOUGHT AND MANNER

According to Existentialism: (1) Existence is always particular and individual—always *my* existence, *your* existence, *his* existence. (2) Existence is primarily the problem of existence (*i.e.,* of its mode of being); it is, therefore, also the investigation of the meaning of Being. (3) This investigation is continually faced with diverse possibilities, from

among which the existent (*i.e.,* man) must make a selection, to which he must then commit himself. (4) Because these possibilities are constituted by man's relationships with things and with other men, existence is always a being-in-the-world—*i.e.,* in a concrete and historically determinate situation that limits or conditions choice. Man is therefore called *Dasein* ("there being") because he is defined by the fact that he exists, or is in the world and inhabits it.

Contrasts and directions of thought

With respect to the first point, that existence is particular, Existentialism is opposed to any doctrine that views man as the manifestation of an absolute or of an infinite substance. It is thus opposed to most forms of Idealism, such as those that stress Consciousness, Spirit, Reason, Idea, or Oversoul. Secondly, it is opposed to any doctrine that sees in man some given and complete reality that must be resolved into its elements in order to be known or contemplated. It is thus opposed to any form of objectivism or scientism since these stress the crass reality of external fact. Thirdly, Existentialism is opposed to any form of necessitarianism; for existence is constituted by possibilities from among which man may choose and through which he can project himself. And, finally, with respect to the fourth point, Existentialism is opposed to any solipsism (holding that I alone exist) or any epistemological Idealism (holding that the objects of knowledge are mental), because existence, which is the relationship with other beings, always extends beyond itself, toward the being of these entities; it is, so to speak, transcendence.

Starting from these bases, Existentialism can take diverse and contrasting directions. It can insist on the transcendence of Being with respect to existence, and, by holding this transcendence to be the origin or foundation of existence, it can thus assume a theistic form. On the other hand, it can hold that human existence, posing itself as a problem, projects itself with absolute freedom, creating itself by itself, thus assuming to itself the function of God. As such, Existentialism presents itself as a radical atheism. Or it may insist on the finitude of human existence—*i.e.,* on the limits inherent in its possibilities of projection and choice. As such, Existentialism presents itself as a humanism.

Diversity of interests and sources

From 1940 on, with the diffusion of Existentialism through continental Europe, its directions have developed in terms of the diversity of the interests to which they are subject: the religious interest, the metaphysical (or nature of Being) interest, the moral and political interest. This diversity of interests is rooted, at least in part, in the diversity of sources on which Existentialism has drawn. One such source has been the subjectivism of the 4th–5th-century theologian St. Augustine, who exhorted man not to go outside himself in the quest for truth, for it is within him that truth abides. "If you find that you are by nature mutable," he wrote, "transcend yourself." Another source has been the Dionysian Romanticism of Nietzsche, who exalted life in its most irrational and cruel features and made this exaltation the proper task of the "higher man," who exists beyond good and evil. Still another source has been the nihilism of Dostoyevsky, who, in his novels, presented man as continually defeated as a result of his choices and as continually placed by them before the insoluble enigma of himself. As a consequence of the diversity of these sources, Existentialist doctrines have focussed on several aspects of existence.

They have focussed, first, on the problematic character of the human situation, through which man is continually confronted with diverse possibilities or alternatives, among which he may choose and on the basis of which he can project his life.

Second, the doctrines have focussed on the phenomena of this situation and especially on those that are negative or baffling, such as the concern or preoccupation that dominates man because of the dependence of all his possibilities upon his relationships with things and with other men; the dread of death or of the failure of his projects; the "shipwreck" upon insurmountable "limit situations" (death, the struggle and suffering inherent in every form of life, the situation in which everyone daily finds himself); the guilt inherent in the limitation of choices and in the responsibilities that derive from making them; the boredom from the repetition of situations; the absurdity of man's dangling between the infinity of his aspirations and the finitude of his possibilities.

Third, the doctrines have focussed on the intersubjectivity that is inherent in existence and is understood either as a personal relationship between two individuals, I and thou, such that the thou may be another man or God, or as an impersonal relationship between the anonymous mass and the individual self deprived of any authentic communication with others.

Approaches to the general meaning of Being

Fourth, Existentialism focusses on ontology, on some doctrine of the general meaning of Being, which can be approached in any of a number of ways: through the analysis of the temporal structure of existence; through the etymologies of the most common words—on the supposition that in ordinary language Being itself is disclosed, at least partly (and thus is also hidden); through the rational clarification of existence by which it is possible to catch a glimpse, through ciphers or symbols, of the Being of the world, of the soul, and of God; through existential psychoanalysis that makes conscious the fundamental "project" in which existence consists; or, finally, through the analysis of the fundamental modality to which all the aspects of existence conform—*i.e.,* through the analysis of possibility.

There is, in the fifth place, the therapeutic value of existential analysis that permits, on the one hand, the liberating of human existence from the beguilements or debasements to which it is subject in daily life and, on the other, the directing of human existence toward its authenticity; *i.e.,* toward a relationship that is well-grounded on itself, and with other men, with the world, and with God.

The various forms of Existentialism may also be distinguished on the basis of language, which is an indication of the cultural traditions to which they belong and which often explains the differences in terminology among the various authors. The principal representatives of German Existentialism are Martin Heidegger and Karl Jaspers; those of French personalistic Existentialism are Gabriel Marcel and Jean-Paul Sartre; that of French Phenomenology is Maurice Merleau-Ponty; that of Spanish Existentialism is José Ortega y Gasset; that of Russian Idealistic Existentialism is Nikolay Berdyayev (who, however, lived half of his adult life in France); and that of Italian Existentialism is Nicola Abbagnano. The linguistic differences, however, are not decisive for a determination of philosophical affinities. For example, Marcel and Sartre are farther apart than Heidegger and Sartre; and there is greater affinity between Abbagnano and Merleau-Ponty than between Merleau-Ponty and Marcel.

HISTORICAL SURVEY OF EXISTENTIALISM

Many of the theses that Existentialists defend or illustrate in their analyses are drawn from the wider philosophical tradition.

Socrates, Montaigne, and Pascal

Precursors of Existentialism. The problem of what man is in himself can be discerned in the Socratic imperative "know thyself," as well as in the work of Montaigne and Pascal, a religious philosopher and mathematician. Montaigne had said: "If my mind could gain a foothold, I would not write essays, I would make decisions; but it is always in apprenticeship and on trial." And Pascal had insisted on the precarious position of man situated between Being and Nothingness: "We burn with the desire to find solid ground and an ultimate sure foundation whereon to build a tower reaching to the Infinite. But our whole groundwork cracks, and the earth opens to abysses."

The stance of the internal tribunal—of man's withdrawal into his own spiritual interior—which reappears in some Existentialists (in Marcel and Sartre, for example) already belonged, as earlier noted, to St. Augustine. In early 19th-century French philosophy, it was defended by a reformed *Idéologue,* Marie Maine de Biran, who wrote: "Even from infancy I remember that I marvelled at the sense of my existence. I was already led by instinct to look within myself in order to know how it was possible that I could be alive and be myself." From then on, this posture inspired a considerable part of French philosophy.

The theme of the irreducibility of existence to reason, common to many Existentialists, was also defended by a leading German Idealist, F.W.J. von Schelling, as he argued against Hegel in the last phase of his philosophy, and Schelling's polemic, in turn, inspired the thinker usually cited as the father of Existentialism, the religious Dane Søren Kierkegaard.

The requirement to know man in his particularity and, therefore, in terms of a procedure different from those used by science to obtain knowledge of natural objects was confronted by Wilhelm Dilthey, an expounder of historical reason, who viewed "understanding" as the procedure and thus as the proper method of the human sciences. Understanding, according to Dilthey, consists in the reliving and reproducing of the experience of others. Hence it is also a feeling together with others and a sympathetic participation in their emotions. Understanding, therefore, accomplishes a unity between the knowing object and the object known.

The immediate background and founding fathers. The theses of Existentialism found a particular relevance during World War II, when Europe found itself threatened alternately by material and spiritual destruction. Under those circumstances of uncertainty, the optimism of Romantic inspiration, by which the destiny of man is infallibly guaranteed by an infinite force (such as Reason, the Absolute, or Mind) and propelled by it toward an ineluctable progress, appeared to be untenable. Existentialism was moved to insist on the instability and the risk of all human reality, to acknowledge that man is "thrown into the world"—*i.e.,* abandoned to a determinism that could render his initiatives impossible—and to hold that his very freedom is conditioned and hampered by limitations that could at any moment render it empty. The negative aspects of existence, such as pain, frustration, sickness, and death—which 19th-century optimism refused to take seriously because they do not touch the infinite principle that these optimists believed to be manifest in man—become for Existentialism the essential features of human reality.

The thinkers who, by virtue of the negative character of their philosophy, constituted the exception to 19th-century Romanticism thus became the acknowledged masters of the Existentialists. Against Hegelian necessitarianism, Kierkegaard interpreted existence in terms of possibility: dread—which dominates existence through and through—is "the sentiment of the possible." It is the feeling of what can happen to a man even when he has made all of his calculations and taken every precaution. Despair, on the other hand, discovers in possibility its only remedy, for "If man remains without possibilities, it is as if he lacked air." Karl Marx, in holding that man is constituted essentially by the "relationships of work and production" that tie him to things and other men, had insisted on the alienating character that these relationships assume in capitalistic society, where private property transforms man from an end to a means, from a person to the instrument of an impersonal process that subjugates him without regard for his needs and his desires. Nietzsche had viewed the *amor fati* ("love of fate") as the "formula for man's greatness." Freedom consists in desiring what is and what has been and in choosing it and loving it as if nothing better could be desired.

Kierkegaard, Marx, and Nietzsche

Emergence as a movement. Contemporary Existentialism reproduces these ideas and combines them in more or less coherent ways. Human existence is, for all the forms of Existentialism, the projection of the future on the basis of the possibilities that constitute it. For some Existentialists (the Germans Heidegger and Jaspers, for example), the existential possibilities, inasmuch as they are rooted in the past, merely lead every project for the future back to the past, so that only what has already been chosen can be chosen (Nietzsche's *amor fati*). For others (such as Sartre), the possibilities that are offered to existential choice are infinite and equivalent, such that the choice between them is indifferent; and for still others (Abbagnano and Merleau-Ponty), the existential possibilities are limited by the situation, but they neither determine the choice nor render it indifferent. The issue is one of individuating, in every concrete situation and by means of a specific inquiry, the real possibilities offered to man. For all the Existentialists, however, the choice among possibilities—*i.e.,* the projection of existence—implies risks, renunciation, and limitation. Among the risks, the most serious is man's descent into inauthenticity or into alienation, his degradation from a person into a thing. Against this risk, for the theological forms of Existentialism (as in Gabriel Marcel, a Socratic dramatist; Karl Barth, a Swiss Neo-orthodoxist; Rudolf Bultmann, a biblical interpreter), there is the guarantee of the transcendent help from God, which in its turn is guaranteed by faith.

Existentialism, consequently, by insisting on the individuality and nonrepeatability of existence (following Kierkegaard and Nietzsche), is sometimes led to regard one's coexistence with other people (held to be, however, an ineluctable fact of the human situation) as a condemnation or alienation of man. Marcel has said that all that exists in society beyond the individual is "expressible by a minus sign," and Sartre has affirmed in his major work *L'Être et le néant* (1943; *Being and Nothingness,* 1956) that "the Other is the hidden death of my possibilities." For the other forms of Existentialism, however, a coexistence that is not anonymous (as that of a mob) but is grounded on personal communication conditions man's authentic existence.

Broader cultural ramifications

Existentialism has had ramifications in various areas of contemporary culture. In literature, Franz Kafka, author of haunting novels, walking in Kierkegaard's footsteps, described human existence as the quest for a stable, secure, and radiant reality that continually eludes it (*Das Schloss* [1926; *The Castle,* 1930]); or he described it as threatened by a guilty verdict about which it knows neither the reason nor the circumstances but against which it can do nothing—a verdict that ends with death (*Der Prozess* [1925; *The Trial,* 1937]).

The theses of contemporary Existentialism were then diffused and popularized by the novels and plays of Sartre, by the writings of the French novelists and dramatists Simone de Beauvoir and Albert Camus. In *L'Homme révolté* (1951; *The Rebel,* 1953), Camus described the "metaphysical rebellion" as "the movement by which a man protests against his condition and against the whole of creation." In art, the analogues of Existentialism may be considered to be Surrealism, Expressionism, and in general those schools that view the work of art not as the reflection of a reality external to man but as the free immediate expression of human reality.

Existentialism made its entrance into psychopathology through Karl Jaspers' *Allgemeine Psychopathologie* (1913; *General Psychopathology,* 1965), which was inspired by the need to understand the world in which the mental patient lives, by means of a sympathetic participation in his experience. Later, Ludwig Binswanger, a Swiss psychiatrist of the *Daseinsanalyse* school, in one of his celebrated works, *Über Ideenflucht* (1933; "On the Flight of Ideas"), inspired by Heidegger's thought, viewed the origin of mental illness as a failure in the existential possibilities that constitute human existence (*Dasein*). From Jaspers and Binswanger, the Existentialist current became diffused and variously stated in contemporary psychiatry.

In theology, Barth's *Römerbrief* (1919; *The Epistle to the Romans,* 1933) started the "Kierkegaard revival," the emblem of which was expressed by Barth himself; it is "the relation of this God with this man; the relation of this man with this God—this is the only theme of the Bible and of philosophy." Within the bounds of this current, on the one hand, there was an insistence upon the absolute transcendence of God with respect to man, who could place himself in relationship with God only by denying himself and by abandoning himself to a gratuitously granted faith. On the other hand, there was the requirement to demythologize the religious content of faith, particularly of the Christian faith, in order to allow the message of the eschatological event (of salvation) to emerge from among the existential possibilities of man.

METHODOLOGICAL ISSUES IN EXISTENTIALISM

The methods that the Existentialists employ in their interpretations have a presupposition in common: the imme-

diacy of the relationship between the interpreter and the interpreted, between the interrogator and the interrogated, between the problem of being and Being itself. The two terms coincide in existence; for the man who poses the question "What is Being?" cannot but pose it to himself and cannot respond without starting from his own being.

Phenomenology, clarification, psychoanalysis

This common ground notwithstanding, each Existentialist thinker has defended and worked out his own method for the interpretation of existence. Heidegger, an Existentialist with ontological (nature of Being) concerns, availed himself of the philosophy of Edmund Husserl, founder of Phenomenology, which, as *logos* of the *phainomenon,* employs speech that manifests or discloses what it is that one is speaking about and that is true—in the etymological use of the Greek word *alētheia* (*i.e.,* the sense of uncovering or manifesting what was hidden). The phenomenon is, from Heidegger's point of view, not mere appearance, but the manifestation or disclosure of Being in itself. Phenomenology is thus capable of disclosing the structure of Being and hence is an ontology of which the point of departure is the being of the one who poses the question about Being, namely man.

Jaspers, an authority in psychopathology as well as in the philosophy of human existence, on the other hand, employed the method of the rational clarification of existence; he maintained that existence, as the quest for Being, is man's effort of rational self-understanding, or universalizing, of communicating—a method that presupposes that existence and reason are the two poles of man's being. Reason is possible existence; *i.e.,* existence that, as Jaspers writes in his *Vernunft und Existenz* (1935; *Reason and Existence,* 1955), becomes "manifest to itself and as such real, if, *with, through* and *by* another existence, it arrives at itself." This activity, however, is never consummated; thus, when the impossibility of its achievement is recognized, it is changed into faith, into the recognition of transcendence as providing the only possibility of its final achievement.

According to the views of Sartre, the foremost philosopher of mid-20th-century France, the method of philosophy is existential psychoanalysis; *i.e.,* the analysis of the "fundamental project" in which man's existence consists. In contrast to the precepts of Freudian psychoanalysis, which stop short at the irreducibility of the libido, or primitive psychic drive, existential psychoanalysis tries to determine the "original choice" through which man constructs his world and decides in a preliminary way upon particular choices (which, however, may place in crisis the primordial choice itself).

According to Marcel, a Christian Existentialist philosopher and dramatist, the method of philosophy depends upon a recognition of the mystery of Being (*Le Mystère de l'être* [1951; *The Mystery of Being,* 1950–51]); *i.e.,* on the impossibility of discovering Being through objective or rational analyses or demonstrations. Philosophy should lead man up, however, to the point of making possible for him "the productive illumination of Revelation."

Mystery of Being and the structures of existence

Finally, according to humanistic Existentialism, as represented by Abbagnano, the leading Italian Existentialist, and by Merleau-Ponty, a French Phenomenologist, the method of philosophy consists of the analysis and the determination—by employing all available techniques including those of science—of the structures that constitute existence; *i.e.,* of the relations that connect man with other beings and that figure, therefore, not only in the constitution of man but in the constitution of the other beings as well.

SUBSTANTIVE ISSUES IN EXISTENTIALISM

Fundamental concepts and contrasts. Both the ontology and manner of human existence are of concern to Existentialism.

Ontic structure of human existence. The fundamental characteristic of Existentialist ontology is the primacy that that study of the nature of existence gives to the concept of possibility. This priority dominated the philosophy of Kierkegaard and also was amply utilized by Husserl, who had explicitly affirmed the ontological priority of possibility over reality. Possibility, however, is not understood by the Existentialists in the purely logical sense as absence of contradiction nor in the sense of traditional metaphysics as potentiality destined to become actuality but, rather, in the sense of ontic or objective possibility, which is the very structure of human existence; it is thus the specific modality of man's being.

Another way of expressing this thesis is the affirmation of Heidegger and Sartre that "existence precedes essence," which signifies that man does not have a nature that determines his modes of being and acting but that, rather, these modes are simply possibilities from which he may choose and on the basis of which he can project himself. In this sense, Heidegger has said that "*Dasein* is always its own possibility," and Sartre has written: "It is true that the possible is—so to speak—an option on being, and if it is true that the possible can come into the world only through a being which is its own possibility, this implies for human reality the necessity of being its being in the form of an option on its being."

As possibility, human existence is the anticipation, the expectation, the projection of the future. The future is its fundamental temporal dimension, to which the present and the past are subordinate and secondary; existence is always stretched out toward the future. As possibility, existence is also transcendence, being beyond, because all of its constitutive possibilities organize it beyond itself toward the other beings of the world and toward the world in its totality. To transcend thus means to move toward something that is not one's own existence; *i.e.,* toward things and toward other men, with which man is related in every situation in which he finds himself.

Yet for some Existentialists, the being of these other entities has a modality that differs from the being of man's existence: their existence is not possible being but real or factual being. To existence, Heidegger contrasts the presence of the things in the world—a presence that assumes, as man takes notice of these things for his needs, the aspect of utilizability. But utilizability is not a simple quality of things; it is their very being. Analogously, Sartre distinguishes the for-itself—the mode of being of man's existence that he identifies, following Descartes and Husserl, with consciousness—from the in-itself, the being or reality of things that he identifies with their utilizability. According to Jaspers, over against the existence of the possible (man, *Dasein*) stands the world as the infinite horizon that encompasses within itself each possible existence and, therefore, cannot itself be encompassed by any one of them. This is a world that is a reality of fact, at the origin of which there is a Being that is pure transcendence and that, therefore, never reveals itself.

Similarly, the religious forms of Existentialism insist on transcendence, considering it to be the property of the Being that is beyond the existential possibilities and that can enter among them solely under the form of mystery (Marcel) and of the extratemporal revelation of faith (Barth, Jaspers). Marcel, in this regard, has contrasted Being, which is a mystery, with having, which is the condition of man in the world; that is to say, man has objects before him that are foreign to his subjectivity. He tries to organize them and discover the bond that ties them together so as to control and use them.

In all of these doctrines, there is the dominating theme of the contrast between the modality proper to existence, which is possibility, and the modality proper to Being, which is reality or facticity. As a result of this contrast, existence (as possibility) appears as the nothingness of Being, as the negation of every reality of fact. In a brief but famous essay, *Was ist Metaphysik?* (1929), Heidegger affirmed that "Human existence cannot have a relationship with being unless it remains in the midst of nothingness." Rudolf Carnap, a semanticist and leading Logical Positivist, in an equally famous essay, "Überwindung der Metaphysik durch die logische Analyse der Sprache" (1931; "The Elimination of Metaphysics Through Logical Analysis of Language"), criticized this hypostatization (or making real) of Nothingness as one of the grosser fallacies of metaphysics. In truth, Nothingness is, for the Existentialists, possible existence, as the negation of the reality of fact. Sartre has written: "The possible is the *something*

Existence as the nothingness of Being

which the For-itself lacks *in order* to be itself"; it is what the subject lacks in order to be an object; thus it does not exist except as a lacking.

This is also true of value, which is such insofar as it does not exist. For even when value occurs or is perceived in certain acts, it lies beyond them and constitutes the limit or the goal toward which they aim. Analogously, knowledge, in which the object (the in-itself) presents itself to consciousness (the for-itself), is a relationship of nullification, because the object cannot be offered to consciousness except as that which is not consciousness. Furthermore, another existence is such insofar as it is not mine; thus this negation is "the constitutive structure of the being-of-others."

But this reduction of existence to Nothingness can lead in two directions: it can lead to insisting on the lack of meaning—*i.e.,* on the absurdity of existence and of every possible project—as it does in Sartre, in Camus, and in atheistic Existentialism; or it can lead toward the quest for a more direct relationship of existence with Being, beyond the constitutive possibilities of existence, so that Being reveals itself, at least partly, in existence—through language or through faith or through some mystical form of religiousness, as happens in the later phase of Heidegger's thought, in Jaspers, and in all of the forms of theological Existentialism.

Manner and style of human existence. Existentialism is never a solipsism in the proper sense of the term (that I alone exist), because every existential possibility relates man to things and to other men. Sometimes it is presented as humanism in the sense that it places human destiny in the hands of men themselves. But this version is rejected by all of the currents of the movement that, starting with Heidegger, insist on the priority and the initiative of Being with regard to human existence. The opposition between these two points of view depends on how the different Existentialists solve the problem of freedom.

Destiny versus freedom of choice

Man always finds himself in a situation in which his constitutive possibilities are rooted. For Heidegger and Jaspers, this situation determines the choice that he makes among these possibilities; for Sartre, conversely, the situation is determined by the choice. Existentialism fluctuates in this way between the concept of a destiny in which, like Nietzsche's *amor fati,* man accepts what has already been chosen and the concept of a radical freedom whereby the choices are offered to man in an absolute indifference. From the first point of view, every project of life falls back on or is reduced to the situation from which it starts; thus the possibility of being, of acting, of willing, of choosing is really, as Jaspers points out in his *Philosophie* (1932), the impossibility of being, acting, willing, and choosing in a manner different from the way things are; *i.e.,* from the factual conditions of the situation. From the second point of view, the fundamental project, which is the primordial choice, has no conditions; as Sartre says: "Since I am free, I project my total possible, but I thereby posit that I am free and that I can always nihilate this first project and make it past." From the first, or deterministic, point of view, the past determines the future and assimilates it to itself; from the second, or libertarian, point of view, the meaning of the past depends upon the present project. In the latter instance, freedom is a kind of damnation: as Sartre affirms: "We said that freedom is not free not to be free and that it is not free not to exist."

A choice, however, is offered to man even from the destinarian point of view: that between understanding and not understanding one's own nothingness. According to Heidegger, a man achieves what he calls "authentic existence" when he understands the impossibility of all of the possibilities of existence—the impossibility of which the sign or term is death. Jaspers affirms, in his turn, that the only choice offered to man is that between accepting or rejecting the situation with which he is identified. The rejection of it, however, is a betrayal that plunges him back into the situation itself.

Existentialist ontology thus fluctuates between Being and Nothingness and concludes by regarding Nothingness as the only possible revelation of Being. In the atheistic version, it is man, as Sartre affirms, who "strives to be God" and consumes himself vainly in the effort. In the cosmological or theological version, it is Being that intervenes, in a way that is more or less mysterious or hidden, to redeem man from Nothingness.

Problems of Existentialist philosophy. The key problems for Existentialism are those of man himself, of his situation in the world, and of his more ultimate significance.

Man and human relationships. Existentialist anthropology is strictly connected with its ontology. The traditional distinction between soul and body is completely eliminated; thus the body is a lived-through experience that is an integral part of man's existence in its relationship with the world. According to Sartre, "In each project of the For-itself, in each perception the body is there; it is the immediate Past in so far as it still touches on the Present which flees it." As such, however, the body is not reduced to a datum of consciousness, to subjective representation. Consciousness, according to Sartre, is constant openness toward the world, a transcendent relationship with other beings and thereby with the in-itself. Consciousness is existence itself, or, as Jaspers says, it is "the manifestation of being." In order to avoid any subjectivistic equivocation, Heidegger went so far as to renounce the use of the term consciousness, preferring the term *Dasein,* which is more appropriate for designating human reality in its totality. For the same reasons, the traditional opposition between subject and object, or between the self and the nonself, loses all sense. *Dasein* is always particular and individual. It is always a self; but it is also always a project of the world that includes the self, determining or conditioning its modes of being.

Self, body, and relationship with the world

All of these modes of being thus arise, as Heidegger shows in his masterpiece *Sein und Zeit* (1927; *Being and Time,* 1962), from the relationship between the self and the world. Heidegger has regarded concern (in the Latin sense of the term) to be the fundamental aspect of this relationship, insofar as it is man's concern to obtain the things that are necessary for him and even to transform them with his work as well as to exchange them so as to make them more suitable to his needs. Concern demonstrates that man is "thrown into the world," into the midst of other beings, so that in order to project himself he must exist among them and utilize them. Being thrown means, for man, being abandoned to the whirling flow of things in the world and to their determinism.

This happens inevitably, according to Heidegger, in inauthentic existence—day-to-day and anonymous existence in which all behaviour is reduced to the same level, made "official," conventional, and insignificant. Chatter, idle curiosity, and equivocation are the characteristics of this existence, in which "One says this" and "One does that" reign undisputed. Anonymous existence amounts to a simple "being together" with others, not a true coexistence, which is obtained only through the acceptance of a common destiny (see below).

Difficulty of intersubjective communication

All of the Existentialists are in agreement on the difficulty of communication; *i.e.,* of well-grounded intersubjective relationships. Jaspers has perhaps been the one to insist most on the relationship between truth and communication. Truths are and can be different from existence. But if fanaticism and dogmatism (which absolutize a historical truth) are avoided on the one hand while relativism and skepticism (which affirm the equivalence of all truths) are avoided on the other, then the only other way is a constant confrontation between the different truths through an always more extended and deepened intersubjective communication.

Sartre, however, denies that there is authentic communication. According to him, consciousness is not only the nullification of things but also the nullification of the other person as other. To look at another person is to make of him a thing. This is the profound meaning of the myth of Medusa. Sexuality itself, which Sartre holds to be an essential aspect of existence, fluctuates between sadism and masochism, in which either the other person or oneself is merely a thing. On this basis, the intersubjective relationship is obviously impossible.

The human situation in the world. Heidegger has pointed to the foundation of the intersubjective relation-

ship in dread. When a man decides to escape from the banality of anonymous existence—which hides the nothingness of existence, or the nonreality of its possibilities, behind the mask of daily concerns—his understanding of this nothingness leads him to choose the only unconditioned and insurmountable possibility that belongs to him: death. The possibility of death, unlike the possibilities that relate him to other things and to other men, isolates him. It is a certain possibility, not through its apodictic evidence but because it continuously weighs upon existence. To understand this possibility means to decide for it, to acknowledge "the possibility of the impossibility of any existence at all" and to live for death. The emotive tonality that accompanies this understanding is dread, through which man feels himself to be "*face to face* with the 'nothing' of the possible impossibility of [his] existence."

But neither the understanding of death nor its emotive accompaniment opens up a specific task for man, a way to transform his own situation in the world. They enable him only to perceive the common destiny to which all men are subject; and they offer to him, therefore, the possibility of remaining faithful to this destiny and of freely accepting the necessity that all men share in common. In this fidelity consists the historicity of existence, which is the repetition of tradition, the return to the possibilities from which existence had earlier been constituted, the wanting for the future what has been in the past. And in this historicity participate not only man but all of the things of the world, in their utilizability and instrumentality, and even the totality of Nature as the locus of history.

Dread, therefore, is not fear in the face of a specific danger. It is rather the emotive understanding of the nullity of the possible, or, as Jaspers says, of the possibility of Nothingness. It has, therefore, a therapeutic function in that it leads human existence to its authenticity. From the fall into factuality into which every project plunges him, man can save himself only by projecting not to project; *i.e.*, either by abandoning himself decisively to the situation in which he finds himself or by being indifferent to any possible project—with regard to which Sartre says, "Thus it amounts to the same thing whether one gets drunk alone or is a leader of nations."

The pivotal point of that conclusion—the conclusion most widely held among the Existentialists and the one in fact often identified with Existentialism—is the antithesis between possibility and reality. On the one hand, existence is interpreted in terms of possibilities that are not purely logical possibilities or manifestations of a man's ignorance of what exists but are, rather, effective, or ontic, possibilities that constitute man as such; on the other hand, contrasted to possibilities in this sense is a reality, a for-itself, a world, a transcendence that is a factual presence, insurmountable and oppressive, with respect to which possibility is a pure Nothingness. The contradiction to which this antithesis leads becomes clear when the same reality is interpreted in terms of possibility: when the being of things, for example, is reduced to their possibility of being utilized; when the being of other men is reduced to the possibility of anonymous or personal relationships that the individual can have with them; and when the being of transcendence, or of God, is reduced to the possibility of the relationship, although ineffable and mysterious, between transcendence, or God, and man.

The antithesis between possibility and reality

It has been said that a coherent Existentialism should avoid the constant mortal leap between Being and Nothingness; should not confuse the problematic character of existence with the fall into factuality; should not confuse the finitude of possibilities with resignation to the situation, choice with determinism; freedom conditioned by the limits of the situation with the acknowledgment of the omnipresent necessity of the Whole. In this inquiry, it is held, Existentialism could well benefit from a more attentive consideration of science, which it has viewed until now only as a preparatory, imperfect, and objectifying knowledge in comparison with the authentic understanding of Being, which it considers to be a more fundamental mode of the being of man in the world. Science, it is submitted, offers today the example of an extensive and coherent use of the concept of the possible in the key notions that it employs, especially in those branches that are interdisciplinary—among them such notions as indeterminacy, chance, probability, field, model, project, structure, and conditionality.

Some steps in this direction have been taken by Abbagnano and by Merleau-Ponty. According to the latter, considerations of probability are rooted in the being of man, inasmuch as he is situated in the world and invested with the ambiguity of his events. Merleau-Ponty has written in his *Phénoménologie de la perception* (1945):

> Our freedom does not destroy our situation, but is engaged with it. The situation in which we live is open. This implies both that it appeals to modes of privileged resolution and that it is of itself powerless to obtain one of them.

From this point of view, there is always a certain freedom in situations, although its degree varies from situation to situation.

Significance of Being and transcendence. Among the thinkers most frequently mentioned here, the concept of the necessity of Being prevails as the basis of their metaphysical or theological orientations. Heidegger has come more and more to insist on the massive presence of Being in the face of human existence, by attributing to Being all initiative and to man only the possibility of abandoning himself to Being and to the things that are the modes of the language of Being. For Heidegger, Being is interpreted better through the etymology of those words that designate the most common things of daily life than through the analysis of existential possibilities. Jaspers has seen the revelation of transcendence in ciphers—*i.e.*, in persons, doctrines, or poems—all of which can be interpreted as symbols of existential situations and above all of limit situations, the insurmountability of which, in provoking the total "shipwreck" of human possibilities, makes man feel the presence of absolute transcendence. In a less philosophically elaborate form, Being has been understood as mystery by Marcel; as the perfect actuality that guarantees the existential possibilities by Louis Lavelle, a leader of the French *philosophie de l'esprit;* and as the absolute value that man encounters in his own spiritual intimacy by René Le Senne, also of the *philosophie de l'esprit.*

Problems of Existentialist theology. Existentialism has a theological dimension. Though Heidegger rejects the label of atheist, he also denies to the Being of which he speaks the essential qualifications of divinity, inasmuch as it is not the ultimate cause and the Good. But Jaspers, in his last writings, emphasized more and more the religious character of faith in transcendence. Faith is the way to withdraw from the world and to resume contact with the Being that is beyond the world. Faith is life itself, in that it returns to the encompassing Whole and allows itself to be guided and fulfilled by it. Jaspers has even developed a theology of history. He speaks of an axial age, which he places between the 8th and 2nd centuries before Christ, the age in which the great religions and the great philosophers of the Orient arose—Confucius and Lao-tzu, the *Upaniṣads,* Buddha, Zoroaster, the great prophets of Israel—and in Greece the age of Homer and of classical philosophy as well as Thucydides and Archimedes. In this age, for the first time, man became aware of Being in general, of himself, and of his limits. The age in which man now lives, that of science and technology, is perhaps the beginning of a new axial age that is the authentic destiny of man but a destiny that is far off and unimaginable.

For Bultmann, the theologian of the demythologization of Christianity, inauthentic existence is tied to the past, to fact, to the world, while authentic existence is open to the future, to the nonfact, to the nonworld; *i.e.*, to the end of the world and to God. Thus, authentic existence is not the self-projection of man in the world but, rather, the self-projection of man in the love of and obedience to God. But this self-projection is no longer the work of human freedom; it is the saving event that enters miraculously through faith into the future possibilities of man.

In these theological speculations and in others that are comparable, the common presupposition of the Existentialists is recognized—*i.e.*, the gap between human existence and Being. There is either an acknowledgment of that gap, with existence assuming the role of the demonic

(the alternative that Sartre and others have all illustrated above all in their literary works), or an acknowledgment of the hidden participation of human existence in Being through a gratuitous initiative on the part of Being.

Aesthetic, ethical, and religious levels

Kierkegaard had earlier distinguished three stages of existence between which there is neither development nor continuity but gaps and jumps: the aesthetic stage is the one in which one lives for the pleasure of the moment; the ethical stage is the one based on the stability and continuity of life in work and in matrimony; and the religious stage is the one characterized by faith, which is always a "dreadful certainty"—*i.e.*, a dread that becomes certain of a hidden relationship with God.

The ethical and religious stages correspond roughly to what Heidegger and Jaspers call, respectively, the inauthenticity and the authenticity of existence. Art is not as a rule recognized by contemporary Existentialists as an autonomous stage; it is almost always for them an essential manifestation of existence itself. For Jaspers, it is a mode of reading in nature, in history, and in men the cipher of transcendence; *i.e.*, the negative symbol in which transcendence is revealed. According to Camus, it is an aspect of man's revolt against the world. The artist tries to remake the sketch of the world that is before him and to give it the style—that is to say, the coherence and unity—that it lacks. For this purpose, he selects the elements of the world and freely combines them in order to create a value that escapes man continuously but that the artist perceives and tries to salvage from the flux of history.

From this point of view, art would be a way of reshaping the world beyond its factual forms, in order that it might show their negative and troublesome characteristics. The directions of contemporary art that have deliberately forsaken the imitation of reality find their justification in this point of view.

SOCIAL AND HISTORICAL PROJECTIONS OF EXISTENTIALISM

The metaphysical or theological dimension of Existentialism does not leave man with nothing to do. Once the nullity of the existential possibilities is recognized, man cannot but resign himself to Being, which, in one of its new manifestations in the world or beyond it, conducts him to a new epoch. Even someone like José Ortega y Gasset, the leading Spanish Existentialist and writer, who, in examining the social aspects of existence, has characterized the present epoch by the advent of the masses and the socialization of man, has halted at the recognition of the crisis and the total uncertainty that dominates the future of man (*La rebelión de las masas* [1929; *The Revolt of the Masses,* 1932]).

On the other hand, humanistic Existentialism has recognized the positive and the to-some-degree determining function that man may have in history. It has insisted, as in Merleau-Ponty, on man's duty to assume the responsibility of an effective action for the transformation of society and, in general, of the world that he inhabits.

Along this line of assuming responsibility, Existentialism has moved toward Marxism, with which it shares the diagnoses of existence as the primordial and ineradicable relationship of man with nature and with society. In the *Critique de la raison dialectique* (1960; "Critique of Dialectical Reason"). Sartre attempted a synthesis between Existentialism and Marxism by modifying the notion of "project" that he defended in *L'Être et le néant* and by utilizing the notion of dialectic as understood by Marx. The project of which existence consists is not the result of an arbitrary choice (as Sartre had previously maintained); it is, instead, that of a conditioning by the objective possibilities that Sartre identifies (as does Marx) with "the material conditions of existence." The project remains, however, that of the particular individual of a unique consciousness—but of a consciousness that tries to become totalized; *i.e.*, to enter into relationship with others so as to constitute, with others, human groups that are more and more comprehensive. In this manner it tends toward a complete and definitive totalization without appeals. Dialectical reason would be precisely such a process of growing totalization; and it becomes, moreover, the true protagonist of history and becomes that with which the interior freedom of any individuals who participate in history is identified.

Sartre's synthesis of Existentialism and Marxism

From the defense of the freedom of the individual, Sartre has thus moved to the defense of the absolute dialectical necessity of history despite its being interiorized and lived by individuals. A historical project of human life that tries to remove the characteristics of inauthenticity or of alienation from existence—a project that may bring Existentialism and Marxism close together—thus ends by losing, in this form, its risky and problematic character and the awareness of the conditions and the modalities of its realization. These features are also lost in the "transcendental project" of a new society elaborated by one of the leaders of the New Left, the German-born American Herbert Marcuse. While insisting on the requirement that the "transcendental project" be "in accord with the real possibilities open at the attained level of the material and intellectual culture," Marcuse entrusts its realization to an impersonal and contemplative Reason, which cannot but invite the "great refusal" of contemporary society.

Having developed in different and contrasting directions, Existentialism has furnished philosophy and the whole of contemporary culture with conceptual tools, of which the nature and techniques of employment have still not been clarified—as, for example, terms like "problematicity," "chance," "condition," "choice," "freedom," and "project." Although these tools can be employed usefully for the interpretation of existence—*i.e.*, to orient philosophical inquiry in the fields of epistemology, ethics, aesthetics, education, and politics—it is nonetheless indispensable that the pivot on which they turn, "possibility," be granted its own ontological status that does not reduce it either to Nothingness or to Being. It is indispensable, moreover, that a positive datum be perceived in possibility, a datum that is verifiable with appropriate techniques and that, while not offering infallible guarantees, allows man to project and to act in the world with calculated risks and with a prudent trust. (N.Ab.)

Idealism

As a philosophical term, Idealism refers to any view that stresses the central role of the ideal or the spiritual in man's interpretation of experience. It may hold that the world or reality exists essentially as spirit or consciousness, that abstractions and laws are more fundamental in reality than sensory things, or, at least, that whatever exists is known to man in dimensions that are chiefly mental—through and as ideas. Thus the two basic forms of Idealism are metaphysical Idealism, which asserts the ideality of reality, and epistemological Idealism, which holds that in the knowledge process the mind can grasp only the psychic or that its objects are conditioned by their perceptibility. In its metaphysics, Idealism is thus directly opposed to Materialism, the view that the basic substance of the world is matter and that it is known primarily through and as material forms and processes; and in its epistemology, it is opposed to Realism, which holds that in human knowledge objects are grasped and seen as they really are—in their existence outside and independently of the mind. As a philosophy often expressed in bold and expansive syntheses, Idealism is also opposed to various restrictive forms of thought: to Skepticism, with occasional exceptions as in the British Hegelian F.H. Bradley (1846–1924); to Positivism, which stresses observable facts and relations as opposed to ultimates and therefore spurns the speculative "pretensions" of every metaphysic; and often to atheism, since the Idealist commonly extrapolates the concept of mind to embrace an infinite Mind. The essential orientation of Idealism can be sensed through some of its typical tenets: "Truth is the whole, or the Absolute"; "to be is to be perceived"; "reality reveals its ultimate nature more faithfully in its highest qualities (mental) than in its lowest (material)"; "the Ego is both subject and object."

Contrasts with opposing schools

APPROACHES TO UNDERSTANDING IDEALISM

What Idealism is may be clarified by approaching it in three ways: through its basic doctrines and principles,

through its central questions and answers, and through its significant arguments.

Six core concepts

Basic doctrines and principles. Six common, basic conceptions distinguish Idealistic philosophy:

The union of individuality and universality. Abstract universals, such as "caniness," which express the common nature or essence that the members of a class (*e.g.*, individual dogs or wolves) share with one another, are acknowledged by all philosophers. Many Idealists, however, emphasize the concept of a concrete universal, one that is also a concrete reality, such as "mankind" or "literature," which can be imagined as gatherable into one specific thing. As opposed to the fixed, formal, abstract universal, the concrete universal is essentially dynamic, organic, and developing. Thus universality and individuality merge.

The contrast between contemporaneity and eternity. While most philosophers tend to focus on matters of contemporary concern, Idealists always seek a much wider perspective that embraces epochs and eras in the broad sweep of history. In the words of the 17th-century Rationalist Spinoza, they strive to view the contemporary world "under the aspect of eternity." Thus, in spite of the extensive formative influence of culture, Idealists claim that their philosophy transcends the parochialism of a particular culture; and Idealisms are found, in fact, in all of the major cultures of the world.

The doctrine of internal relations and the coherence theory of truth. It seems natural to suppose, as non-Idealists usually do, that the consideration of two things in their relatedness to one another can have no effect on the things themselves—*i.e.*, that a relation is something in addition to the things or terms related and is thus external. On this basis, truth would be defined as a relation of correspondence between a proposition and a state of affairs. The Idealist believes, however, that reality is more subtle than this. The relationship between a mineral deposit and the business cycle, for example, is an internal one: the deposit changes to an ore when prices render it profitable to mine the mineral. Similarly, it is part of the essence of a brick that it is related to a wall or pavement. Thus terms and relations logically determine one another. Ultimate reality is therefore a system of judgments or propositions, and truth is defined in terms of the coherence of these propositions with one another to form a harmonious whole. Thus a successful spy is judged either a hero or a villain only in relation to a total system of international relations, an accepted philosophy of history, and the moral judgments involved. There are therefore degrees of reality and degrees of truth within a system of truth cohering by internal relations, and the truth of a judgment reflects its place in this system.

The dialectical method. Idealism seeks to overcome contradictions by penetrating into the overall coherent system of truth and continually creating new knowledge to be integrated with earlier discoveries. Idealism is thus friendly to all quests for truth, whether in the natural or behavioral sciences or in art, religion, and philosophy. It seeks the truth in every positive judgment and in its contradictory as well. Thus it uses the dialectical method of reasoning to remove the contradictions characteristic of human knowledge. Such removal leads to a new synthetic judgment that incorporates in a higher truth the degree of truth that was present in each of the two lower judgments.

The centrality of mind in knowledge and being. Idealism is not reductive, as are opposing philosophies that identify mind with matter and reduce the higher level of reality to the protons and electrons of mathematical physics. On the contrary, Idealism defends the principle that the lower is explained by the higher—specifically, that matter can be explained by mind but that mind cannot be explained by matter. The word spirit can be substituted for "mind" or even placed above it; and "Spiritualism" is often used, especially in Europe, as a synonym for Idealism.

The transmutation of evil into good. Nearly all Idealists accept the principle that the evils with which man has to deal may become ingredients in a larger whole that overcomes them. The eminent American Hegelian Josiah Royce (1855–1916) held that the larger whole is the Absolute Mind, which keeps evils under control as a man might hold a viper under the sole of his boot. Along with this doctrine of the sublimation or transmutation of evil, Royce incorporated into his metaphysics a point from the 19th-century irrationalism of Schopenhauer, itself a voluntaristic form of Idealism, viz., that "the world is my idea." Schopenhauer, however, was probably the only Idealist who defended the converse principle that good is transmuted into evil.

Basic questions and answers. In defining philosophical Idealism in its historical development as a technical metaphysical doctrine, three most difficult and irreducible questions arise. From the efforts to answer these questions there has been created an extensive literature that is the corpus of philosophical Idealism.

Metaphysical and logical issues

Ultimate reality. The first of the three questions is metaphysical: What is the ultimate reality that is given in human experience? Historically, answers to this question have fallen between two extremes. On the one hand is the Skepticism of the 18th-century Empiricist David Hume, who held that the ultimate reality given in experience is the moment by moment flow of events in the consciousness of each individual. This concept compresses all of reality into a solipsistic specious present—the momentary sense experience of one isolated percipient. At the other extreme, followers of the 17th-century Rationalist Spinoza adopted his definition of ultimate substance as that which can exist and can be conceived only by itself. According to the first principle of his system of pantheistic Idealism, God, or Nature, or Substance is the ultimate reality given in human experience. Hegel said that this dogmatic absolutism was the lion's den into which all tracks enter and from which none ever returns. In answering the first question, most philosophical Idealists steer between Hume and Spinoza and in so doing create a number of types of Idealism, which will be discussed below.

The given. The second question to arise in defining Idealism is: What is given? What results can be obtained from a logical interpretation and elaboration of the given? According to Idealists the result, though it is frequently something external to individual experience, is, nevertheless, a concrete universal, an order system (like the invisible lattice structure of a crystal), or an ideality in the sense explained earlier. In Hegel's words: "What is real is rational, and what is rational is real." Idealists believe that the collective human spirit of intellectual inquiry has discovered innumerable order systems that are present in external, nonhuman reality, or nature, and that this collective creative intelligence has produced the various sciences and disciplines. This production has required a long period of time called history. But history was antedated by the achievements of ancestors who created languages and religions and other primitive institutions. Consequently, the logical interpretation and elaboration of the given is actually the complete transformation of the earth by its various inhabitants; so that the moon flights portend a similar transformation of the planetary system. An inherent part of the collective intelligence is the spiritual force that Idealists call the spirit of philosophy.

Change. The third question is: What position or attitude is a thinker to take toward temporal becoming and change, and toward the presence of ends and values within the given? According to Idealists, reason not only discovers a coherent order in nature but also creates the state and other cultural institutions, which together constitute the cultural order of a civilized society. Idealistic political philosophers recognize the primacy of this cultural order over the private order or family and over the public order—the governing agencies and economic institutions. The conservation and enhancement of the values of all three orders is the basic moral objective of every civilized people. A useful distinction drawn by the German philosopher Ernst Cassier (1874–1946), a member of the late 19th- and 20th-century Marburg school of Neo-Kantianism (see below), between the efficient energies and the formative energies of a people emphaszes the way in which these moral forces function: the efficient energies are the conserving, and the formative are the creative forces in society. It is on the basis of this distinction that Idealists have made a contribution to international ethics,

Order and value in nature and society

which charges that no nation has a right to use its efficient energies to exercise power over another civilized people except to further the formative energies of that people, to enrich their cultural order. Ethically, then, there can be no power over without power for; economic exploitation is wrong.

Modern Idealists have also created an Idealistic philosophy of history. An eminent early 20th-century Italian Idealist, Benedetto Croce (1866–1952), expressed it in the formula "every true history is contemporary history"; and at the same time in France a subjective Idealist, Léon Brunschvicg (1869–1944), agreed. There are close relations between the philosophy of history and the philosophy of values.

Basic arguments. Idealists delight in arguments. They agree with Socrates and Plato in thinking that every philosopher should follow the argument wherever it leads, and, like them, they believe that it will eventually lead to some type of Idealism. Four basic arguments found in the literature of Idealism may be briefly summarized.

Epistemological issues

The esse est percipi (to be is to be perceived) argument of Berkeley. According to this argument all of the qualities attributed to objects are sense qualities. Thus hardness is the sensing of a resistance to a striking action, and heaviness is a sensation of muscular effort when holding the object in one's hand, just as blueness is a quality of visual experience. But these qualities exist only while they are being perceived by some subject or spirit equipped with sense organs. A classical 18th-century British Empiricist, George Berkeley (1685–1753), rejected the idea that sense perceptions are caused by material substance, the existence of which he denied. Intuitively he grasped the truth that "to be is to be perceived." The argument is a simple one, but it has provoked an extensive and complicated literature, and to some contemporary Idealists it seems irrefutable.

The reciprocity argument. Closely related to the *esse est percipi* argument is the contention that subject and object are reciprocally dependent upon each other. It is impossible to conceive of a subject without an object, since the essential meaning of being a subject is being aware of an object and that of being an object is being an object to a subject, this relation being absolutely and universally reciprocal. Consequently, every complete reality is always a unity of subject and object—*i.e.,* an immaterial ideality, a concrete universal.

The mystical argument. In the third argument, the Idealist holds that in man's most immediate experience, that of his own subjective awareness, the intuitive self can achieve a direct apprehension of ultimate reality, which reveals it to be spiritual. Thus the mystic bypasses normal cognition, feeling that, for metaphysical probings, the elaborate processes of mediation interposed between sense objects and their perceptions reduces its reliability as compared to the direct grasp of intuition.

It is significant that the claims of this argument have been made by numerous thinkers, in varying degrees Idealistic and mystical, living in different periods and in different cultures. In ancient Greece, for example, it was made by Plato, to whom the final leap to the Idea of the Good was mystical in nature. In Indian Hindu Vedānta philosophy it was made by the 9th-century monistic theologian Śaṅkara, by the 12th-century dualistic Brahmin theist Rāmānuja, and by the recent philosopher-president of India Sarvepalli Radhakrishnan. In Buddhism the claims were made by the sometimes mystical, extreme subjectivism of the Vijñānavāda school of Mahāyāna (represented by Aśvaghoṣa in the 1st and Asaṅga in the 4th century) and in China by the Ch'an school and by the 7th-century scholar Hui-neng, author of its basic classic *The Platform Scripture.* In Islāmic lands it was made by Ṣūfīs (mystics)—in particular, by the 13th-century Persian writer Jalāl ad-Dīn ar-Rūmī. And in the recent West it was made by several distinguished Idealists: in Germany, by the seminal modern theologian Friedrich Schleiermacher (1768–1834); in France, by the evolutionary intuitionist Henri Bergson (1859–1941), by the philosopher of action Maurice Blondel (1861–1949), and by the Jewish religious Existentialist Martin Buber (1878–1965); and in English-speaking countries, by the Scottish metaphysician James Frederick Ferrier (1808–64) and the American Hegelian William E. Hocking (1873–1966).

The ontological argument. This famous argument originated as a proof of the existence of God. It came to the 11th-century Augustinian, St. Anselm of Canterbury, as an intuitive insight from his personal religious experience that a being conceived to be perfect must necessarily exist, for otherwise he would lack one of the essentials of perfection. God's perfection requires his existence. Some Idealist philosophers have generalized the argument to prove Idealism. They distinguish conceptual essences that exist only in the intellect from categorial essences that actually exist *in re* (in the thing). Every actual reality, therefore, is a unity of one or more categorial essences and existence; and again, this means that it is an immaterial ideality or concrete universal. According to Hegel "the ideality of the finite" is "the main principle of philosophy."

TYPES OF PHILOSOPHICAL IDEALISM

Several types of Idealism have already been distinguished. Some modern types should now be mentioned, classified first by cultures and then by branches of philosophy.

Types classed by culture. Cultural differences suggest a division into Western and Oriental Idealisms.

Subjective, transcendental, objective, and absolute Idealisms

Western types. Berkeley's Idealism is called subjective Idealism because he reduced reality to spirits (his name for subjects) and the ideas entertained by spirits. In Berkeley's philosophy the apparent objectivity of the world outside the self was accommodated to his subjectivism by claiming that its objects are ideas in the mind of God. The foundation for a series of more objective Idealisms was laid in the late 18th century by Immanuel Kant, whose epochal work *Kritik der reinen Vernunft* (2nd ed., 1787; *Critique of Pure Reason,* 1929) presented a formalistic or transcendental Idealism, so named because Kant thought that the human self, or "transcendental ego," constructs knowledge out of sense impressions, upon which are imposed certain universal concepts that he called categories. Three systems constructed in the early 19th century by, respectively, the moral Idealist J.G. Fichte, the aesthetic Idealist F.W.J. Schelling, and the dialectical Idealist G.W.F. Hegel, all on a foundation laid by Kant, are called objective Idealisms in contrast to Berkeley's subjective Idealism. The designations, however, are not consistent; and when the contrast with Berkeley is not at issue, Fichte himself is often called a subjective Idealist, inasmuch as he exalted the subject above the object, employing the term Ego to mean God in the two memorable propositions: "The Ego posits itself" and "The Ego posits the non-Ego (or nature)." And in contrast now to the subjective Idealism of Fichte, Schelling's is called an objective Idealism and Hegel's an absolute Idealism.

All of these terms form backgrounds for contemporary Western Idealisms, most of which are based either on Kant's transcendental Idealism or on those of Fichte, Schelling, and Hegel. Exceptions are those based on other great Idealists of the past—Plato, Plotinus, Spinoza, Leibniz, and others. A revised form of Spinoza's spiritual monism, for example, which held that reality is one Substance to be identified with God, has been formulated by the Idealist logician H.H. Joachim (1868–1938), a follower of the British Hegelian F.H. Bradley.

Kantian and theistic Idealisms

Unwilling to accept any of the above titles, one school of modern Idealists adopted the motto "Back to Kant" and are thus called Kantian Idealists. Edward Caird (1835–1908), who imported German Idealism into England, and the German philosopher of "As If," Hans Vaihinger (1852–1933), who held that much of man's so-called knowledge reduces to pragmatic fictions, were Kantian Idealists or transcendentalists. On this tradition are based the Idealism of the austerely religious essayist Thomas Carlyle (1795–1881) in *Sartor Resartus* (1833–34) and the New England transcendentalism of Ralph Waldo Emerson (1803–82). It must be stated, however, that Kant preferred the name critical Idealism to that of transcendental Idealism.

Another group of Idealists, adopting the motto "From Kant forward," founded the so-called Marburg school of Neo-Kantian, or scientific, Idealism. They rejected the

Idealisms of Fichte, Schelling, and Hegel and the classical Newtonian dynamics presupposed by Kant and built instead upon the new quantum and relativity theories of modern physics. Founded by Hermann Cohen (1842–1918), champion of a new interpretation of Kant, and his colleague, the Platonic scholar Paul Natorp (1854–1924), who applied Kant's critical method to humanistic as well as to scientific studies, this school underwent a remarkable development, especially under the leadership of Ernst Cassirer, noted for his profound analyses of man defined as that animal that creates culture through a unique capacity for symbolic representation. The Russian novelist Boris Pasternak, in his *Autobiography,* tells of enrolling in Cohen's graduate seminar on Kant at the University of Marburg. Undoubtedly this type of Idealism continues to wield considerable influence on intellectuals in Soviet Russia.

Theistic Idealism was founded by the medical instructor R.H. Lotze (1817–81), who became a broadly learned metaphysician and whose theory of the world ground, in which all things find their unity, has been widely accepted by theistic philosophers and Protestant theologians. To Lotze, the world ground is the transcendent synthesis of an evolutionary world process, which is both mechanical and teleological (purposive); it is an infinite spiritual being, or God. In England, the absolute Idealism of T.H. Green (1836–82), a philosopher influenced chiefly by Plato and Kant, was shared by his disciple, the more Hegelian thinker Bernard Bosanquet (1848–1923), whose views are based upon Lotze's Idealism, and by the somewhat skeptical metaphysician of the movement, F.H. Bradley (1846–1924).

Theistic absolutism is represented by a pioneer of contemporary philosophical theology, F.R. Tennant (1866–1957), and by the eminent German-American theologian Paul Tillich (1886–1956). It differs from the personalistic form of absolute Idealism in accepting the traditional theological monotheism that is essential to the Jewish, Christian, and Islāmic religions. It revives classic arguments for the existence of God that were rejected by Kant and uses recent advances in the physical, biological, and behavioral sciences to support these revisions. The cosmological argument, for example, is restated as the continuing relation of the cosmos to a world ground that is spiritual in essence; thus the concept of God as a first cause is rejected. The concept of the fitness of the environment to life and to human history and other recent scientific concepts are used to modernize the teleological argument. Nevertheless, all of this revision is kept within the framework of Idealistic metaphysics and epistemology. A theistic spiritual pluralism, which interprets reality in terms of a multitude of interacting psychic monads (elementary units), was developed by the English philosopher James Ward (1843–1925). On the other hand, an atheistic spiritual pluralism, which holds that reality consists entirely of individual minds and their contents, was espoused by the Cambridge Hegelian J.M. Ellis McTaggert (1866–1925).

Hegelian Idealism, Personalism, and other Idealisms

During the late 19th century a movement known as American Hegelian Idealism arose in the United States. The movement found vigorous early expression in the work of W.T. Harris (1835–1909), central figure in a midwestern group of scholars known as the St. Louis school and editor of its *Journal of Speculative Philosophy,* and finds current expression in the recently organized Hegel Society of America. In its later development, American Idealism split into two branches: one, of the aforementioned Bradley–Bosanquet type, and a second, of the Royce–Hocking type, so called because it was founded by one of America's most distinguished philosophers, the absolute Idealist and personal pantheist Josiah Royce (1855–1916), and developed by his disciple W.E. Hocking (1873–1966). The American philosopher of religion Borden Parker Bowne (1847–1910) founded another important American school, that of Personalism, a Kantian- and Lotzean-based variety of theistic Idealism similar to the spiritual pluralism of Ward. Whereas most previous Idealisms had stressed the rational as the highest category of reality and hence as its paradigm, Personalism saw in the centred structures of personhood, both finite and infinite, an even higher category, displaying dimensions richer than the rational alone. Personalism has had an influential development in America, most notably through the Methodist philosopher E.S. Brightman (1884–1953), known for his defense of the doctrine of a finite God, and through *The Personalist,* edited by one of Bowne's disciples, R.T. Flewelling (1871–1960). Personalism is also found in the French philosopher C.B. Renouvier (1815–1903) and in several Latin American philosophers.

To the above types should be added the vitalism or creative evolutionism of the French anti-intellectualist Henri Bergson (1859–1941), which first found in the apprehension of subjective time a more valid insight into reality than in that of an objective space–time order and then, extending this metaphysics to the cosmic level, discerned there an Idealistic *élan vital* (or vital impetus) that is more fundamental than matter, which subsequently appeared in the role of a husk born of the mechanization of the *élan.* In this same tradition, the voluntarism of Maurice Blondel (1861–1949), a unique theory of belief in God as a live option that must be deliberately willed by the self before it can be found to be true in experience, is an important contribution to Idealistic philosophy. Miguel de Unamuno y Jugo (1864–1936), a Spanish philosopher, developed a unique type of Idealism, more literary than philosophical. He stressed the significance of each individual and argued for personal immortality.

Eastern types. For centuries, philosophical Idealism has dominated the philosophy of India. An Idealism that is quite influential in Japan is that of Nishida Kitarō, a distinguished Berlin-trained philosopher. Prior to World War II, Kitarō created a system of absolute Idealism that employed the dialectical method of Hegel to clarify the Zen Buddhist doctrine of nothingness, which, in his view, is that of which all phenomenal existences are determinations and in which they all appear.

Some classical types of Indian and Chinese Idealism were considered above (see *The mystical argument*). A number of gifted Indian and Chinese scholars have restated and revitalized the principles and arguments of classic Oriental Idealisms in an extensive literature.

Indian Idealism

Probably the major recent proponent of Indian Idealism has been Radhakrishnan, who has spent a long lifetime expounding and defending its mystical types and has presented authoritative analyses of all of its classical systems. He saw his modernized Idealism as destined to save civilization from exploitation by Western commercial technology. Surendranath Dasgupta, an outstanding Sanskrit and Pāli scholar, in a monumental work, has revived the classic systems of Indian Idealism, concluding that "Idealism has not only been one of the most dominant phases of Indian thought in metaphysics, epistemology, and dialectics, but it has also very largely influenced the growth of the Indian ideal as a whole." Ghose Aurobindo, reinterpreting the Indian Idealistic heritage in the light of his own Western education, rejected the *māyā* doctrine of illusion, replacing it with the concept of evolution. Arguing that the "illumination of individuals will lead to the emergence of a divine community," Aurobindo founded the influential Pondicherry Ashram, a religious and philosophical community, and headed it until his death. Late in the 19th century, Swami Vivekananda, a spiritual monist, promulgated the Idealistic philosophy of mystical Brahmanism in lectures on the Vedānta delivered and published widely.

The inwardness of subjectivity of Indian Idealism has been contrasted with the outwardness of Western objective Idealism, and a synthesis of the two has been advocated in comparative studies made by P.T. Raju, an Indian philosopher who has taught both in Indian universities and in the U.S.

Prior to World War II, Sir Rabindranath Tagore, a distinguished Hindu Idealist poet and Nobel laureate, contributed to what Dasgupta has called the "Indian ideal as a whole." A selection from Tagore's aphorisms will convey its spirit:

> Let your life lightly dance on the edges of Time like dew on the tip of a leaf.
> Our little heaven, where dwell only two immortals, is too absurdly narrow.

Is it then true that the mystery of the Infinite is written on this little forehead of mine?
Where is this hope for union except in thee my God?
Raise my veil and look at my face proudly, O Death, my Death.
All is done and finished in the eternal Heaven. But Earth's flowers of illusion are kept externally fresh by death.
If my claims to immortal fame after death are shattered, make me immortal while I live.
This I know that the moment my God has created me he has made himself mine.

Chinese Idealism

In addition to the Ch'an and Hui-neng schools mentioned above (see *The mystical argument*), three other notable Idealistic schools have flourished in China. Representing one wing of the Neo-Confucian movement of the 11th and 12th centuries, Ch'eng Hao and his disciple, the rationalist Chu Hsi, developed a dualistic philosophy that has been compared to Cartesianism. In this view, however, reason takes precedence over matter and the two together are the primary cause of the universe or the absolute; thus this view is essentially Idealistic. At the turn of the 15th century, a more purely Idealistic school arose—forming the other wing of Neo-Confucianism—under the leadership of Wang Yang-ming, who, having had an inner experience of enlightenment, sought to understand the cosmos within his own mind and heart. The third school is that of the 20th-century Idealist Hsung Shih-li, who, borrowing to some extent from Wang Yang-ming, proclaimed a "new doctrine of consciousness only," of which the basic ideas are the unity of substance and function and the primacy of the original Mind. To Hsung Shih-li, reality and all of its manifestations are one, and the original Mind is will and consciousness as well as reason.

Types classed by branches of philosophy. Another way of classifying Idealisms is to use branches of philosophy to distinguish the various types. Such types, however, overlap those given above.

Metaphysical, epistemological, and axiological Idealisms

A term that covers several of the above types (the spiritual, theistic, and Hegelian; Personalism; vitalism) is metaphysical Idealism. A.N. Whitehead (1861–1947), noted for his collaboration with Bertrand Russell in mathematical logic and for his process metaphysics, who was profoundly influenced by Bradley, created an original Idealistic philosophy of science, a highly complicated form of metaphysical Idealism; and the leading metaphysician Charles Hartshorne (1897–) may be regarded as a representative of Whiteheadian Idealism, although rightly claiming originality. Epistemological Idealism, of which the Kantian scholar N.K. Smith's (1872–1958) *Prolegomena to an Idealist Theory of Knowledge* is an excellent example, covers all Idealistic theories of epistemology, or knowledge. Aesthetic Idealism is devoted to philosophical theories of beauty in nature and in all forms of art. Because Schelling claimed that art is the best approach to an understanding of philosophy, his system is designated aesthetic Idealism. Axiological Idealism is a name referring to such philosophies as those of Wilbur M. Urban (1873–1952) and others who have developed Idealistic theories of value and valuation. Ethical Idealism deals with moral values, rights, and obligations. Several of the above-mentioned philosophers, such as Fichte and Green, as well as the Plato scholar A.E. Taylor (1864–1945), the theistic pluralist Hastings Rashdall (1858–1924), and the absolutist W.R. Sorley (1855–1935), could be called ethical Idealists in the sense that they have produced well-thought-out systems of ethics. The writings of the German philosopher of life and action Rudolf Eucken (1846–1926) provide an excellent example of ethical Idealism.

These classifications are not exhaustive. The actual existence of so many types of philosophical Idealism, however, proves its fertility and ubiquity.

CRITICISM AND APPRAISAL

Obviously, some of the types of Idealism in the above classifications conflict with one another. For example, spiritual monism and spiritual pluralism are opposite types; Personalism rejects absolute Idealism; and atheistic spiritual pluralism is in sharp conflict with theistic spiritual pluralism. These and other debatable issues keep Idealists in dialogue with each other, but each type tends to preserve itself.

Perry, Moore, Positivism, Marxism

Over against these internal disputes stand the criticisms of the anti-Idealists. The wide-ranging Realist Ralph Barton Perry's (1876–1957) article "The Ego-Centric Predicament" (1910) is a widely discussed criticism. Perry admitted that the primary approach of every philosopher to the problem of ultimate reality must be through his own thought, using his own ideas; but this is a human predicament that has been unjustifiably exploited by the Idealists, according to Perry, and turned into the "fallacious" *esse est percipi* argument.

The famous "Refutation of Idealism" prepared by the meticulous Cambridge philosopher G.E. Moore (1873–1958) and a similar refutation by the Realist Bertrand Russell (1872–1970) rest upon the distinction between a subject's act of perceiving and the perceptual object of this act, which they both called a "sense datum." They claimed that Berkeley's *esse est percipi* argument is vitiated by his failure to make this distinction.

Logical Positivism claims that a basic weakness in Idealism is its rejection of the doctrine of empirical verifiability, according to which every proposition that claims to be true must be verified by searching out the sense experience in which its terms originated. Linguistic philosophy attacks Idealism by making a detailed analysis of its more technical terms in an effort to prove that they are full of ambiguities and double meanings. Critics have also severely attacked the ontological and the mystical arguments for Idealism. Karl Marx (1818–83) and his followers borrowed and adapted the dialectical argument of Hegel and used it effectively to develop dialectical Materialism, an archenemy of all Idealisms. Buttressed by the political endorsements of various Communist regimes, Marxism (*q.v.*) poses a formidable opposition to Idealism; and even in the non-Communist countries of Europe it presents a significant cultural alternative to spiritualism and Thomism.

Idealists consider all of the foregoing criticisms to be external. Instead of answering them in detail, some Idealists prefer to challenge the critics to make really constructive efforts to build an adequate substitute for Idealism—a system to be reached by seriously working at the problems from within philosophy. So far a satisfactory substitute has not been achieved. To produce such a substitute would require careful reconsideration of the arguments of at least some of the above Idealistic systems.

Prognosis

In evaluating the effects of these criticisms and attacks, the question remains: Will they succeed in eradicating philosophical Idealism? Although it is now on the wane, at least in Western culture, the great Idealist tradition has survived many other historic periods of turmoil and has often been reborn in prolonged periods of settled and peaceful social conditions. Will it rise again? Only the future holds the answer. But Idealism shows evidence of being, perhaps, a reflection of some permanent aspect of the human spirit, and it may then be a perennial philosophy. In any case, it seems highly unlikely that such a rich heritage of philosophical thought will vanish entirely.

(D.S.R.)

Materialism

The word Materialism has been used in modern times to refer to a family of metaphysical theories (*i.e.*, theories on the nature of reality) that can best be defined by saying that a theory tends to be called Materialism if it is felt sufficiently to resemble a paradigmatic theory that will here be called mechanical Materialism. This section covers the various types of Materialism and the ways by which they are distinguished and traces the history of Materialism from the Greeks and Romans to modern and contemporary Materialisms.

TYPES OF MATERIALIST THEORY

Mechanical Materialism is the theory that the world consists entirely of hard, massy material objects, which, though perhaps imperceptibly small, are otherwise like such things as stones. (A slight modification is to allow

the void—or empty space—to exist also in its own right.) These objects interact in the sort of way that stones do: by impact and possibly also by gravitational attraction. The theory denies that immaterial or apparently immaterial things (such as minds) exist or else explains them away as being material things or motions of material things.

Physicalistic, deistic, emergent, and epiphenomenal types

Types distinguished by departures from the paradigm. In modern physics (if interpreted realistically), however, matter is conceived as made up of such things as electrons, protons, and mesons, which are very unlike the hard, massy, stonelike particles of mechanical Materialism. In it the distinction between matter and energy has also broken down. It is therefore natural to extend the word Materialist beyond the above paradigm case (of mechanical Materialism) to cover anyone who bases his theory on whatever it is that physics asserts ultimately to exist. This sort may be called physicalistic Materialism. Such a Materialist allows the concept of material thing to be extended so as to include all of the elementary particles and other things that are postulated in fundamental physical theory—perhaps even continuous fields and points of space-time. Inasmuch as some cosmologists even try to define the elementary particles themselves in terms of the curvature of space-time, there is no reason why a philosophy based on such a geometricized cosmology should not be counted as Materialist, provided that it does not give an independent existence to nonphysical things such as minds.

Another sort of departure from the paradigm leads in the direction of what might be called a deistic Materialism. In this view it would be allowed that, although there is a spiritual Creator of the universe, he does not interfere with the created universe, which is itself describable in terms of mechanical or physicalist Materialism.

Still another departure from the paradigm is the theory that holds that everything is composed of material particles (or physical entities generally) but also holds that there are special laws applying to complexes of physical entities, such as living cells or brains, that are not reducible to the laws that apply to the fundamental physical entities. (To avoid inconsistency, such a theory may have to allow that the ordinary laws of physics do not wholly apply within such complex entities.) Such a theory, which could be called "emergent Materialism," can shade off, however, into theories that one would not wish to call Materialist, such as hylozoism, which ascribes vital characteristics to all matter, and panpsychism, which attributes a mindlike character to all constituents of material things.

Another common relaxation of the paradigm is that which allows as compatible with Materialism such a theory as epiphenomenalism, according to which sensations and thoughts do exist in addition to material processes but are nonetheless wholly dependent on material processes and without causal efficacy of their own. They are related to material things somewhat in the way that a man's shadow is related to the man. A similar departure from the paradigm is a form of what might be called "double-aspect Materialism," according to which in inner experience men are acquainted with nonphysical *properties* of material processes, though these properties are not causally effective. A form of double-aspect theory in which these properties were allowed to be causally effective would be a species of emergent Materialism.

Of course, more than one of these qualifications might be made at the same time: thus a person might wish to speak of "physicalist deistic epiphenomenalist Materialism." If no other qualifications are intended, it is convenient to use the word extreme and to speak, for example, of "extreme physicalist Materialism"—which is probably the type most discussed among professional philosophers in English-speaking countries.

Dialectical Materialism

Type distinguished by its view of history. In the wider world, however, the word Materialism most commonly brings to mind dialectical Materialism, which is the orthodox philosophy of Communist countries. This is most importantly a theory of how changes arise in human history, though a general metaphysical theory lies in the background. Dialectical Materialists contrast their view with what they call "vulgar" Materialism; and it does, indeed, appear that their theory is not an extreme Materialism, whether mechanical or physicalist. They seem to hold merely that mental processes are dependent on or have evolved from material ones. Though they might be akin to emergent Materialists, it is hard to be sure; their assertion that something new emerges at higher levels of organization might refer only to such things as that a wireless receiver is different from a mere heap of the same components. And if so, even an extreme physicalistic Materialist could acquiesce in this view. The distinctive features of dialectical Materialism would, thus, seem to lie as much in its being dialectical as in its being Materialist. Its dialectical side may be epitomized in three laws: (1) that of the transformation of quality into quantity, (2) that of the interpenetration of opposites, and (3) that of the negation of the negation. Nondialectical philosophers find it hard, however, to interpret these laws in a way that does not make them into either platitudes or falsehoods.

Perhaps because of the historical determinism implicit in dialectical Materialism, and perhaps because of memories of the mechanical Materialist theories of the 18th and 19th centuries, when physics was deterministic, it is popularly supposed that Materialism and determinism must go together. This is not so. As indicated below, even some ancient Materialists were indeterminists, and a modern physicalist Materialism must be indeterministic because of the indeterminism that is built into modern physics. Modern physics does imply, however, that macroscopic bodies behave in a way that is effectively deterministic, and, because even a single neuron (nerve fibre) is a macroscopic object by quantum mechanical standards, a physicalistic Materialist may still regard the human brain as coming near to being a mechanism that behaves in a deterministic way.

Central-state Materialism and analytical behaviourism

Types distinguished by their account of mind. A rather different way of classifying Materialist theories, which to some extent cuts across the classifications already made, emerges when the theories are divided according to the way in which a Materialist accounts for minds. A central-state Materialist identifies mental processes with processes in the brain. An analytical behaviourist, on the other hand, argues that, in talking about the mind, one is not talking about an actual entity, whether the brain or an immaterial soul, but, rather, one is somehow talking about the way in which people would behave in various circumstances. According to the analytical behaviourist, there is no more of a problem for the Materialist in having to identify mind with something material than there is in identifying such an abstraction as the average plumber with some concrete entity. Analytical behaviourism differs from psychological behaviourism, which is merely a methodological program to base theories on behavioral evidence and to eschew introspective reports. The analytical behaviourist usually has a not too plausible theory of introspective reports according to which they are what are sometimes called "avowals": roughly, he contends that to say "I have a pain" is to engage in a verbal surrogate for a wince. Epistemic Materialism is a theory that can be developed either in the direction of central-state Materialism or in that of analytical behaviourism and that rests on the contention that the only statements that are intersubjectively testable are either observation reports about macroscopic physical objects or statements that imply such observation reports (or are otherwise logically related to them).

Before leaving this survey of the family of Materialistic theories, a quite different sense of the word Materialism should be noted in which it denotes not a metaphysical theory but an ethical attitude. A person is a Materialist in this sense if he is interested mainly in sensuous pleasures and bodily comforts and hence in the material possessions that bring these about. A man might be a Materialist in this ethical and pejorative sense without being a metaphysical Materialist, and conversely. An extreme physicalistic Materialist, for example, might prefer a Beethoven record to a comfortable mattress for his bed; and a person who believes in immaterial spirits might opt for the mattress.

HISTORY OF MATERIALISM

Greek and Roman Materialism. Though Thales of Miletus (*c.* 580 BC) and some of the other Pre-Socratic philoso-

phers have some claims to being regarded as Materialists, the Materialist tradition in Western philosophy really begins with Leucippus and Democritus, Greek philosophers who were born in the 5th century BC. Leucippus is known only through his influence on Democritus. According to Democritus, the world consists of nothing but atoms (indivisible chunks of matter) in empty space (which he seems to have thought of as an entity in its own right). These atoms can be imperceptibly small, and they interact either by impact or by hooking together, depending on their shapes. The great beauty of atomism was its ability to explain the changes in things as due to changes in the configurations of unchanging atoms. The view may be contrasted with that of the earlier philosopher Anaxagoras (*c.* 480 BC), who thought that when, for example, the bread that a person eats is transformed into human flesh, this must occur because bread itself already contains hidden within itself the characteristics of flesh. Democritus thought that the soul consists of smooth, round atoms and that perceptions consist of motions caused in the soul atoms by the atoms in the perceived thing (see above *Atomism*).

Democritus and Epicurus

Because Epicurus' philosophy was expounded in a lengthy poem by Lucretius, a Roman philosopher of the 1st century BC, Epicurus (died 270 BC) was easily the most influential Greek Materialist. He differed from Democritus in that he postulated an absolute up–down direction in space, so that all atoms fall in roughly parallel paths. To explain their impacts with one another, he then held that the atoms are subject to chance swerves—a doctrine that was also used to explain free will. Epicurus' Materialism therefore differed from that of Democritus in being an indeterministic one. Epicurus' philosophy contained an important ethical part, which was a sort of enlightened egoistic hedonism (see above *Epicureanism*). His ethics, however, were not Materialistic in the pejorative sense of the word.

Modern Materialism. Materialism languished throughout the medieval period, but the Epicurean tradition was revived in the first half of the 17th century in the atomistic Materialism of the French Catholic priest Pierre Gassendi. In putting forward his system as a hypothesis to explain the facts of experience, Gassendi showed that he understood the method characteristic of modern science, and he may well have helped to pave the way for corpuscular hypotheses in physics. Gassendi was not thoroughgoing in his Materialism inasmuch as he accepted on faith the Christian doctrine that men have immortal souls. His contemporary, the English philosopher Thomas Hobbes, also propounded an atomistic Materialism and was a pioneer in trying to work out a mechanistic and physiological psychology. Holding that sensations are corporeal motions in the brain, Hobbes skirted, rather than solved, the philosophical problems about consciousness that had been raised by another contemporary, the great French philosopher René Descartes. Descartes's philosophy was dualistic, making a complete split between mind and matter. In his theory of the physical world, however, and especially in his doctrine that animals are automata, Descartes's own system had a mechanistic side to it that was taken up by 18th-century Materialists, such as Julien de La Mettrie, the French physician whose appropriately titled *L'Homme machine* (1747; *Man a Machine,* 1750) applied Descartes's view about animals to man himself. Denis Diderot, an 18th-century French Encyclopaedist, supported a broadly Materialist outlook by considerations drawn from physiology, embryology, and the study of heredity; and his friend Paul, baron d'Holbach, published his *Système de la nature* (1770), which expounded a deterministic type of Materialism in the light of evidence from contemporary science, reducing everything to matter and to the energy inherent in matter. He also propounded a hedonistic ethics as well as an uncompromising atheism, which provoked a reply even from the Deist Voltaire.

Gassendi through Büchner and Vogt

The 18th-century French Materialists had been reacting against orthodox Christianity. In the early part of the 19th century, however, certain writers in Germany—usually with a biological or medical background—reacted against a different orthodoxy, the Hegelian and Neo-Hegelian tradition in philosophy, which had become entrenched in German universities. Among these were Ludwig Büchner and Karl Vogt. The latter is notorious for his assertion that the brain secretes thought just as the liver secretes bile. This metaphor of secretion, previously used by P.-J.-G. Cabanis, a late 18th-century French Materialist, is seldom taken seriously, because to most philosophers it does not make sense to think of thought as a *stuff.* The Hobbesian view, also espoused by Büchner, that thought is a *motion* in the brain is usually viewed as a more promising one.

The synthesis of urea (the chief nitrogenous end product of protein metabolism), discovered in 1828, broke down the discontinuity between the organic and the inorganic in chemistry, which had been a mainstay of nonmaterialistic biology. Materialist ways of thinking were later strengthened enormously by the Darwinian theory of evolution, which not only showed the continuity between man and other living things right back to the simplest organisms but also showed how the apparent evidences of design in natural history could be explained on a purely causal basis. There still seemed to be a gap, however, between the living and the nonliving, though E.H. Haeckel, a 19th-century German zoologist, thought that certain simple organisms could have been generated from inorganic matter and, indeed, that a certain simple sea creature may well be in process of generation in this way even now. Though Haeckel was wrong, 20th-century biologists have proposed much more sophisticated and more plausible theories of the evolution of life from inorganic matter. Haeckel and his contemporary, the British zoologist T.H. Huxley, did much to popularize philosophical accounts of the world that were consonant with the scientific thought of their time, but neither could be regarded as an extreme Materialist.

Darwin, Haeckel, and Huxley

Contemporary Materialism. Perhaps because recent developments in biochemistry and in physiological psychology have greatly increased the plausibility of Materialism, there has lately been a resurgence of interest in the philosophical defense of central-state Materialism. Central-state Materialists have proposed their theories partly because of dissatisfaction with the analytical behaviourism of the Oxford philosopher Gilbert Ryle. Ryle himself is reluctant to call himself a Materialist, partly because of a dislike of all "isms" and partly because he thinks that the notion of matter has meaning only by contrast with that of mind, which he thinks to be an illegitimate sort of contrast. Nevertheless, it would seem that analytical behaviourism could be used to support a physicalist Materialism that would go on to explain human behaviour by means of neural mechanisms. (Ryle himself is suspicious of mechanistic accounts of biology and psychology.) Analytical behaviourism has been felt to be unsatisfactory, however, chiefly because of its account of introspective reports as avowals (see above), which most philosophers have found to be unconvincing.

Philosophers have distinguished two forms of central-state Materialism, namely, the translation form and the disappearance form. The translation form is the view that mentalistic discourse can be translated into discourse that is neutral between physicalism and dualism, so that the truth of a man's introspective reports is compatible with the objects of these reports being physical processes. The disappearance form is the view that such a translation cannot be done and that this fact, however, does not refute physicalism but shows only that man's ordinary introspective reports are contaminated by false theories.

Translation central-state theories. Among the philosophers who have advocated the translation form is the U.S. philosopher Herbert Feigl, earlier a member of the Vienna Circle, who, in an influential monograph (see *Bibliography*), did the most to get contemporary philosophers to treat central-state Materialism as a serious philosophical theory. Against the objection that, for example, "visual sensation" does not *mean* "process in the visual cortex," advocates of the translation form point out that "the morning star" does not mean the same as "the evening star," and yet the morning star as a matter of fact *is* the evening star. The objection confuses meaning and reference. Against the objection that a purely physical process (a dance of

Feigl, Place, Smart, Armstrong

electrons, protons, and so on) cannot have the sensory quality of greenness that is observed in a visual experience of seeing grass, say, they reply that to talk of the sensory experience of something looking green (or having a green mental image) is not to talk of anything that is literally green, but is simply to report that some internal process is of the sort that normally goes with seeing something, such as a lawn, which really is green. Though an immaterialist might say that the sort of process in question is a spiritual process, the Materialist can equally claim that it is a material process in the brain. The analysis of the introspective report is neutral between these two contentions; the Materialist, however, opts for his contention on various grounds. The British Materialist U.T. Place does so on the ground of normal scientific methodology; and the Australian Materialist J.J.C. Smart does so with a metaphysical application of the principle (called "Ockham's razor") that entities should not be multiplied beyond necessity. A physicalistic Materialist has, of course, an obligation to go on to give a suitable account of such apparently nonphysicalist qualities as the greenness of grass. At one time Smart analyzed colours in terms of the discriminatory behaviour of human beings. Another Australian Materialist, D.M. Armstrong, holds, on the other hand, that colours are as a matter of fact properties of objects, such properties being of the sort describable in the theoretical terms of physics. Feigl, in turn, is to some extent (and rather reluctantly) a double-aspect theorist. He qualifies the position taken by the other translation theorists, conceding that the translations do leave something out, viz., the immediately introspectable properties of "raw feels," such as that of hearing the tone of middle C. He holds, however, that such properties are irrelevant to causal explanations of phenomena. The translation form of central-state Materialism thus has some affinities with the earlier epistemic Materialism of the Positivist philosophers Rudolf Carnap and Hans Reichenbach, Germans who settled in the United States. Thus Carnap has suggested that mental predicates be treated as applying to material entities: for example, "Carnap sees green" could be taken as meaning "the body Carnap is in the state of green-seeing," the state of green-seeing being a purely physical state that explains the behavioral facts that led one to ascribe the predicate "sees green" to Carnap in the first place. David K. Lewis, a United States philosopher of science and language, has developed a translation form of central-state Materialism on the basis of a theory regarding the definition of theoretical terms in science. According to this theory, entities such as electrons, protons, and neutrons are defined in terms of the causal roles that they play in relation to observational phenomena—*e.g.*, phenomena in cloud chambers—but the method of definition is able to do justice to the causal and other interrelations between the theoretical entities themselves. Lewis applies this account to commonsense psychology. Since mental entities, such as pains, are defined in commonsense psychology in terms of their causal roles (in relation to observable behaviour) and since there is empirical reason to ascribe the same causal roles to brain processes, Lewis identifies mental events, processes, and states with brain events, processes, and states.

Disappearance central-state theories. The disappearance form of central-state Materialism is the sort of theory held by P.K. Feyerabend, a U.S. philosopher, who denies that the Materialist can give a neutral analysis of introspective reports. In Feyerabend's view, commonsense introspective reports are irreducibly immaterialist in content. He argues, however, that this admission does not show the untenability of Materialism. Ordinary mentalistic discourse, he holds, is comparable to the medieval discourse about epileptics as being "possessed by the devil." If one now "identified" demon possession with a certain medical condition of the brain, this would really be an assertion that there is no such thing as a demon-possessed state: the medieval way of looking at the matter is thus rejected. It is in this sort of way that Feyerabend wants to "identify" the mind with the brain: he simply rejects the ordinary mentalistic conceptual scheme and so feels no obligation to show its compatibility with Materialism.

Feyerabend, Quine, Sellars, Wittgenstein

The influential American philosophers W.V. Quine and Wilfrid Sellars also hold theories that could be regarded as disappearance forms of physicalistic Materialism, though there is a Kantian twist to Sellars' philosophy that makes it hard to classify. Sellars holds that mentalistic concepts cannot be eliminated from man's commonsense picture of the world, which he calls "the manifest image." In a way reminiscent of Kant he holds that, although the manifest image is inescapable, it does not give metaphysical truth about the world as it really is in itself. This truth is given, instead, by "the scientific image"—*i.e.*, by theoretical science, which is physicalist. In the case of Quine, there is a certain Platonism in that he believes in the objective reality of nonspatiotemporal entities, viz., those that are the subject matter of pure mathematics. Because he holds that the reason for believing mathematics is that it is needed as part of physical theory, his reasons for believing in numbers and the like are not in principle different from those for believing in electrons; thus Quine's Platonism does not really compromise his physicalism.

The Austrian philosopher Ludwig Wittgenstein, who lived to the mid-20th century and was professor of philosophy at Cambridge University, has sometimes been interpreted as a behaviourist, though his insistence that an inner process stands in need of outward criteria could possibly be interpreted as a sort of epistemic and central-state Materialism. Nevertheless, to count Wittgenstein as a Materialist would be to take considerable liberties with him; for, while displaying at times a certain mystical attitude, he also held very strongly that the business of a philosopher is not to put forward any metaphysical theory but to clear up conceptual confusions—to show the fly the way out of the fly bottle.

Eastern Materialism. This historical survey has been concerned with Materialism in Western philosophy. On the whole, Materialism is contrary to the spirit of both Indian and traditional Chinese philosophy, though the Cārvāka school of Materialists flourished from the 6th century BC until medieval times in India. Mention should also be made of the strong naturalistic tendency in Theravāda Buddhism, as also in certain schools of Chinese philosophy that exalt *ch'i* ("ether" or "material force") above principle and mind.

SUBSTANTIVE ISSUES IN MATERIALISM

Reductionism, consciousness, and brain. The main attraction of Materialism today is the way in which it fits in with a unified picture of science—a picture that has become very plausible. Thus, chemistry is reducible to physics inasmuch as there is a quantum-mechanical theory of the chemical bond. Biology is mainly an application of physics and chemistry to the structures described in natural history (including the natural history that one can explore through powerful microscopes). Increasingly, biological explanations resemble explanations in engineering, in which material structures are described and then the laws of physics and chemistry are used to explain the behaviour of these structures. (In the biological case, of course, these structures are often dynamic in the sense that their molecules are continually being replaced.) Through the influence of neurophysiology and also cybernetics (the science of information and control, which can be applied also to artificial automata), scientific psychology is also fitting well into the same mechanistic scheme. The recalcitrant residue appears in the phenomena of consciousness. Here mental events seem, indeed, to be correlated with physical events: but, if the mental events are not the very same as the physical events, one is left with apparently ultimate (or irreducible) physical–mental laws that do not fit happily into unified science, and one is thus faced with a situation unlike that of the rest of science. Looking at science generally, one expects *ultimate* laws to relate simple entities, such as fundamental particles. A physical–mental law, however, would have to refer to something very complex—a brain process involving perhaps millions of neurons, with each neuron being itself an almost fantastically complex entity. There would be a multitude of physical–mental laws, which would look like excrescences on the face of science. Because they would not fit into

Unified science and physical–mental laws

the network of scientific laws, Herbert Feigl has called them "nomological danglers." To get rid of these danglers is one of the chief attractions of Materialism. Of course, an immaterialist might assert that mental entities exist and also that there are no physical–mental laws. But it might be hard for him to reconcile this position with the empirical evidence; and in any case he would be faced with the problem of how to distinguish the free exercise of such anomalous physical–mental interaction from mere chance behaviour.

Artificial intelligence and generative grammar

The development of computers and other devices to take over much of the more routine sort of human behaviour has led to attempts on the part of scientists and technologists, such as the American M.L. Minsky, to develop real artificial intelligence. So far, the success that these scientists hoped for has not been achieved. An American linguistic theorist, Noam Chomsky, has argued on the basis of his theories of generative grammar that the brain is quite unlike any already-understood type of mechanism. Indeed, any physicalistic Materialist must certainly concede that there are very deep problems about the brain, which apparently can no longer be thought of as a bundle of conditioned reflex mechanisms or the like, as it often has seemed to be to many psychologists. The physicalist can stress, however, that the investigator's ignorance need not lead him to assume that he will never be able to find an explanation of intelligence and of linguistic abilities in terms consonant with his present notion of a physical mechanism. (There is also the possibility that physical laws not yet discovered might be needed to explain the workings of the brain. So long as these turned out to be basic laws of physics, such discoveries would not imply a shift to emergentist Materialism.)

Logic, intentionality, and psychical research. Some philosophers, such as the Oxford philosopher J.R. Lucas, have tried to produce positive arguments against a mechanistic theory of mind by employing certain discoveries in mathematical logic, especially Gödel's theorem, which implies that no axiomatic theory could possibly capture all arithmetical truths. In general, philosophers have not found such attempts to extract an antimaterialist philosophy from mathematical logic to be convincing. Nevertheless, the problems of mechanizing intelligence, including the mathematical abilities of human beings, do pose unsolved problems that the Materialist is obliged to take seriously.

Perhaps the most common challenge to Materialism comes from philosophers who hold that it cannot do justice to the concept of intentionality, which Franz Brentano, a pre-World War I German philosopher, made the distinguishing mark between the mental and the nonmental. (A related objection is that Materialism cannot do justice to the distinction between behaviour and mere bodily movements.) Brentano held that mental events and states somehow point toward objects beyond themselves (or have a "content"). Many contemporary philosophers agree with Brentano that purely physical entities cannot have this property. If it is said, for example, that punched holes on the tape of a computer can refer beyond themselves in the way that thoughts do, then it is commonly replied that, in themselves, the holes on the tape have no reference or content—for this belongs only to the thoughts in the mind of a person who reads the tape. The Materialist reply may be to argue, however, that there is a fundamental unclarity in the very notion of intentionality (this is roughly Quine's position) or else to argue that purely physical systems can, after all, possess intentionality.

The alleged spiritualistic and other phenomena reported in psychical research are sometimes adduced against Materialism. The Materialist, however, can well afford to postpone discussion of these phenomena until such time as they are accepted by the general scientific community, which on the whole still remains skeptical of them.

At present, there are reputable philosophers who accept Materialism, and there are also reputable philosophers who either reject it as false or hold that it is not the business of a philosopher to propound any sort of metaphysical system. Perhaps Materialists are still in a minority; but at any rate there is much less tendency than there was a generation ago for this type of theory to be thought philosophically naive. (J.J.C.S.)

Phenomenology

In the 20th century, Phenomenology is mainly used as the name for a philosophical movement the primary objective of which is the direct investigation and description of phenomena as consciously experienced, without theories about their causal explanation and as free as possible from unexamined preconceptions and presuppositions. The word itself is much older, going back at least to the 18th century, when the Swiss-German mathematician and philosopher Johann Heinrich Lambert applied it to that part of his theory of knowledge that distinguishes truth from illusion and error. In the 19th century the word became associated chiefly with the *Phänomenologie des Geistes* (1807; *Phenomenology of Mind,* 2nd ed., 1931), by G.W.F. Hegel, who traced the development of the human spirit from mere sense experience to "absolute knowledge." The so-called Phenomenological movement did not get under way, however, until early in the 20th century. But even this new Phenomenology includes so many varieties that a comprehensive characterization of the subject requires their consideration.

CHARACTERISTICS OF PHENOMENOLOGY

In view of the spectrum of Phenomenologies that have issued directly or indirectly from the original work of the Austrian-born German philosopher Edmund Husserl, it is not easy to find a common denominator for such a movement beyond its common source. But similar situations occur in other philosophical as well as non-philosophical movements.

Essential features and variations. Although, as seen from Husserl's last perspective, all departures from his own views could only appear as heresies, a more generous assessment will show that all those who consider themselves Phenomenologists subscribe, for instance, to his watchword, *Zu den Sachen selbst* ("To the things themselves"), by which they meant the taking of a fresh approach to concretely experienced phenomena, an approach, as free as possible from conceptual presuppositions, and the attempt to describe them as faithfully as possible. Moreover, most adherents to Phenomenology hold that it is possible to obtain insights into the essential structures and the essential relationships of these phenomena on the basis of a careful study of concrete examples supplied by experience or imagination and by a systematic variation of these examples in imagination. Some Phenomenologists also stress the need for studying the ways in which the phenomena appear in men's object-directed ("intentional") consciousness.

"To the things themselves"

Beyond this merely static aspect of appearance, some also want to investigate its genetic aspect, exploring, for instance, how the phenomenon intended—for example, a book—shapes ("constitutes") itself in the typical unfolding of experience. Husserl himself believed that such studies require a previous suspension of belief in the reality of these phenomena, whereas others consider it not indispensable but helpful. Finally, in existential Phenomenology, the meanings of certain phenomena (such as anxiety) are explored by a special interpretive ("hermeneutic") Phenomenology, the methodology of which needs further clarification.

Contrasts with related movements. It may also be helpful to bring out the distinctive essence of Phenomenology by confronting it with some of its philosophical neighbours. In contrast to Positivism and to traditional Empiricism, from which Husserl's teacher at Vienna, Franz Brentano, had started out and with which Phenomenology shares an unconditional respect for the positive data of experience ("We are the true positivists," Husserl claimed in his *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie* [1913; "Ideas for a Pure Phenomenology and Phenomenological Philosophy"]), Phenomenology does not restrict these data to the range of sense experience but admits on equal terms such non-sensory ("categorial") data as relations and values, as long as they present themselves

intuitively. Consequently, Phenomenology does not reject universals; and, in addition to analytic a priori statements, whose predicates are logically contained in the subjects and the truth of which is independent of experience (*e.g.*, "All material bodies have extension"), and the synthetic a posteriori statements, whose subjects do not logically imply the predicate and the truth of which is dependent on experience (*e.g.*, "My shirt is red"), it recognizes knowledge of the synthetic a priori, a proposition whose subject does not logically imply the predicate but one in which the truth is independent of experience (*e.g.*, "Every colour is extended"), based on insight into essential relationships within the empirically given.

In contrast to phenomenalism, a position in the theory of knowledge (epistemology) with which it is often confused, Phenomenology—which is not primarily an epistemological theory—accepts neither the rigid division between appearance and reality nor the narrower view that phenomena are all that there is (sensations or permanent possibilities of sensations). These are questions on which Phenomenology as such keeps an open mind—pointing out, however, that phenomenalism overlooks the complexities of the intentional structure of men's consciousness of the phenomena.

In contrast to a Rationalism that stresses conceptual reasoning at the expense of experience, Phenomenology insists on the intuitive foundation and verification of concepts and especially of all a priori claims; in this sense it is a philosophy from "below," not from "above."

In contrast to an Analytic philosophy that substitutes simplified constructions for the immediately given in all of its complexity and applies "Ockham's razor," Phenomenology resists all transforming reinterpretations of the given, analyzing it for what it is in itself and on its own terms.

Phenomenology shares with Linguistic Analysis a respect for the distinctions between the phenomena reflected in the shades of meaning of ordinary language as a possible starting point for phenomenological analyses. Phenomenologists, however, do not think that the study of ordinary language is a sufficient basis for studying the phenomena, because ordinary language cannot and need not completely reveal the complexity of phenomena.

In contrast to an Existential philosophy that believes that human existence is unfit for phenomenological analysis and description, because it tries to objectify the unobjectifiable, Phenomenology holds that it can and must deal with these phenomena, however cautiously, as well as other intricate phenomena outside the human existence.

(H.Sp.)

ORIGIN AND DEVELOPMENT OF HUSSERL'S PHENOMENOLOGY

Husserl's earlier development

Basic principles. Phenomenology was not founded; it grew. Its fountainhead was Edmund Husserl, who held professorships at Göttingen and Freiburg im Breisgau and who wrote *Die Idee der Phänomenologie* (*The Idea of Phenomenology,* 1964) in 1906. Yet, even for Husserl, the conception of Phenomenology as a new method destined to supply a new foundation for both philosophy and science developed only gradually and kept changing to the very end of his career. Trained as a mathematician, Husserl was attracted to philosophy by Franz Brentano, whose descriptive psychology seemed to offer a solid basis for a scientific philosophy. The concept of intentionality, the directedness of the consciousness toward an object, which is a basic concept in Phenomenology, was already present in Brentano's *Psychologie vom empirischen Standpunkte* (1874): "And thus we can define psychic phenomena by saying that they are those phenomena which, precisely as intentional, contain an object in themselves." Brentano dissociated himself here from Sir William Hamilton, known for his philosophy of the "unconditioned," who had attributed the character of intentionality to the realms of thought and desire only, to the exclusion of that of feeling.

The point of departure of Husserl's investigation is to be found in the treatise *Der Begriff der Zahl* (1887; "The Concept of Number"), which was later expanded into *Philosophie der Arithmetik: Psychologische und logische Untersuchungen* (1891). Numbers are not found ready-made in nature but result from a mental achievement. Here Husserl was preoccupied with the question of how something like the constitution of numbers ever comes about. This treatise is important to Husserl's later development for two reasons: first, because it contains the first traces of the concepts "reflection," "constitution," "description," and the "founding constitution of meaning," concepts that later played a predominant role in Husserl's philosophy; and second, because it reflected two events—(1) a criticism of his book by Gottlob Frege, a seminal thinker in logic, who had charged him with confusing logical and psychological considerations, and (2) Husserl's discovery of the *Wissenschaftslehre* (1837; *Logic and Scientific Methods,* 1971) by Bernard Bolzano, a Bohemian mathematician, theologian, and social moralist, and his view concerning "truths in themselves"—which led Husserl to an analysis and critical discussion of psychologism, the view that psychology could be used as a foundation for pure logic, which he clearly felt to be no longer possible.

In the first volume of *Logische Untersuchungen* (1900–01; *Logical Investigations,* 1970), entitled *Prolegomena,* Husserl began with a criticism of psychologism. And yet he continued by conducting a careful investigation of the psychic acts in and through which logical structures are given; these investigations, too, could give the impression of being descriptive psychological investigations, though they were not conceived of in this way by the author. For the issue at stake was the discovery of the essential structure of these acts. Here Brentano's concept of intentionality received a richer and more refined signification. Husserl distinguished between perceptual and categorical intuition and stated that the latter's theme lies in logical relationships. The real concern of Phenomenology was clearly formulated for the first time in his *Logos* article, "Philosophie als strenge Wissenschaft" (1910–11; *Philosophy as Rigorous Science,* 1965). In this work Husserl wrestled with two unacceptable views: naturalism and historicism.

Anti-naturalism

Naturalism attempts to apply the methods of the natural sciences to all other domains of knowledge, including the realm of consciousness. Reason becomes naturalized. Although an attempt is then made to find a foundation for the human sciences (*Geisteswissenschaften*) by means of experimental psychology, it proves to be impossible, because in so doing one is unable to grasp precisely what is at stake in knowledge as found in the natural sciences.

What a philosopher must examine is the relationship between consciousness and Being; and in doing so, he must realize that from the standpoint of epistemology, Being is accessible to him only as a correlate of conscious acts. He must thus pay careful attention to what occurs in these acts. This can be done only by a science that tries to understand the very essence of consciousness; and this is the task that Phenomenology has set for itself. Because clarification of the various types of objects must follow from the basic modes of consciousness, Husserl's thought remained close to psychology. In contradistinction to what is the case in psychology, however, in Phenomenology, consciousness is thematized in a very special and definite way, viz., just insofar as consciousness is the locus in which every manner of constituting and founding meaning must take place. In man's intuition, conscious occurrences must be given immediately in order to avoid introducing at the same time certain interpretations. The nature of such processes as perception, representation, imagination, judgment, and feeling must be grasped in immediate self-givenness. The call "To the things themselves" is not a demand for realism, because the things at stake are the acts of consciousness and the objective entities that get constituted in them: these things form the realm of what Husserl calls the phenomena.

Thus, the objects of Phenomenology are "absolute data grasped in pure, immanent intuition," and its goal is to discover the essential structures of the acts (*noesis*) and the objective entities that correspond to them (*noema*).

Anti-historicism

On the other hand, Phenomenology must also be distinguished from historicism, a philosophy that stresses the immersion of all thinkers within a particular historical

setting. Husserl objected to historicism because it implies relativism. He gave credit to Wilhelm Dilthey, author of "Entwürfe zur Kritik der historischen Vernunft" ("Outlines for the Critique of Historical Reason," in *Gesammelte Schriften,* 12 vol. [1914–36], vol. 5, 6), for having developed a typification of world views, but he doubted and even rejected the skepticism that flows necessarily from the relativity of the various types. History is concerned with facts, whereas Phenomenology deals with the knowledge of essences. To Husserl, Dilthey's doctrine of world views was incapable of achieving the rigour required by genuine science. Contrary to all of the practical tendencies found in world views, Husserl demanded that philosophy be founded as a rigorous science. Its task implies that nothing be accepted as given beforehand but that the philosopher should try to find the way back to the real beginnings. This is tantamount to saying, however, that he must try to find the way to the foundations of meaning that are found in consciousness. Just as for Immanuel Kant the empirical has merely relative validity and never an absolute, or apodictic, validity, so for Husserl, too, what is to be searched for is a scientific knowledge of essences in contradistinction to a scientific knowledge of facts.

Basic method. The basic method of all Phenomenological investigation, as Husserl developed it himself—and on which he worked throughout his entire lifetime—is the "reduction": the existence of the world must be put between brackets, not because the philospher should doubt it but merely because this existing world is not the very theme of Phenomenology; its theme is rather the manner in which knowledge of the world comes about. The first step of the reduction consists in the phenomenological reduction, through which all that is given is changed into a phenomenon in the sense of that which is known in and by consciousness; for this kind of knowing—which is to be taken in a very broad sense as including every mode of consciousness, such as intuition, recollection, imagination, and judgment—is here all-important. There are several reasons why Husserl gave a privileged position to intuition: among them is the fact that intuition is that act in which a person grasps something immediately in its bodily presence and also that it is a primordially given act upon which all of the rest is to be founded. Furthermore, Husserl's stress on intuition must be understood as a refutation of any merely speculative approach to philosophy.

This reduction reverses—"re-flects"—man's direction of sight from a straightforward orientation toward objects to an orientation toward consciousness.

Eidetic and transcendental reductions

The second step is to be found in the eidetic reduction. To get hold of consciousness is not sufficient; on the contrary, the various acts of consciousness must be made accessible in such a way that their essences—their universal and unchangeable structures—can be grasped. In the eidetic reduction one must forego everything that is factual and merely occurs in this way or that. A means of grasping the essence is the *Wesensschau,* the intuition of essences and essential structures. This is not a mysterious kind of intuition. Rather, one forms a multiplicity of variations of what is given, and while maintaining the multiplicity, one focusses attention on what remains unchanged in the multiplicity; *i.e.,* the essence is that identical something that continuously maintains itself during the process of variation. Husserl, therefore, called it the invariant.

To this point, the discussion of reduction has remained within the realm of psychology, albeit a new—namely, a phenomenological—psychology. The second step must now be completed by a third, the transcendental reduction. It consists in a reversion to the achievements of that consciousness that Husserl, following Kant, called transcendental consciousness, although he conceived of it in his own way. The most fundamental event occurring in this consciousness is the creation of time awareness through the acts of protention (future) and retention (past), which is something like a self-constitution. To do phenomenology was for Husserl tantamount to returning to the transcendental ego as the ground for the foundation and constitution (or making) of all meaning (German *Sinn*). Only when a person has reached this ground can he achieve the insight that makes his comportment transparent in its entirety and makes him understand how meaning comes about, how meaning is based upon meaning like strata in a process of sedimentation.

Husserl worked on the clarification of the transcendental reduction until the very end of his life. It was precisely the further development of the transcendental reduction that led to a division of the Phenomenological movement and to the formation of a school that refused to become involved in this kind of system of problems (see below *Phenomenology of essences*).

Basic concepts. In an effort to express what it is to which this method gives access, Husserl wrote:

Intentionality and constitution theory

> In all pure psychic experiences (in perceiving something, judging about something, willing something, enjoying something, hoping for something, etc.) there is found inherently a being-directed-toward .... Experiences are intentional. This being-directed-toward is not just joined to the experience by way of a mere addition, and occasionally as an accidental reaction, as if experiences could be what they are without the intentional relation. With the intentionality of the experiences there announces itself, rather, the essential structure of the purely psychical.

The phenomenological investigator must examine the different forms of intentionality in a reflective attitude because it is precisely in and through the corresponding intentionality that each domain of objects becomes accessible to him. Husserl took as his point of departure mathematical entities and later examined logical structures, in order finally to achieve the insight that each being must be grasped in its correlation to consciousness, because each datum becomes accessible to a person only insofar as it has meaning for him. From this position, regional ontologies, or realms of being, develop: for instance, those dealing with the region of "nature," the region of "the psychic," or the region of "the spirit." Moreover, Husserl distinguished formal ontologies—such as the region of the logical—from material ontologies.

In order to be able to investigate a regional ontology, it is first necessary to discover and examine the founding act by which realities in this realm are constituted. For Husserl, constitution does not mean the creation or fabrication of a thing or object by a subject; it means the founding constitution of its meaning. There is meaning only for consciousness. All founding constitution of meaning is made possible by transcendental consciousness. Speaking of this transcendental motif, Husserl wrote:

> It is the motif of questioning back to the last source of all achievements of knowledge, of reflection in which the knower reflects on himself and his knowing life, in which all the scientific constructs which have validity for him, occur teleologically, and as permanent acquisitions are kept and become freely available to him.

In the realm of such transcendental problems, it is necessary to examine how all of the categories in and through which one understands mundane beings or purely formal entities originate from specific modes of consciousness. In Husserl's view, the temporalization must be conceived as a kind of primordial constitution of transcendental consciousness itself.

Understood in this way, Phenomenology does not place itself outside the sciences but, rather, attempts to make understandable what takes place in the various sciences and, thus, to thematize the unquestioned presuppositions of the sciences.

In his last publication, *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie: Eine Einleitung in die phänomenologische Philosophie* (1936; *The Crisis of European Sciences and Transcendental Phenomenology,* 1970), Husserl arrived at the life-world—the world as shaped within the immediate experience of each man—by questioning back to the foundations that the sciences presuppose. In *Die Krisis* he analyzed the grounds that had led to the European crisis of culture and philosophy, which found its immediate expression in the contrast between the great successes of the sciences of nature and the failure of the sciences of man. In the modern era, scientific knowledge had become fragmented into an objectivistic-physicalist and a transcendental knowledge. Until recently this split could not be overcome. It led, rather, to the attempt to develop the sciences of man in

Husserl's notion of the life-world

accordance with the procedures used in the exact sciences of nature (naturalism)—an attempt doomed to failure. In opposition to this attempt, Husserl wished to show that in the new approach one must reflect on the activities of the scientists.

> As the immediately given world, this merely subjective world, was forgotten in the scientific thematization, the accomplishing subject, too, was forgotten and the scientist himself was not thematized.

Husserl demonstrated this point by using the example of Galileo and his mathematization of our world. The truth characteristic of the life-world is by no means an inferior form of truth when compared to the exact, scientific truth but is, rather, always a truth already presupposed in all scientific research. That is why Husserl claimed that an ontology of the life-world must be developed—*i.e.*, a systematic analysis of the constitutive achievements the result of which is the life-world, a life-world that, in turn, is the foundation of all scientific constitutions of meaning. The stimulating change that occurred here consists in the fact that truth is no longer measured after the criterion of an exact determination. For what is decisive is not the exactness but, rather, the part played by the founding act.

It is in this connection that, rather abruptly, historicity, too, became relevant for Husserl. He began to reflect upon the emergence of philosophy among the Greeks and on its significance as a new mode of scientific knowledge oriented toward infinity; and he interpreted the philosophy of René Descartes, often called the father of modern philosophy, as the point at which the split into the two research directions—physicalist objectivism and transcendental subjectivism—came about. Phenomenology must overcome this split, he held, and thus help mankind to live according to the demands of reason. In view of the fact that reason is the typical characteristic of man, mankind must find itself again through Phenomenology.

LATER DEVELOPMENTS

Phenomenology of essences. A different type of Phenomenology, the Phenomenology of essences, developed from a tangential continuation of that of the *Logische Untersuchungen.* Its supporters were Husserl's students in Göttingen and a group of young philosophers in Munich, originally students of Theodor Lipps, a Munich psychologist and philosopher—students who had turned away from Lipp's psychologism and discovered powerful support in Husserl. The Phenomenological movement, which then began to take shape, found its most tangible expression in the publication of the *Jahrbuch für Philosophie und phänomenologische Forschung* (1913–30), a Phenomenological yearbook with Husserl as its main editor, the preface of which defined Phenomenology in terms of a return to intuition (*Anschauung*) and to the essential insights (*Wesenseinsichten*) derived from it as the ultimate foundation of all philosophy.

Fruits of the movement in broader application

The 11 volumes of the *Jahrbuch* contained, in addition to Husserl's own works, the most important fruits of the movement in its broader application. Of the co-editors, Alexander Pfänder contributed chiefly to the development of phenomenological psychology and pure logic but developed also the outlines of a complete Phenomenological philosophy. Moritz Geiger applied the new approach particularly to aesthetics and Adolf Reinach to the philosophy of law. The most original and dynamic of Husserl's early associates, however, was Max Scheler, who had joined the Munich group and who did his major Phenomenological work on problems of value and obligation. A Polish philosopher, Roman Ingarden, did major work in structural ontology and analyzed the structures of various works of art in its light; Hedwig Conrad-Martius, a cosmic Realist at the University of Munich, worked intensively in the ontology of nature; and others made comparable contributions in other fields of philosophy. None of these early Phenomenologists, however, followed Husserl's road to transcendental Idealism; and some tried to develop a Phenomenology along the lines of Realism.

Heidegger's hermeneutic Phenomenology. Martin Heidegger, one of Germany's foremost philosophers at the middle of the 20th century, was inspired to philosophy through Brentano's work *Von der mannigfachen Bedeutung des Seienden nach Aristoteles* (1862; "On the Multifarious Meaning of Being According to Aristotle"). While he was still studying theology, from 1910 to 1911, Heidegger encountered Husserl's *Logische Untersuchungen.* From then on he pursued the course of Phenomenology with the greatest interest, and from 1916 he belonged to the narrow circle of students and followers of the movement. The typical character of the Phenomenological intuition was at that time the focus of Husserl's seminar exercises. To be sure, there appeared very early a difference between Husserl and Heidegger. Discussing and absorbing the works of the important philosophers in the history of metaphysics was, for Heidegger, an indispensable task, whereas Husserl repeatedly stressed the significance of a radically new beginning and—with few exceptions (among them Descartes, Locke, Hume, and Kant)—wished to bracket the history of philosophy.

Heidegger's basic work, *Sein und Zeit* (1927; *Being and Time,* 1962), which was dedicated to Husserl, strongly acknowledged that its author was indebted to Phenomenology. In it, Phenomenology was understood as a methodological concept—a concept that was conceived by Heidegger in an original way and resulted from his questioning back to the meanings of the Greek concepts of *phainomenon* and *logos. Phainomenon* is "that which shows itself from itself," but together with the concept of *logos,* it means "to let that which shows itself be seen from itself in the very way in which it shows itself from itself." This conception of Phenomenology, which relied more on Aristotle than on Husserl, constituted a change that was later to lead to an estrangement between Husserl and Heidegger. For in *Sein und Zeit* there is no longer a phenomenological reduction, a transcendental ego, or an intuition of essences in Husserl's sense. Heidegger's new beginning was, at the same time, a resumption of the basic question of philosophy: that concerning the meaning (*Sinn*) of Being. His manner of questioning can be defined as hermeneutical in that it proceeds from the interpretation of man's situation. What he thematized is, thus, the explanation of what is already understood.

Heidegger's return to ontology

At the heart of *Sein und Zeit* lies Heidegger's analysis of the one (the man) who asks the question—who is capable of asking the question—concerning Being, who precisely through this capability occupies a privileged position in regard to all other beings, viz., that of *Dasein* (literally, "being there"). By conceiving of *Dasein* as being-in-the-world, Heidegger made the ancient problem concerning the relationship between subject and object superfluous. The basic structures of *Dasein* are primordial moodness (*Befindlichkeit*), understanding (*Verstehen*), and logos (*Rede*). These structures are, in turn, founded in the temporalization of *Dasein,* from which future, having-been (past), and present originate. The two basic possibilities of man's existing (from the Latin *ex* and *sistere,* "standing out from") are those in which *Dasein* either comes to its self (called authenticity) or loses itself (called inauthenticity); *Dasein* is inauthentic, for example, when it lets the possibilities of the choice for its own "ek-sisting" be given to it by others instead of deciding for itself. Heidegger's concept of care (*Sorge, cura*) has nothing to do with distress (*Bekümmernis*) but includes the unity of the articulated moments of man's being-in-the-world.

The hermeneutic character of Heidegger's thought manifested itself also in his interpretation of poetry, in which he discovered a congenial spirit in Friedrich Hölderlin, one of Germany's greater poets, of whose poetry he inaugurated a completely new interpretation; but it manifested itself equally well in his interpretation of metaphysics, which Heidegger tried to envision as an occurrence determined by the forgottenness of Being, an occurrence in the centre of which man finds himself and of which the clearest manifestation is to be found in "technicity," the attempt of modern man to dominate the earth by controlling beings that are considered as objects.

The concept of transcendental consciousness, which was central for Husserl, is not found in Heidegger—which clearly shows how Heidegger, in *Sein und Zeit,* had already dissociated himself from Husserl's Phenomenology.

Other developments. Eugen Fink, for several years Husserl's collaborator, whose essay "Die phänomenologische Philosophie Edmund Husserls in der gegenwärtigen Kritik" (1933) led to a radicalization of Husserl's philosophical, transcendental Idealism, later turned in another direction, one that approached Heidegger's position and divorced itself at the same time from that of Husserl.

Ludwig Landgrebe, who was Husserl's personal assistant for many years, published in 1938 *Erfahrung und Urteil* ("Experience and Judgment"), the first of Husserl's posthumous works devoted to the genealogy of logic. Among German-language scholars, Landgrebe remained closest to Husserl's original views and has developed them consistently in several works.

DISSEMINATION OF PHENOMENOLOGY

Phenomenology in various countries. Following upon the work of Husserl, Phenomenology spread into a worldwide movement.

In France. One of the first French authors to become familiar with Husserl's philosophy was Emmanuel Lévinas, a pluralistic Personalist, who combined ideas from Husserl and Heidegger in a very personal way. Similarly, Jean-Paul Sartre, the leading Existentialist of France, took his point of departure from the philosophies of Husserl and Heidegger. His first works, *L'Imagination* (1936; *Imagination: A Psychological Critique,* 1962) and *L'Imaginaire: Psychologie phénoménologique de l'imagination* (1940; *The Psychology of Imagination,* 1950), remain completely within the context of Husserl's analyses of consciousness. Sartre explains the distinction between perceptual and imaginative consciousness with the help of Husserl's concept of intentionality, and he frequently employs the method of ideation (*Wesensschau*).

Jean-Paul Sartre's use of the method

In *L'Être et le néant* (1943; *Being and Nothingness,* 1956), an essay on Phenomenological ontology, it is obvious that Sartre borrowed from Heidegger. Some passages from Heidegger's *Was ist Metaphysik?* (1929; *What Is Metaphysics?,* 1949), in fact, are copied literally. The meaning of nothingness, which Heidegger in this lecture made the theme of his investigations, became for Sartre the guiding question. Sartre departs from Heidegger's analytic of *Dasein* and introduces the position of consciousness (which Heidegger had overcome).

The distinction between being-in-itself (*en-soi*) and being-for-itself (*pour-soi*) pervades the entire investigation. The in-itself is the opaque, matter-like substance that remains the same, whereas the for-itself is consciousness permeated by nothingness. The influence of the Idealist G.W.F. Hegel becomes apparent when the author tries to interpret everything in a dialectical way; *i.e.,* through a tension of opposites. The dialectic of men's being-with-one-another is central: thus, seeing and being-seen correspond to dominating and being-dominated. The basic characteristic of being-for-itself is bad faith (*mauvaise foi*), which cannot be overcome because facticity (being-already) and transcendence (being-able-to-be) cannot be combined.

The Phenomenological character of Sartre's analyses of consciousness consists in the way in which he elucidates certain modes of behaviour: love, hatred, sadism, masochism, and indifference. Although Sartre sees and describes these forms of behaviour strikingly and precisely, he limits himself to those modes that fit his philosophical interpretation. The significance of psychology, recognized by Husserl, emerges again in Sartre and leads to a demand for an Existential psychoanalysis.

Sartre's definition of man as a being of possibilities that finds or loses itself in the choice that it makes in regard to itself refers to Heidegger's definition of *Dasein* as a being that has to materialize itself. For Sartre, freedom is the basic characteristic of man; thus Sartre belongs to the tradition of the great French moralist philosophers.

In his later works, as in his *Critique de la raison dialectique* (1960; *Search for a Method,* 1963), Sartre turned to Marxism, though he developed a method of understanding that was influenced by hermeneutics. Here the choice made by the individual is limited by social and psychological conditions. Sartre's outstanding two-volume interpretation of Flaubert, *L'Idiot de la famille; Gustave Flaubert de 1821–1857* (1971), is an example of this new method of understanding and interpretation, which combines Marxist elements with interpretations of a highly personal nature taken from depth psychology.

The Phenomenology of Maurice Merleau-Ponty

Maurice Merleau-Ponty (died 1961), who, together with Sartre and his associate Simone de Beauvoir, a writer and novelist, was an important representative of French Existentialism, was at the same time the most important French Phenomenologist. His works, *La Structure du comportement* (1942; *Structure of Behaviour,* 1963) and *Phénoménologie de la perception* (1945; *Phenomenology of Perception,* 1962), were the most original further developments and applications of Phenomenology to come from France. Merleau-Ponty gave a new interpretation of the meaning of the human body from the viewpoint of Phenomenology and, connected with this, of man's perception of space, the natural world, temporality, and freedom.

Starting from Husserl's later phenomenology of the life-world, Merleau-Ponty anchored the phenomena of perception in the phenomenology of the lived body (the body as it is experienced and experiences), in which the perceiving subject is incarnated as the mediating link to the phenomenal world. Such a phenomenology of human "presence" in the world was also to offer an alternative to the rigid dichotomy between Idealism and Realism, in which consciousness and world could be reciprocally related. Phenomenology thus became a way of showing the essential involvement of human existence in the world, starting with everyday perception.

Although it is true that Merleau-Ponty was originally close to Husserl in his thought, he later developed noticeably in the direction of Heidegger, a change that became particularly manifest in *L'Oeil et l'esprit* (1964; "Eye and Mind," in *The Primacy of Perception,* 1964).

Paul Ricoeur, a student of the volitional experience, whose translation of Husserl's *Ideen zu einer reinen Phänomenologie* brought Husserl closer to the younger French generation, writes in a Phenomenological vein but with the intention of further developing Husserl's conception of Phenomenology. Ricoeur's two-volume *Philosophie de la volonté* (1950–60; "Philosophy of the Will") also deals with the problems involved in the theological concept of guilt.

Suzanne Bachelard, who in 1957 translated Husserl's *Formale und transzendentale Logik: Versuch einer Kritik der logischen Vernunft,* has pointed to the significance of Husserl for modern logic; and Jacques Derrida, an original French thinker on the limits of thought and language, has combined Phenomenology and Structuralism in his interpretation of literature.

In Germany. After World War II, interest in Phenomenology sprang up again in its own homeland. The influence of Ludwig Landgrebe in Cologne has been particularly felt, as have the activities of the Husserl Archives in Cologne, with editions by Walter Biemel, who also published *Philosophische Analysen zur Kunst der Gegenwart* (1968; "Philosophical Analyses of Contemporary Art") and essays on the relationships between Husserl and Heidegger. The circle around Gerhard Funke in Mainz, author of *Phänomenologie—Metaphysik oder Methode?* (1966), has also had a positive influence.

Husserl scholarship at Cologne and Louvain

In other European countries. In Belgium, at the Catholic University of Louvain, are located the entire posthumous works of Husserl, as well as his personal library. Thanks to the initiative of H.L. Van Breda, founder of the Husserl Archives, several scholars worked intensively on the manuscripts for several decades. By 1972, 12 volumes of collected works had been published. Van Breda was also the director of the *Phaenomenologica* series—totalling 42 volumes by 1972—in which the most important publications in the field of Phenomenology (taken in a very broad sense) have been published. Thus, mainly through Van Breda's efforts, Louvain has become the most important centre for Phenomenology. Van Breda also organized international colloquia on Phenomenology. The influence of Alphonse de Waelhens, a Belgian philosopher of fresh experience and author of *Phénoménologie et vérité* (1953) and *Existence et signification* (1958), also bears mentioning.

In The Netherlands, Stephan Strasser, oriented particularly toward phenomenological psychology, has been especially influential. And in Italy, the Phenomenology circle has centred around Enzo Paci. The Husserl scholar Jan Patocka, a prominent expert in Phenomenology as well as in the metaphysical tradition, was influential in Czechoslovakia; in Poland, Roman Ingarden represented the cause of Phenomenology; and there have also been important representatives in such countries as Portugal, England, South America, Japan, and India.

In the United States. Phenomenology in the United States has lived a rather marginal existence for quite some time, notwithstanding the meritorious journal of *Philosophy and Phenomenological Research* founded by Husserl's student Marvin Farber, who is also the author of *The Foundation of Phenomenology* (1943). More recently, however, a noticeable change has taken place, chiefly because of the work of two scholars at the New School for Social Research in New York, Alfred Schütz, an Austrian-born sociologist and student of human cognition (died 1959), and Aron Gurwitsch, a Lithuanian-born philosopher. Schütz came early to Phenomenology, developing a social science on a phenomenological basis. Gurwitsch, author of *Théorie du champ de la conscience* (1957; *The Field of Consciousness,* 1964), came to Phenomenology through his study of the Gestalt psychologists Adhemar Gelb and Kurt Goldstein. While in Paris, Gurwitsch influenced Merleau-Ponty. The essays on Phenomenology published by Gurwitsch in the United States are among the best. His comprehensive knowledge ranges from mathematics, via the natural sciences, to psychology and metaphysics. The work *The Phenomenological Movement* (2nd ed., 1965), by Herbert Spiegelberg, an Alsatian-American Phenomenologist, is the movement's first encompassing historical presentation.

The work of Farber, Schütz, and Gurwitsch

Phenomenology in other disciplines. Of greater significance is the role of Phenomenology outside of philosophy proper in stimulating or reinforcing phenomenological tendencies in such fields as mathematics and the biological sciences. Much stronger was its impact on psychology, in which Franz Brentano and the German Carl Stumpf had prepared the ground and in which the U.S. psychologist William James, the Würzburg school, and the Gestalt psychologists had worked along parallel lines. But Phenomenology has probably made its strongest contribution in the field of psychopathology, in which the German Karl Jaspers, a foremost contemporary Existentialist, stressed the importance of phenomenological exploration of a patient's subjective experience. Jaspers was followed by the Swiss Ludwig Binswanger and several others. The Phenomenological strand is also very pronounced in American Existential psychiatry and has affected sociology, history, and the study of religion.

CONCLUSION

At the turn of the fourth quarter of the 20th century, it remained to be seen whether Phenomenology could make solid contributions to philosophical knowledge. To this end, it needed to develop rigorous standards, which had not always been observed by some of its most brilliant practitioners, such as Max Scheler, and which were likely to be violated in a philosophy the ultimate appeal of which had to be made to intuitive verification. With this proviso, Phenomenology may well be qualified not only to become a bridge for better international communication in philosophy but also to shed new light on philosophical problems old and new, to reclaim for philosophy parts of man's quotidian world that have been abandoned by science as too private and too subjective, and, finally, to give access to layers of man's experience unprobed in everyday living, thus providing deeper foundations for both science and life. (W.Bi.)

Positivism and Logical Empiricism

As a philosophical ideology and movement, Positivism first assumed its distinctive features in the work of the French philosopher Auguste Comte, who also named and systematized the science of sociology. It then developed through several stages known by various names, such as Empiriocriticism, Logical Positivism, and Logical Empiricism, and finally, in the mid-20th century, flowed into the movement known as Analytic and Linguistic philosophy.

Affirmations and negations

The basic affirmations of Positivism are (1) that all knowledge regarding matters of fact is based on the "positive" data of experience, and (2) that beyond the realm of fact is that of pure logic and pure mathematics, which were already recognized by the Scottish Empiricist and Skeptic David Hume as concerned with the "relations of ideas" and, in a later phase of Positivism, were classified as purely formal sciences. On the negative and critical side, the Positivists became noted for their repudiation of metaphysics; *i.e.,* of speculation regarding the nature of reality that radically goes beyond any possible evidence that could either support or refute such "transcendent" knowledge claims. In its basic ideological posture, Positivism is thus worldly, secular, antitheological, and antimetaphysical. Strict adherence to the testimony of observation and experience is the all-important imperative of the Positivists. This imperative is reflected also in their contributions to ethics and moral philosophy, and most Positivists have been Utilitarians to the extent that something like "the greatest happiness for the greatest number of people" was their ethical maxim. It is notable, in this connection, that Auguste Comte was the founder of a short-lived religion, in which the object of worship was not the deity of the monotheistic faiths but humanity.

There are distinct anticipations of Positivism in ancient philosophy. Though the relationship of Protagoras—a 5th-century-BC Sophist—for example, to later Positivistic thought was only a distant one, there was a much more pronounced similarity in the classical Skeptic Sextus Empiricus, who lived at the turn of the 3rd century AD, and in Pierre Bayle, his 17th-century reviver. Moreover, the medieval Nominalist William of Ockham had clear affinities with modern Positivism. An 18th-century forerunner who had much in common with the Positivistic antimetaphysics of the following century was the German thinker Georg Lichtenberg.

Positivism clearly has its proximate roots, however, in the French Enlightenment, which stressed the clear light of reason, and in the 18th-century British Empiricism, particularly that of Hume and of Bishop George Berkeley, which stressed the role of sense experience. Comte was influenced specifically by the Enlightenment Encyclopaedists (such as Denis Diderot, Jean d'Alembert, and others) and, especially in his social thinking, was decisively influenced by the founder of French Socialism, Claude-Henri, comte de Saint-Simon, whose disciple he had been in his early years and from whom the very designation Positivism stems.

THE SOCIAL POSITIVISM OF COMTE AND MILL

Comte's law of three phases

Comte's Positivism was posited on the assertion of a so-called law of the three phases (or stages) of intellectual development. There is a parallel, as Comte saw it, between the evolution of thought patterns in the entire history of man, on the one hand, and in the history of an individual's development from infancy to adulthood, on the other. In the first, or so-called theological, stage, natural phenomena are explained as the results of supernatural or divine powers. It matters not whether the religion is polytheistic or monotheistic; in either case, miraculous powers or wills are believed to produce the observed events. This stage was criticized by Comte as anthropomorphic; *i.e.,* as resting on all-too-human analogies. Generally, animistic explanations—made in terms of the volitions of soullike beings operating behind the appearances—are rejected as primitive projections of unverifiable entities.

The second phase, called metaphysical, is in some cases merely a depersonalized theology: the observable processes of nature are assumed to arise from impersonal powers, occult qualities, vital forces, or entelechies (internal perfecting principles). In other instances, the realm of observable facts is considered as an imperfect copy or imitation of eternal ideas, as in Plato's metaphysics of pure Forms. Again, Comte charged that no genuine explanations result; questions concerning ultimate reality,

first causes, or absolute beginnings are thus declared to be absolutely unanswerable. The metaphysical quest can lead only to the conclusion expressed by the German biologist and physiologist, Emil du Bois-Reymond: *"Ignoramus et ignorabimus"* ("We are and shall be ignorant"); it is a deception through verbal devices and the fruitless rendering of concepts as real things.

The sort of fruitfulness that it lacks can be achieved only in the third phase, the scientific, or "positive," phase—hence the title of Comte's magnum opus: *Cours de philosophie positive* (1830–42; *The Positive Philosophy of Auguste Comte,* 1853)—because it claims to be concerned only with positive facts. The task of the sciences, and of knowledge in general, is to study the facts and regularities of nature and society and to formulate the regularities as (descriptive) laws; explanations of phenomena can consist in no more than the subsuming of special cases under general laws. Mankind reached full maturity of thought only after abandoning the pseudoexplanations of the theological and metaphysical phases and substituting an unrestricted adherence to scientific method.

Classification of the sciences; sociology

In his three stages Comte combined what he considered to be an account of the historical order of development with a logical analysis of the levelled structure of the sciences. By arranging the six basic and pure sciences one upon the other in a pyramid, Comte prepared the way for Logical Positivism to "reduce" each level to the one below it. He placed at the fundamental level the science that does not presuppose any other sciences—viz., mathematics—and then ordered the levels above it in such a way that each science depends upon, and makes use of, the sciences below it on the scale; thus arithmetic and the theory of numbers are declared to be presuppositions for geometry and mechanics, astronomy, physics, chemistry, biology (including physiology), and sociology. Each higher level science, in turn, adds to the knowledge content of the science or sciences on the levels below, thus enriching this content by successive specialization. Psychology is conspicuously missing in Comte's system of the sciences. Anticipating some ideas of 20th-century Behaviourism and physicalism, Comte assumed that psychology should become a branch of biology (especially of brain neurophysiology), on the one hand, and of sociology, on the other. As the "father" of sociology, Comte maintained that the social sciences should proceed from observations to general laws, very much as (in his view) physics and chemistry do. He was skeptical of introspection in psychology, being convinced that, in attending to one's own mental states, these states would be irretrievably altered and distorted. In thus insisting on the necessity of objective observation, he was close to the basic principle of the methodology of 20th-century Behaviourism.

Among Comte's disciples or sympathizers were Cesare Lombroso, an Italian psychiatrist and criminologist, and Paul-Emile Littré, J.-E. Renan, and Louis Weber.

Despite some basic disagreements with Comte, the 19th-century English philosopher John Stuart Mill, also a logician and economist, must be regarded as one of the outstanding Positivists of his century. In his *System of Logic* (1843), he developed a thoroughly Empiricist theory of knowledge and of scientific reasoning, going even so far as to regard logic and mathematics as empirical (though very general) sciences. The broadly synthetic philosopher Herbert Spencer, author of a doctrine of the "unknowable" and of a general evolutionary philosophy, was, next to Mill, an outstanding exponent of a Positivistic orientation.

THE CRITICAL POSITIVISM OF MACH AND AVENARIUS

The influences of Hume and of Comte were also manifest in important developments in German Positivism, just prior to World War I. The outstanding representatives of this school were a philosophical critic of the physics of Newton, an Austrian, Ernst Mach, who was also an original thinker as a physicist and excelled as a historian of mechanics, thermodynamics, and optics, and Richard Avenarius, founder of a philosophy known as Empiriocriticism.

Mach's phenomenalism

Mach, in the introductory chapter of his book *Beiträge zur Analyse der Empfindungen* (1886; *Contributions to the Analysis of the Sensations,* 1897), reviving the Humean antimetaphysics, contended that all factual knowledge consists of a conceptual organization and elaboration of what is given in the elements; *i.e.,* in the data of immediate experience. Very much in keeping with the spirit of Comte, he repudiated the transcendental Idealism of Kant. For Mach, the most objectionable feature in Kant's philosophy was the doctrine of the *Dinge an sich—i.e.,* of the "things-in-themselves"—the ultimate entities underlying phenomena, which Kant had declared to be absolutely unknowable though they must nevertheless be conceived as partial causes of man's perceptions. Hermann von Helmholtz, a wide-ranging scientist and philosopher and one of the great minds of the 19th century, by contrast, held that the theoretical entities of physics are, precisely, the things-in-themselves—a view which, though generally Empiricist, was thus clearly opposed to the Positivist doctrine. Theories and theoretical concepts, according to the Positivist understanding, were merely instruments of prediction. From one set of observable data, theories formed a bridge over which the investigator could pass to another set of observable data. Positivists generally maintained that theories might come and go, whereas the facts of observation and their empirical regularities constituted a firm ground from which scientific reasoning could start and to which it must always return in order to test its validity. In consequence, most Positivists were reluctant to call theories true or false but preferred to consider them merely as more or less useful.

The task of the sciences, as it earlier had been expressed by the German physicist Gustav Kirchhoff, was the pursuit of a compendious and parsimonious description of observable phenomena. Concern with first causes or final reasons was to be excluded from the scientific endeavour as fruitless, or hopeless (if not meaningless). Even the notion of explanation became suspect and was at best taken (as already in Comte) to be no more than an ordering and connecting of observable facts and events by empirically ascertainable laws.

Anti-atomism of Mach and Ostwald

Mach and, along with him, Wilhelm Ostwald, the originator of physical chemistry, were the most prominent opponents of the atomic theory in physics and chemistry. Ostwald even attempted to derive the basic chemical laws of constant and multiple proportions without the help of the atomic hypothesis. To the Positivist the atom, since it could not be seen, was to be considered at best a "convenient fiction" and at worst an illegitimate ad hoc hypothesis. Hans Vaihinger, a subjectivist who called himself an "idealistic Positivist," pursued the idea of useful fictions to the limit, and was convinced that the concept of the atom, along with the mathematical concepts of the infinite and the infinitesimal, and those of causation, free will, the economic man, and the like, were altogether fictitious, some of them even containing internal contradictions.

The anti-atomistic strand in the thought of the Positivists was an extreme manifestation of their phobia regarding anything unobservable. With the undeniably great success of the advancing microtheories in physics and chemistry, however, the Positivist ideology was severely criticized, not only by some contemporary philosophers but also by outstanding scientists. The Austrian Ludwig Boltzmann and the German Max Planck, for example, both top-ranking theoretical physicists, were in the forefront of the attack against Mach and Ostwald. Boltzmann and Planck, outspoken Realists, were deeply convinced of the reality of unobservable microparticles, or microevents, and were clearly impressed with the ever-growing and converging evidence for the existence of atoms, molecules, quanta, and subatomic particles. Nevertheless, the basic Positivist attitude was tenaciously held by many scientists, and striking parallels to it appeared in American Pragmatism and instrumentalism; in parts of the work of the Pragmatists Charles Sanders Peirce, William James, and John Dewey, there is a philosophy of pure experience essentially similar to that of Mach.

The Positivism of Avenarius

Though Richard Avenarius has not become widely known, he too anticipated a good deal of what the American Pragmatists propounded. His Positivism, like that of Mach, comprised a biologically oriented theory of knowl-

edge. From the needs of organisms in their adaptation to the exigencies of their environment develop the conceptual tools needed for prediction of future conditions. In Avenarius' view, the raw material of the construction of the concepts of common sense and of the sciences, however, was "the given"; *i.e.*, the data of immediate sensory experience. Just as Mill in the 19th century considered ordinary physical objects as "permanent possibilities of sensation," so Mach and Avenarius construed the concepts pertaining to what men commonsensically regard as the objects of the real world as "complexes of sensations." Thus, it was maintained that a stone, for example, is no more than a collection of such sensory qualities as hardness, colour, and mass. The traditional assumption that there must be an underlying substance that has these properties was repudiated. To the question "What would be left over if all of the perceptible qualities were stripped (in thought) away from an observable object?" these Positivists answered: "Precisely nothing." Thus the concept of substance was declared not only superfluous but meaningless as well.

In similar fashion, the concept of causation was explicated not as a real operating principle but as regularity of succession or as functional dependency among observable or measurable variables. Because these dependencies are not logically necessary, they are contingent and ascertained by observation, and especially by experimentation and inductive generalization.

The Newtonian doctrine according to which space and time are absolute or substantive realities had been incisively criticized by the 17th-century Rationalist Gottfried Leibniz and was subjected by Mach to even more searching scrutiny. While Leibniz had already paved the way for the conception of space and time as exclusively a matter of relations between events, Mach went still further in attacking the arguments of Newton in favour of a dynamic and absolute space and time. In particular, the inertial and centrifugal forces that arise in connection with accelerated or curvilinear motions had been interpreted by Newton as effects of such motions with respect to a privileged reference medium imagined as an absolute Cartesian mesh system graphed upon a real space. In a typically Positivistic manner, however, Mach found the idea quite incredible. How, he asked, could an absolutely empty space have such powerful effects? Mach conjectured that any privileged reference system must be generated not by an imperceptible grid but by material reality—specifically, by the total mass of the universe (galaxies and fixed stars), an idea that later served as an important starting point for Einstein's general theory of relativity and gravitation.

Influence of Mach and Avenarius

The Positivist theory of knowledge, as proposed by Mach and Avenarius, impressed many scholars, most notable among whom was probably the leading British logician and philosopher Bertrand Russell in one of the earlier phases of his thought. In a work entitled *Our Knowledge of the External World* (1914), Russell analyzed the concept of physical objects as comprising classes of (perceptual) aspects or perspectives, an idea that later stimulated the work of Rudolf Carnap, an outstanding philosophical semanticist and Analyst, entitled *Der logische Aufbau der Welt* (1928; *The Logical Structure of the World*, 1967). Mach remained the most influential thinker among Positivists for a long time, though some of his disciples, like Josef Petzoldt, are now largely forgotten. But *The Grammar of Science* (1892), written by Karl Pearson, a scientist, statistician, and philosopher of science, still receives some attention; and in France it was Abel Rey, also a philosopher of science, who, along the lines of Mach, severely criticized the traditional mechanistic view of nature. In the United States, John Bernard Stallo, a German-born American philosopher of science (also an educator, jurist, and statesman), developed a Positivistic outlook, especially in the philosophy of physics, in his book *The Concepts and Theories of Modern Physics* (1882), in which he anticipated to a degree some of the general ideas later formulated in the theory of relativity and quantum mechanics.

LOGICAL POSITIVISM AND LOGICAL EMPIRICISM

A first generation of 20th-century Viennese Positivists began its activities, strongly influenced by Mach, around 1907. Notable among them were a physicist, Philipp Frank, mathematicians Hans Hahn and Richard von Mises, and an economist and sociologist, Otto Neurath. This small group was also active during the 1920s in the Vienna Circle of Logical Positivists, a seminal discussion group of gifted scientists and philosophers that met regularly in Vienna, and in the related Berlin Society for Empirical Philosophy.

The Vienna Circle and the Berlin Society

These two schools of thought, destined to develop into an almost worldwide and controversial movement, were built on the Empiricism of Hume, on the Positivism of Comte, and on the philosophy of science of Mach. Equally important influences came from several eminent figures who were at the same time scientists, mathematicians, and philosophers—G.F. Bernhard Riemann, the author of a non-Euclidean geometry; Hermann von Helmholtz, a pioneer in a broad range of scientific studies; Heinrich Hertz, the first to produce electromagnetic waves in his laboratory; Ludwig Boltzmann, a researcher in statistical mechanics; Henri Poincaré, equally eminent in mathematics and philosophy of science; and David Hilbert, distinguished for his formalizing of mathematics. Most significant, however, was the impact of Albert Einstein, as well as that of the three great mathematical logicians of the past 100 years—the ground-breaking German Gottlob Frege and the authors of the monumental *Principia Mathematica* (1910–13), Bertrand Russell and Alfred North Whitehead.

The earlier Positivism of Viennese heritage. The confluence of ideas from these sources and the impressions that they made upon the Vienna and Berlin groups in the 1920s gave rise to the philosophical outlook of Logical Positivism—a label supplied in 1931 by A.E. Blumberg and the Minnesota philosopher of science Herbert Feigl. The leader of the Vienna Circle between 1924 and 1936 was Moritz Schlick, who in 1922 succeeded to the chair (previously held by Mach and Boltzmann) for the philosophy of the inductive sciences at the University of Vienna. By 1924 an evening discussion group had been formed with Schlick, Hans Hahn, Otto Neurath, Victor Kraft, Kurt Reidemeister, and Felix Kaufmann as the prominent active participants. The most important addition to the circle was Rudolf Carnap, who joined the group in 1926. One of the early activities was the study and critical discussion of the *Tractatus Logico-Philosophicus* (1922) of Ludwig Wittgenstein, a seminal thinker in Analytical and Linguistic philosophy. It seemed at the time that the views of Carnap and Wittgenstein, although they had been formulated and elaborated quite differently, shared a large measure of basic agreement. Parallel, but not completely independent, developments occurred in the Berlin group, in which Hans Reichenbach, Richard von Mises, Kurt Grelling, and Walter Dubislav were the leading spirits.

Scientific, antimetaphysical orientation

Both the Vienna and Berlin groups consisted mainly of philosophically interested scientists or scientifically trained and oriented philosophers. Schlick had already anticipated some of the basic epistemological tenets of the groups in his *Allgemeine Erkenntnislehre* (1918; "General Theory of Knowledge"). But the philosophical outlook was sharpened and deepened when, in the late 1920s, the Viennese Positivists published a pamphlet *Wissenschaftliche Weltauffassung: Der Wiener Kreis* (1929; "Scientific Conception of the World: The Vienna Circle"), which was to be their declaration of independence from traditional philosophy—and, in the minds of its authors (Carnap, Hahn, and Neurath, aided by Waismann and Feigl), a "philosophy to end all philosophies."

Language and the clarification of meaning. The basic ideas of Logical Positivism were roughly as follows: the genuine task of philosophy is to clarify the meanings of basic concepts and assertions (especially those of science)—and not to attempt to answer unanswerable questions such as those regarding the nature of ultimate reality or of the Absolute. Inasmuch as an extremely ambitious Hegelian type of metaphysics, Idealistic and absolutist in orientation, was still prevalent in the German-speaking countries, there were many who believed that the antidote was urgently needed. Moreover, the Logical Positivists also had

only contempt and ridicule for the ideas of the German Existentialist Martin Heidegger, whose interminable torment regarding such questions as "Why is there anything at all?" and "Why is what there is, the way it is?" and especially his pronouncements about Nothingness seemed to them to be not only sterile but so confused as to be nonsensical. The Logical Positivists viewed metaphysics as a hopelessly futile way of trying to do what great art, and especially poetry and music, already do so effectively and successfully. These activities, they held, are expressions of visions, feelings, and emotions and, as such, are perfectly legitimate as long as they make no claims to genuine cognition or representation of reality. What Logical Positivism recommended positively, on the other hand, was a logic and methodology of the basic assumptions and of the validation procedures of knowledge and of evaluation.

An adequate understanding of the functions of language and of the various types of meaning is another of the fundamentally important contributions of the Logical Positivists. Communication and language serve many diverse purposes: one is the representation of facts, or of the regularities in nature and society; another is the conveying of imagery, the expression and arousal of emotions; a third is the triggering, guidance, or modification of actions. Thus, they distinguished cognitive–factual meaning from expressive and evocative (or emotive) significance in words and sentences. It was granted that, in most utterances of everyday life (and even of science), these two types of meaning are combined or fused. What the Logical Positivists insisted upon, however, was that the emotive type of expression and appeal should not be mistaken for one having genuinely cognitive meanings. In such expressions as moral imperatives, admonitions, and exhortations there is, of course, a factually significant core; viz., regarding the (likely) consequences of various actions. But the normative element—expressed by such words as "ought," "should," "right," and their negations (as in "Thou shalt not. . . .")—is by itself not cognitively meaningful but has primarily emotional and motivative significance.

Cognitive versus emotive meaning

Early statements about moral value judgments, such as those by Carnap or by A.J. Ayer, a more radical British Positivist, seemed shocking to many philosophers, to whom it seemed that, in their careless formulation, moral norms were to be treated like expressions of taste. Equally shocking was their condemnation as nonsense (really non-sense; *i.e.,* complete absence of factual meaning) of all moral, aesthetic, and metaphysical assertions. More adequate and delicate analyses, such as that of the American Positivist Charles Stevenson, were soon to correct and modify those extremes. By proper allocation of the cognitive and the normative (motivative) components of value statements, many thinkers rendered the originally harsh and implausible Positivist view of value judgments more acceptable. Nevertheless, there is—in every Positivistic view—an ineluctable element of basic, noncognitive commitment in the acceptance of moral, or even of aesthetic, norms.

The verifiability criterion of meaning and its offshoots. The most noteworthy, and also most controversial, contribution of the Logical Positivists was the so-called verifiability criterion of factual meaningfulness. In its original form, this criterion had much in common with the earlier Pragmatist analysis of meaning (as in Charles Sanders Peirce and William James). Schlick's rather careless formulation: "The meaning of a [declarative sentence] is the method of its verification," which was really intended only to exclude from the realm of the cognitively meaningful those sentences for which it is logically inconceivable that either supporting or refuting evidence can be found, was close to the Pragmatist and, later, to the operationalist slogan that may be paraphrased by: "A difference must make a difference in order to be a difference"—or (more fully explicated): only if there is a difference in principle, open to test by observation, between the affirmation and the denial of a given assertion does this assertion have factual meaning. To take the classical example from Hume's analysis of the concept of causation, there is no difference between saying "*A* is always followed by *B*" and saying "*A* is necessarily always followed by *B*." That all *effects* have causes is true by virtue of the (customary) definitions of "cause" and "effect"; it is a purely formal or logical truth. But to say (instead of speaking of effects) that all *events* have causes is to say something factual—and conceivably false. (It should be noted that these rather crude uses of "causality" and "necessity" were later replaced by much more subtle analyses.)

Untestability of causation, ether, simultaneity

One of the most important examples that stimulated the formulation of the meaning criterion was Einstein's abandonment, in 1905, of the ether hypothesis and of the notion of absolute simultaneity. The hypothesis that there exists a universal ether, as a medium for the propagation of light (and of electromagnetic waves generally), had been quite plausible and was widely accepted by physicists during the second half of the 19th century. To be sure, there were a number of serious difficulties with the idea: the properties that had to be ascribed to the ether were difficult to conceive in a logically compatible manner; and the ether hypothesis in the last stage of its development (by the Dutch physicist Hendrik Lorentz and the Irish physicist George Fitzgerald) had become objectionable in that it sought to provide excuses for the absolute unobservability of that mysterious, allegedly all-pervasive, universal substance. Similarly, it had become impossible, except at the price of intolerably ad hoc hypotheses, to maintain the notions of absolute time and of absolute simultaneity. Thus Einstein, by eliminating these empirically untestable assumptions, was led to his special theory of relativity.

Modifications of the meaning criterion

Several important changes in the formulation of the meaning criterion took place in the ensuing decades from 1930 to 1960. The original version formulated in terms of verifiability was replaced by a more tolerant version expressed in terms of testability or confirmability. Obviously, universal propositions, such as "All cats have claws," being only partially supportable by positive instances (one cannot examine every cat that exists), are not conclusively verifiable. Nevertheless, scientists do accept lawlike statements on the basis of only incomplete, as well as indirect, verification—which is what "confirmation" amounts to. It was in coming to this juncture in his critique of Positivism that Karl Popper, an Austro-English philosopher of science, in his *Logik der Forschung* (1935; *The Logic of Scientific Discovery,* 1959), insisted that the meaning criterion should be abandoned and replaced by a criterion of demarcation between empirical (scientific) and transempirical (nonscientific, metaphysical) questions and answers—a criterion that, according to Popper, is to be testability, or, in his own version, falsifiability; *i.e.,* refutability. Popper was impressed by how easy it is to supposedly verify all sorts of assertions—those of psychoanalytic theories seemed to him to be abhorrent examples. But the decisive feature, as Popper saw it, should be whether it is in principle conceivable that evidence could be cited that would refute (or disconfirm) a given law, hypothesis, or theory. Theories are (often) bold conjectures. It is true that scientists should be encouraged in their construction of theories—no matter how far they deviate from the tradition; it is also true, however, that all such conjectures should be subjected to the most severe and searching criticism and experimental scrutiny of their truth claims. The growth of knowledge thus proceeds through the elimination of error; *i.e.,* through the refutation of hypotheses that are either logically inconsistent or entail empirically refuted consequences.

Despite valuable suggestions in Popper's philosophy of science, the Logical Positivists and Empiricists continued to reformulate their criteria of factual meaningfulness. The Positivist Hans Reichenbach, who emigrated from Germany to California, proposed, in his *Experience and Prediction* (1938), a probabilistic conception. If hypotheses, generalizations, and theories can be made more or less probable by whatever evidence is available, he argued, then they are factually meaningful. In another version of meaningfulness, first adumbrated by Schlick (under the influence of Wittgenstein's thought), the philosopher's attention is focussed on concepts rather than on propositions. If the concepts in terms of which theories are formulated can be related, through chains of definitions, to concepts that are definable ostensibly—*i.e.,* by pointing to or exhibiting items or aspects of direct experience—

then those theories are factually meaningful. This is the version also advocated by Richard von Mises, an Austro-American mathematician and methodologist, in his *Positivism* (1951) and, later, more technically elaborated by Carnap, in *Minnesota Studies in the Philosophy of Science,* vol. 1 (1956).

Other issues. These views of meaningfulness are essentially refinements of the doctrine of so-called protocol sentences, developed in the late 1920s and early 1930s and elaborated especially by Carnap, by Otto Neurath, a polymath sociologist and philosopher, and also (with some differences) by Schlick. Protocol sentences, originally conceived along the lines of an interpretation, developed in the Vienna Circle, of Wittgenstein's elementary propositions, were identified as those sentences that make statements about the data of direct experience. But Neurath—and independently also Popper—warned of the danger that this doctrine might lead to subjective Idealism and recommended that it be given a rational reconstruction on an intersubjective basis. Thus, Neurath and Carnap preferred that a physicalistic thing-language be employed as the starting point and testing ground of all knowledge claims. Propositions in this language would describe objectively existing, directly observable states of affairs or events. Because all objective and intersubjective knowledge was seen, in such a physicalism, to rest on statements representing things and their properties, relations and ongoing processes as they are found in unbiassed, and presumedly theory-free, observation, the physicalists were thus proclaiming a first thesis of the so-called Unity of Science principle. Though Mach had proceeded from the basis of (neutral) immediate experience, his insistence on the unity of all knowledge and all science was retained—at least in general spirit—by the later Positivists. In this view, all classifications of the sciences, or divisions of their subject matter, were seen as artificial, valuable at best only administratively, but without philosophical justification.

Physicalism and the Unity of Science

Sharply to be distinguished from this first thesis of the Unity of Science is a second that formulates a reductionism of a very different type: whereas the first thesis concerns the unity of the observational basis of all the sciences, the second proposes (tentatively) a unity of the explanatory principles of science. Reductions within physics itself, such as that of thermodynamics to the kinetic theory of heat (statistical mechanics) and of optics to electromagnetics; and, beyond that, the explanation of chemical phenomena, with the help of the quantum theory, in terms of atomic and molecular processes; and, furthermore, the progress that has been made in the physical explanation of biological phenomena (especially in the recent development of molecular biology)—all of these encourage the idea of a unitary set of physical premises from which the regularities of all of reality could be derived. But it must be admitted that, in contrast to the first thesis (which, by comparison is almost trivial), the second, being a bold conjecture about future reductions in the sciences, might well prove to be limited in the scope of its validity. The most controversial part of this reductionist ideology, however, concerns the realms of organic life and especially that of mind; it concerns, in other words, the reducibility of biology to physics and chemistry and of psychology to neurophysiology—and (though this is clearly utopian at present) of both ultimately to basic physics.

The most serious alternative to this reducibility thesis of the Unity of Science movement is the theory of emergent evolution, according to which life or mind (or both) are genuinely novel forms of reality that could not possibly have been derived from, or predicted by, any laws or theories of the lower or earlier levels of existence.

Historically, it may be plausible that the notorious perplexities of the traditional problem of how mind relates to body motivated both the phenomenalistic Positivists as well as the Behaviourists and physicalists. In either view, the mind–body problem conveniently disappears; it is branded as a metaphysical pseudoproblem. The phenomenalism of Mach and the early Russell was expressed in a position called neutral monism, according to which both psychological and physical concepts are viewed as logical constructions on the basis of a neutral set of data of immediate experience. There are thus not two realities—the mental and the physical; there are merely different ways of organizing the experiential data. In the Behaviourist–physicalist alternative, on the other hand, the philosopher, considering the concepts that are ordinarily taken to characterize private mental acts and processes, defines them on the basis of publicly (intersubjectively) observable features of the behaviour, including the linguistic behaviour, of man.

Neutral monism and mental states

The absolute privacy of mental events was first criticized, however, by Carnap and later by an Oxford Analytical philosopher, Gilbert Ryle; and Wittgenstein, in an argument against the very possibility of a private language, maintained that, unless men have objective criteria for the occurrence of mental states, they cannot even begin to communicate meaningfully with each other about their direct experiences. Wittgenstein thus repudiated the traditional view according to which a man's knowledge of other persons' minds must be based on analogical inference from his own case. In a similar vein, an American psychologist, B.F. Skinner, tried to account for man's acquisition of subjective terms in his language by a theory of verbal behaviour. A man learns to describe his mental states, explained Skinner, from the utterances of others who ascribe these states to him by virtue of their observation of his behaviour (*e.g.,* in the social context; or when a certain stimulus situation prevails in his environment).

Dispositions as conditionals

Both Carnap and Ryle have emphasized that many mental features or properties have a dispositional character. Dispositional terms, whether used in psychology or more broadly, have to be understood as shorthand expressions for test conditions—or test-result conditionals. Thus, even in ordinary life, a man appraises, for example, the intelligence of a person in the light of what he does, how he does it, how fast he does it, when confronted with various tasks or problems. Just as such physical properties as malleability, brittleness, or thermal or electrical conductivity must be defined in terms of what happens when certain conditions are imposed, so also mental dispositions are to be construed as similarly hypothetical; *i.e.,* as (in the simplest case) stimulus–response relationships.

The later Positivism of Logical Empiricism. Logical Positivism, essentially the doctrine of the Vienna Circle (*c.* 1924–38), underwent a number of important changes and innovations in the middle third of the century, which suggested the need for a new name. The designation Positivism had been strongly connected with the Comte–Mach tradition of instrumentalism and phenomenalism. The emphasis that this tradition had placed, however, on the positive facts of observation and their negative attitude toward the atomic theory and the existence of theoretical entities in general were no longer in keeping with the spirit of 20th-century science. Nevertheless, the requirement that hypotheses and theories be empirically testable, though it became more flexible and tolerant, could not be relinquished. It was natural, then, that the word "empiricism" should occur in any new name. Accordingly, retaining the term "logical" in roughly its same earlier meaning, the new name "Logical Empiricism" was coined.

The status of the formal and a priori. The intention of the word "logical" was to insist on the distinctive nature of logical and mathematical truth. In opposition to Mill's view, according to which even logic and pure mathematics are empirical (*i.e.,* are justifiable or refutable by observation), the Logical Positivists—essentially following Frege and Russell—had already declared mathematics to be true only by virtue of postulates and definitions. Expressed in the traditional terms used by Kant, logic and mathematics were recognized as a priori disciplines (valid independently of experience) precisely because their denial would amount to a self-contradiction, and statements within these disciplines were expressed in what Kant called analytic propositions; *i.e.,* propositions that are true or false only by virtue of the meanings of the terms they contain. In his own way, the broad-ranging scholar Gottfried Leibniz had seen this in the 17th century long before Kant. The truth of such a simple arithmetical proposition as, for example, "2 + 3 = 5" is necessary, universal, a priori, and analytic because of the very meaning of "2," "+,"

The analytic and synthetic in mathematics

"3," "5," and "=." Experience could not possibly refute such truths because their validity is established (as Hume said) merely by the "relation of ideas." Even if—"miraculously"—putting two and three objects together should on some occasion yield six objects, this would be a fascinating feature of those objects (rabbits, perhaps); but it would not in the least tend to refute the purely definitional truths of arithmetic.

The case of geometry is altogether different. Geometry can be either an empirical science of natural space or an abstract system with uninterpreted basic concepts and uninterpreted postulates. The latter is the conception introduced in rigorous fashion by a late 19th-century mathematician, David Hilbert, and, still later, by a United States geometer, Oswald Veblen. In the axiomatizations that they developed, the basic concepts, called primitives, are implicitly defined by the postulates: thus, such concepts as point, straight line, intersection, betweenness, and plane are related to each other in a merely formal manner. The proof of theorems from postulates, and with explicit definitions of derived concepts (such as of triangle, polygon, circle, or conic section), is achieved by strict deductive inference. Very different, however, is geometry as understood in practical life, and in the natural sciences and technologies, in which it constitutes the science of space. Ever since the development of the non-Euclidean geometries in the first half of the 19th century, it has no longer been taken for granted that Euclidean geometry is the only geometry uniquely applicable to the spatial order of physical objects or events. In Einstein's general theory of relativity and gravitation, in fact, a four-dimensional Riemannian geometry with variable curvature was successfully employed, an event that amounted to a final refutation of the Kantian contention that the truths of geometry are "synthetic a priori." With respect to the relation of postulates to theorems, geometry is thus analytic, like any other rigorously deductive discipline; but the postulates themselves, when interpreted—*i.e.,* when construed as statements about the structure of physical space—are indeed synthetic, but also a posteriori; *i.e.,* their adequacy depends upon the results of observation and measurement.

Developments in Linguistic Analysis and their offshoots. Important contributions, beginning in the early 1930s, were made by Carnap, by Kurt Gödel, a Moravian-American mathematical logician, and others to the logical analysis of language. Charles Morris, a Pragmatist concerned with Linguistic Analysis, had outlined the three dimensions of semiotics (the general study of signs and symbolisms): syntax, semantics, and pragmatics (the relation of signs to their users and to the conditions of their use). Syntactical studies, concerned with the formation and transformation rules of language (*i.e.,* its purely structural features), soon required supplementation by semantical studies, concerned with rules of designation and of truth. Semantics, in the strictly formalized sense, owed its origin to Alfred Tarski, a leading member of the Polish school of logicians, and was then developed by Carnap and applied to problems of meaning and necessity. As Wittgenstein had already shown, the necessary truth of tautologies simply amounts to their being true under all conceivable circumstances. Thus the so-called eternal verity of the principles of identity (p is equivalent with itself), of noncontradiction (one cannot both assert and deny the same proposition), and the principle of excluded middle (any given proposition is either true or false; there is no further possibility) is an obvious consequence of the rules according to which the philosopher uses (or decides to use) the words "proposition," "negation," "equivalence," "conjunction," "disjunction," and others. Quite generally, questions regarding the meanings of words or symbols are answered most illuminatingly by stating the syntactical and the semantical rules according to which they are used.

Ordinary versus artificial language

Two different schools of thought originated from this basic insight: (1) the philosophy of "ordinary language" Analysis—initiated by Wittgenstein, especially in his later work, and (following him) developed in differing directions in the works of Gilbert Ryle and John Langshaw Austin, both Oxford philosophers, of the Cambridge Analyst John Wisdom, and others; and (2) the ideology, essentially that of Carnap, usually designated as logical reconstruction, which builds up an artificial language. In the procedures of "ordinary language" Analysis, an attempt is made to trace the ways in which a person commonly expresses himself. In this manner, many of the traditional vexatious philosophical puzzles and perplexities are shown to arise out of deviant uses of language. (Lewis Carroll had already anticipated some of these oddities in his whimsical manner in *Alice in Wonderland.*) The much more rigorous procedures of the second school—of Tarski, Carnap, and many other logicians—rest upon the obvious distinction between the language (and all of its various symbols) that is the object of analysis, called the object language, and that in which the analysis is formulated, called the metalanguage. If needed and fruitful, this process can be repeated, in that the erstwhile metalanguage can become the object of a metametalanguage, and so on—without the danger of a vicious infinite regress.

With the help of semantical concepts, an old perplexity in the theory of knowledge can then be resolved. Positivists have often tended to conflate the truth conditions of a statement with its confirming evidence, a procedure which has led to certain absurdities committed by phenomenalists and operationalists, such as the pronouncement that the meanings of statements about past events consist in their (forthcoming future) evidence. Clearly, the objects—the targets or referents—of such statements are the past events. Thus the meaning of a historical statement is its truth conditions; *i.e.,* the situation that would have to obtain if the historical statement is to be true. The confirmatory evidence, however, may be discovered either in the present or in the future. Similarly, the evidence for an existential hypothesis in the sciences may consist, for example, in cloud-chamber tracks, spectral lines, or the like, whereas the truth conditions may relate to subatomic processes or to astrophysical facts. Or, to take an example from depth psychology, the occurrences of unconscious wishes or conflicts are the truth conditions for which the observable symptoms (Freudian lapses, manifest dream contents, and the like) serve merely as indicators or clues; *i.e.,* as items of confirming evidence.

The third dimension of language (in Morris' view of semiotic)—*i.e.,* the pragmatic aspect—has not been as fully and formally analyzed as the other two dimensions. There is even some disagreement as to whether some of the cognitive activities (such as verifying, refuting, or interpreting), in addition to the noncognitive functions (such as the motivative and persuasive appeals), are to be included in studies of pragmatics.

Undecidability; probability

One of the most surprising and revolutionary offshoots of the metalinguistic (formal) analyses was Kurt Gödel's discovery, in 1931, of an exact proof of the undecidability of certain types of mathematical problems, a discovery that dealt a severe blow to the expectations of the Formalistic school of mathematics championed by Hilbert and his collaborator, Paul Bernays. Before Gödel's discovery, it had seemed plausible that a mathematical system could be complete in the sense that any well-formed formula of the system could be either proved or disproved on the basis of the given set of postulates. But Gödel showed rigorously (what had been only a conjecture on the part of the Dutch Intuitionist L.E.J. Brouwer and his followers) that, for a large class of important mathematical systems, such completeness cannot be achieved.

Both Carnap and Reichenbach, in their very different ways, made extensive contributions to the theory of probability and induction. Impressed with the need for an interpretation of the concept of probability that was thoroughly empirical, Reichenbach elaborated a view that conceived probability as a limit of relative frequency and buttressed it with a pragmatic justification of inductive inference. Carnap granted the importance of this concept (especially in modern physical theories) but attempted, in increasingly refined and often revised forms, to define a concept of degree-of-confirmation that was purely logical. Statements ascribing an inductive probability to a hypothesis are, in Carnap's view, analytic, because they merely formulate the strength of the support bestowed upon a hypothesis by a given body of evidence.

CURRENT CRITICISMS AND CONTROVERSIES

Logical Positivism and Logical Empiricism have from their very beginnings been subjected to searching criticisms. At first it was the verifiability criterion of meaningfulness that produced a storm of oppositions. One group of critics asked whether the criterion was meaningful in the light of its own standard. Carnap replied that the criterion itself was not intended as a factual assertion but rather as a proposal for a better and clearer use of language. Nevertheless, the Logical Empiricists felt that the (tolerant–liberal) formulation of the meaning criterion—far from being an arbitrary injunction—came rather close to what enlightened common sense, and especially the scientific attitude, intended by the difference between sense and non-sense.

Scientific explanation

Other recent and current criticisms concern the Logical Empiricist views regarding the nature of scientific explanation, in regard to which matters are not as simple as they were originally conceived to be. Closer attention to the history of scientific theories has revealed important discontinuities, or revolutions, in the conceptual schemes of the sciences. The significant role of statistical (or probabilistic) explanations in most modern sciences, for example, is receiving increasingly sophisticated analyses.

Nevertheless, the prevalence of scientific revolutions and anarchism or subjectivism in scientific method has been exaggerated in differing ways by several scholars. As has been conceded by all competent philosophers of science and even by the greatest scientist-philosophers of the century—Albert Einstein, Niels Bohr, Erwin Schrödinger, and others—there is no straight logical path, no standard recipe, by which to move from the data of observation and arrive at scientific theories. It may also be admitted that, though scientific creativity has psychologically much in common with artistic creation, the criteria of appraisal are certainly quite different. And, although, from the present critical point of view, all and any scientific assertions are in principle subject to revision, it is nonetheless felt to be grotesque to deny the relative stability of the empirical laws that serve as the testing ground of alternative theories.

(H.Fe.)

Pragmatism

In its broadest and most familiar sense, "pragmatism" refers to the usefulness, workability, and practicality of ideas, policies, and proposals as criteria of their merit and claims to attention. Achieving results, "getting things done" in business and public affairs is often said to be "pragmatic." There is a harsher and more brutal connotation of the term in which any exercise of power in the successful pursuit of practical and specific objectives is called "pragmatic." The character of American business and politics is often so described. In these cases "pragmatic" carries the stamp of justification: a policy is justified pragmatically if it is successful. The familiar and the academic conceptions have in common an opposition to invoking the authority of precedents or of abstract and ultimate principles. Thus in law, judicial decisions that have turned on the weighing of consequences and probable general welfare rather than on being deduced from precedents have been called pragmatic.

The word pragmatism goes back to the Greek πραγμα ("action," "affair"). The Greek historian Polybius (died 118 BC) called his writings "pragmatic," meaning thereby that they were intended to be instructive and useful to his readers. In his introduction to the *Philosophy of History,* Hegel commented on this "pragmatical" approach as the second kind of reflective history, and for this genre he cited Johannes von Müller's *History of the World* (Eng. trans. 1840). As the psychologist and leading Pragmatist William James remarked, "the term is derived from the same Greek word πραγμα meaning action, from which our words 'practice' and 'practical' come." Charles Peirce, another pioneering Pragmatist, who may have been the first to use the word to designate a specific philosophic doctrine, had Kant's German term rather than the Greek word in mind: *Pragmatisch* refers to experimental, empirical, and purposive thought "based on and applying to experience." In the philosophy of education the notion that children learn by doing, that critical standards of procedure and understanding emerge from the application of concepts to directly experienced subject matters, has been called "pragmatic." In semiotics, the general theory of language, that part that studies the relation of the user to the words or other signs that he uses is called pragmatics (as distinct from semantics and syntax).

MAJOR THESES OF PHILOSOPHIC PRAGMATISM

During the first quarter of the 20th century, Pragmatism was the most influential philosophy in America, exerting an impact on the study of law, education, political and social theory, art, and religion. Six fundamental theses of this philosophy can be distinguished. It is, however, unlikely that any one thinker would have subscribed to them all; and even on points of agreement, varying interpretations mark the thought and temper of the major Pragmatists. The six theses are:

1. Responsive to Idealism and evolutionary theory, Pragmatists have emphasized the "plastic" nature of reality and the practical function of knowledge as an instrument for adapting to reality and controlling it. Existence is fundamentally concerned with action, which some Pragmatists exalted to an almost metaphysical level. Change being an inevitable condition of life, Pragmatists have called attention to the ways in which change can be directed for individual and social benefit. They have consequently been most critical of moral and metaphysical doctrines in which change and action are relegated to the "merely practical," on the lowest level of the hierarchy of values. Some Pragmatists anticipated the more concrete and life-centred philosophy of Existentialism by arguing that only in acting—confronted with obstacles, compelled to make choices, and concerned to give form to experience—is man's being realized and discovered.

2. Pragmatism is a continuation of critical Empiricism in emphasizing the priority of actual experience over fixed principles and a priori reasoning in critical investigation. For James this meant that the Pragmatist

> turns away from abstraction and insufficiency, from verbal solutions, from bad *a priori* reasons, from fixed principles, closed systems, and pretended absolutes and origins. He turns towards concreteness and adequacy, towards facts, towards action . . . . It means the open air and possibilities of nature, as against . . . dogma, artificiality, and the pretence of finality in truth.

3. The pragmatic meaning of an idea, belief, or proposition is said to reside in the distinct class of specific experimental or practical consequences that result from the use, application, or entertainment of the notion. As Peirce commented: "Our idea of anything is our idea of its sensible effects." Two propositions for which no different effects can be discerned have merely a verbal appearance of dissimilarity, and a proposition for which no definite theoretical or practical consequences can be determined is pragmatically meaningless. For Pragmatists "there is no distinction of meaning so fine as to consist in anything but a possible difference of practice." Meaning thus has a predictive component, and some Pragmatists came close to identifying the meaning of a term or proposition with the process of its verification.

4. While most philosophers have defined truth in terms of a belief's "coherence" within a pattern of other beliefs or as the "correspondence" between a proposition and an actual state of affairs, Pragmatism has, in contrast, generally held that truth, like meaning, is to be found in the process of verification. Thus truth *is* the verification of a proposition, or the successful working of an idea. Crudely, truth is "what works." Less crudely and more theoretically, truth is in Peirce's words, the "limit towards which endless investigation would tend to bring scientific belief." For John Dewey, founder of the "Instrumentalist" school of Pragmatism, these are beliefs "warranted" by inquiry.

5. In keeping with their understanding of meaning and truth, Pragmatists have interpreted ideas as instruments and plans of action. In contrast to the conception of ideas as images and copies of impressions or of external objects, Pragmatist theories have emphasized the functional character of ideas: ideas are suggestions and anticipations of

possible conduct; they are hypotheses or forecasts of what will result from a given action; they are ways of organizing behaviour in the world rather than replicas of the world. Ideas are thus analogous in some respects to tools; they are efficient, useful, and valuable, or not, depending on the role that they play in contributing to the successful direction of behaviour.

6. In methodology, Pragmatism is a broad philosophical attitude toward the formation of concepts, hypotheses, and theories and their justification. To Pragmatists man's interpretations of reality are motivated and justified by considerations of efficacy and utility in serving his interests and needs; the molding of language and theorizing are likewise subject to the critical objective of maximum usefulness according to man's various purposes.

HISTORY OF PRAGMATISM

Idealism, Empiricism, Kantianism

Antecedents in modern philosophy. Pragmatism was a part of a general revolt against the overly intellectual, somewhat fastidious, and closed systems of Idealism in 19th-century philosophy. These boldly speculative Idealists had expanded man's subjective experience of mind till it became a metaphysical principle of cosmic explanation. To the Idealist, all of reality was one fabric, woven from parts that cohered by virtue of the internal relations that they bore to one another; and this reality was often interpreted in abstract and fixed intellectual categories. The theory of evolution, then still new, seemed to the Pragmatists, on the other hand, to call for a new, non-Idealist interpretation of nature, life, and reason—one that challenged the long-established conceptions of fixed species. The new emphasis was on the particular variations and struggles of life in adapting to the environment. Philosophically, the fact of growth and the development of techniques for instituting changes favourable to life became the significant factors rather than the Idealist's ambitious rationalistic account of human goals and of the universe in general, and important developments in natural science and logic also encouraged a critical attitude toward earlier systems.

There were two main influences on the early formation of Pragmatism: One was the tradition of British Empiricism in the work of John Stuart Mill, Alexander Bain, and John Venn, which had stressed the role of experience in the genesis of knowledge—and particularly their analyses of belief as being intimately tied in with action and, indeed, as definable in terms of one's disposition and motive to act. The work of George Berkeley, an important 18th-century empirical Idealist, which presented a theory of the practical and inferential nature of knowledge, of sensations as signs (and thus predictive) of future experience, led Peirce to refer to him as "the introducer of Pragmatism." The other major influence came from modern German philosophy: from Kant's analysis of the purposive character of belief and of the roles of will and desire in forming belief and his doctrine of "regulative ideas," such as God or the Soul, which guide the understanding in achieving systematic completeness and unity of knowledge; from Romantic Idealists, for whom all reason is "practical" in expanding and enriching human experience; and from Hegel's historical and social conception of changing and developing subject matters. In sum, Peirce was profoundly impressed by Kant and by the Scottish philosophy of common sense, James by British Empiricism and by the voluntarisms (stressing the role of choice or will) of the genetic epistemologist James Ward and the relativistic French Personalist Charles Renouvier, Dewey by Coleridge's version of Kant's active conception of mind and by neo-Kantian and Hegelian Idealism.

Finally, to these influences must be added that of American social experience in the 19th century: the rapid expansion of industry and trade and a popular optimism, with its roots in Puritan theology, holding that hard work and virtue are bound to be rewarded. Both the precariousness of frontier life, however, and the rapidly expanding economy weakened the prevailing Calvinistic belief in a predestined future and encouraged the emergence of inventiveness, a sense of living still in the New World experiment, and adoption of the ideal of "making good."

The Metaphysical Club. It was in the critical group discussions of the "Metaphysical Club" in the 1870s in Cambridge, Massachusetts, that Pragmatism first received philosophic expression. In addition to Peirce and James, membership in the club included Chauncey Wright, F.E. Abbot, and Justice Oliver Wendell Holmes, Jr. A version of Peirce's now classic paper "The Fixation of Belief" (November 1877) seems to have been presented at the club. But James also published a paper in 1878, "Spencer's Definition of Mind as Correspondence," in which his Pragmatism and analysis of thought and belief are clearly discernible. It was in a lecture delivered 20 years later, however, that James introduced Pragmatism, then fully crediting the idea to Peirce. It was primarily James's exposition that became famous and was received by the world at large.

Peirce's Pragmatism

The classical Pragmatists. The Pragmatic philosophy of Charles Sanders Peirce was part of a more general theory of thought and of signs. Thought, or "inquiry," it was held, results from doubt, a state in which habitual actions are blocked or confused and from which organic irritation and irresolution result. Resolution, unobstructed conduct, on the other hand, are products of belief, which is a form of stability and satisfaction. It is the function of scientific thought to produce true beliefs. In a prolonged effort to embed this analysis of doubt and inquiry within a more comprehensive theory of signs in which communication, thought, knowledge, and intelligent conduct could be fully understood, Peirce achieved a wealth of original insights. A sign is a socially standardized way by which something (a thought, word, object) refers man (the community of sign users) to something else (the interpretant), which, in turn, is itself another sign. Peirce's Pragmatism is thus a method for translating certain kinds of signs into clearer signs in order to surmount linguistic or conceptual confusion. Getting at the interpretant involves determining the "effects" or consequences of the signs or ideas in question.

Peirce's Pragmatism is therefore primarily a theory of meaning that emerged from his first-hand reflections on his own scientific work, in which the experimentalist understands a proposition as meaning that, if a prescribed experiment is performed, a stated experience will result. The method has two different uses:

(1) It is a way of showing that when disputes permit no resolution, the difficulties are due to misuses of language, to subtle conceptual confusions. Such questions as whether the physical world is an illusion, whether man's senses always mislead him, or whether his actions are fated are "not real problems."

(2) The method may be employed for clarification. As Peirce wrote:

> Consider what effects, that might conceivably have practical bearings, we conceive the object of our conception to have. Then our conception of these effects is the whole of our conception of the object.

To say, for example, that an object *O* is hard means that if the operation of scratching *O* is performed, *O* will not be scratched by most substances. One thus achieves clarity when one can supply a conditional statement of this kind.

Similarly, in his theory of truth, one means by truth of belief that if a certain operation is the subject of continuous scientific inquiry by the community of investigators, assent to the belief would increase and dissent decrease "in the long run." Consequently, not only is thought purposive but meaning carries a reference to the future. Peirce's concept of the community of sign users and inquirers also has social and moral relevance, for it is nothing less than the ideal of rational democracy.

Witnessing his doctrine undergo a medley of dubious interpretations, Peirce eventually dissociated himself from these by calling his own view "Pragmaticism," a term he called so ugly as to be safe from uninformed use. Parts of the work of Dewey, of the social Pragmatist G.H. Mead, and of the conceptualistic Pragmatist C.I. Lewis are a further development of the logical Pragmatism of Peirce. The English logician F.P. Ramsey and the Italians Giovanni Vailati and Mario Calderoni also undertook significant extensions of Peirce's Pragmatism.

An alternative, though not wholly different, version of Pragmatism was developed by William James. It took a

psychological and moral form largely unforeseen and unintended by Peirce.

James's Pragmatism

A basic difference between Peirce and James is discernible in their respective conceptions of the direction to be taken by Pragmatic analysis. While Peirce construed meaning in general, conditional schema, and interpretants, James focussed upon the distinct contributions that ideas and beliefs make to specific forms of human experience on the living level of practical wants and purposes. Between the two close friends there persisted a fundamental philosophical difference in outlook that affected even their styles. While James was a Nominalist, holding that the full significance of ideas, meanings, and actions lies in their particular concrete existent occurrence, Peirce—as a scholastic Realist—sharply criticized him at this point, arguing that "a thing in general is as real as in the concrete."

The most conspicuous feature of James's writings on Pragmatism is the dominant place given to considerations of value, worth, and satisfaction—consequences of his teleological (purposive) conception of mind (*cf.* his *Principles of Psychology*). James maintained that thought is adaptive and purposive but also suffused with ideal emotional and practical interests—"should-be's"—which, as conditions of action, work to transform the world and create the future, even to "make the truth which they declare." Consequently, truth and meaning are species of value: "*The true is the name of whatever proves itself to be good in the way of belief.*"

James took meaning to be an intimate part of the use of ideas for expediting action. The notion of the difference that a proposition makes in experience was fundamental in James's Pragmatic methodology. He remarked that "it is a good rule in physiology, when we are studying the meaning of an organ," to look to the specific function that it performs. In like manner, the special difference that the presence of mind makes in observable cases, reflected in its unique functioning, defines the use of "mind"; "In particular, the pursuance of future ends and the choice of means . . . are thus the *mark and criterion of the presence of mentality.*"

With his training in medicine and psychology and the influence of Darwin in the background, James considered that the main function of thought is to help us establish "satisfactory relations with our surroundings." Thus man helps to mold the character of reality according to his needs and desires. Indeed, this is fundamental in James's defense of the right to believe in his famous essay "The Will to Believe" (1897). James argued that we may have a reasonable right to hold a religious or metaphysical belief (*e.g.,* that there is a perfect, eternal, and personal aspect of the universe) when the belief in question would supply a vital psychological and moral benefit to the believer, when evidence for and against the belief is equal, and when the decision to believe is forced and momentous. In James's functional conception of truth, the "working," and hence the truth, of ideas is their role in opening up valuable possible directions of thought and action—"a leading that is worth while."

James's "working" view of truth and of a reality that man in part makes by acting out and realizing ideas, and especially his essay "The Will to Believe," were enthusiastically received by F.C.S. Schiller in England and by Giovanni Papini in Italy, and these doctrines became a cause célèbre for Pragmatists and their critics.

Schiller's Pragmatism

An admirer and friend of James, Schiller, now nearly forgotten, was once the most famous Pragmatist in England and Europe. Schiller was initially a humanist in the sense that, for him, both reality and knowledge are reflections of human activity—"the taken" rather than "the given." He first came to appreciate James's "The Will to Believe" in 1897 and subsequently acknowledged its impact on his thinking in an early and important paper, "Axioms As Postulates" (1902). He was a tireless critic of the "closed" systems of Idealism of F.H. Bradley, J.M.E. McTaggart, and Bernard Bosanquet and an advocate of the intellectual freedom that consists in "open," plural, changing, and to some extent never finished philosophical theorizing. According to Schiller, reality and truth are "man-made" rather than eternal verities. The true and the false are basically forms of good and bad and are relative to the private purposes of some particular person. He attempted to describe and analyze the "logic" of the experimental "trying" through which such needs are satisfied. For Schiller, reality is wholly plastic; and, starting out from initial postulates, one proceeds to construct one's schemes for achieving a satisfactory outcome of desire, finally rendering one's unformed possibilities (*hylē*) into a common world of language and action. On this view, all of science derives from and is inescapably guided by the psychological process of human thought; man is the measure of all things.

Dewey's Instrumentalism

John Dewey once noted that "Peirce wrote as a logician and James as a humanist." This distinction characterizes not only the course of Pragmatism but also the shaping of Dewey's own thought. Dewey first felt the influence of James in the 1890s, during the period in which he was struggling to free himself from the hold of Hegelian Idealism. Later, he recognized the value of Peirce's work, which clearly prefigured certain ideas that he had developed independently.

With indefatigable effort and care Dewey reformulated Pragmatism, critically readjusting some of its conflicting doctrines, drawing upon his own work in psychology and education, and finding stimulation in the social Pragmatism of his friend George Herbert Mead. The resulting construction was Instrumentalism, which Dewey conceived as a single coherent theory embracing both the logical and humanistic currents of Pragmatism and thus integrating the methods and conclusions of scientific knowledge with beliefs about values and purposes.

While scientific, moral, and social experiences may differ in subject matter, the method of thought functioning "in the experimental determinations of future consequences" remains the same for all inquiry. Initially provoked by doubtful or problematic conditions, intelligent conduct is addressed to a resolution and settling of these conditions and to a "warranted assertion"—Dewey's version of "truth." Such is the "mediative function" of reason. "Truth" is thus identified with the outcome of competent inquiry. Actions occurring on the organic level, if they be at first confused and obstructed, can then become organized, coherent, and liberated through such inquiry.

Dewey's analysis of the organic, cultural, and formal conditions of intelligent action implies that all reflective conduct issues in an evaluation of a situation with respect to future action and consequences: thus inquiry is essentially an evaluative procedure. This method, most impressively applied in the sciences, is nonetheless a paradigm of moral activity as well. In ethics, "the action needed to satisfy" the situation is not to be found simply by the application of moral codes. The meaning

> has to be searched for [since] there are conflicting desires and alternative apparent goods . . . . Hence inquiry is exacted . . . . The good of the situation has to be discovered, projected and attained on the basis of the exact defect and trouble to be rectified.

In general, for Instrumentalism, moral ideals and "ends" function as means and hypotheses in guiding the deliberative process directed to controlling experience and attaining future goods.

> Not health as an end fixed once for all, but the needed improvement in health—a continual process—is the end and good . . . . Not perfection as a final goal, but the ever-enduring process of perfecting, maturing, refining is the aim of living . . . , Growth itself is the only moral 'end.'

Inquiry possessed a genuine religious significance for Dewey, and in its functioning as a critical, self-corrective social process of human growth he envisaged the working ethic of democracy.

Other American Pragmatists. Two important contributions to American Pragmatism, which have not yet received the attention that they deserve, came from Mead and Lewis.

Mead's orientation was social psychology. He had studied physiological psychology in Germany, had earlier worked under James and Josiah Royce at Harvard, and was also familiar with Peirce's analyses of thought and signs. Dewey

regarded him as one of the most fertile minds in American philosophy.

Mead's social and Lewis' conceptualistic Pragmatism

Mead developed the most comprehensive of the Pragmatist theories of mind. He depicted the evolution of mind and self-consciousness as emerging from social interactions and the use of gestures and "significant symbols" such as words. In contrast to other creatures, an individual regarded as having mind, engaging with others in social acts, can respond to his own gestures as others respond to them—thus taking on social roles and becoming an "other" in respect to himself. It is therefore by means of language, the use of "significant symbols," that mind emerges.

Fundamental to Mead's philosophy is his conception of the social act, in which individuals modify and direct one anothers' activities, work out their purposes, and accordingly transform their environments. In the social act the future controls present conduct, and this is distinctive of consciousness. Since the function of intelligence is to render the world "favourable for conduct," Mead viewed the development of scientific knowledge and the evolutionary process as coinciding.

Lewis's theory of "conceptualistic Pragmatism" was derived partly from his study of modern logic and partly from the influence of Royce and the classic Pragmatists. The critical results of a careful study of Kant are traceable in his work.

Lewis's Pragmatism focusses upon concepts, categories, and principles through which experience is interpreted. Though the sensuously given is "unalterable," how it is taken, how conceptually interpreted, depends on the purposes and initiatives of the mind—the a priori element in knowledge, which, functioning as categorical criteria of reality, is "true no matter what." It is by means of these that a systematic interpretation of reality is developed. However,

> there may be alternative conceptual systems, giving rise to alternative descriptions of experience, which are equally objective and equally valid . . . . When this is so, choice will be determined . . . on pragmatic grounds.

In stressing the purposive character of conceptualization, Lewis is thus in the main course of American Pragmatism.

Pragmatism in Europe. In his preface to *Pragmatism,* James commented that the Pragmatic movement was the focal expression of a number of philosophic tendencies suddenly becoming conscious of themselves and of "their combined mission." He mentioned the French thinkers Maurice Blondel, Édouard Le Roy, and B. de Sailly and the Italian Giovanni Papini. Blondel was the author of *L'Action* (1893) and spokesman for a voluntaristic and activistic theory of knowledge. He was a founder of the "school of action," a liberal Catholic group that was part of the modernist movement (which employed the new historico-critical approach to the Bible and promoted a rationalistic interpretation of the faith). As early as 1888, Blondel appropriated the term *Pragmatisme,* only to abandon it when he learned of American Pragmatism, which was a more naturalistic philosophy than his own. Le Roy, closer to James than other French thinkers, also called his views Pragmatism. In broad respects he was like James in holding that the truth and the full significance of beliefs is found in acting them out. Le Roy was a disciple of Henri Poincaré, who had argued that scientific theories are not mere summaries of data, nor deduced from axioms, but are creative constructions, products of human thought and ingenuity, "conventions." To the question of what limits are imposed on otherwise arbitrary conventions, of what justifies them, Le Roy suggested their convenience in use. James saw similar forms of Pragmatism in Ernst Mach, Wilhelm Ostwald, Pierre Duhem, and Théodore Ruyssen in "the notion that no theory is absolutely a transcript of reality" and that "their great use is to summarize old facts and lead to new ones" so that they are a "man-made language, a conceptual shorthand . . . in which we write our reports of nature."

Blondel, Le Roy, Papini, and others

Another French thinker, Georges Sorel, undertook to reformulate James's Pragmatism into a "useful" doctrine of social criticism. Mussolini later cited Sorel and James as two of his philosophic mentors. He claimed to find in James "that faith in action, that ardent will to live and fight, to which Fascism owes a great part of its success." To the democratic James, no lesson could have been more badly learned.

A more immediate and direct form of James's Pragmatism occurred in Italy with its centre in the journal *Leonardo,* under the leadership of the iconoclastic critic Giovanni Papini. James referred to Papini as "a brilliant, humorous and witty writer." He called him a genius and was addressed in turn by him as "the Master." Papini's Pragmatism, derived from James's "The Will to Believe," became a theory of the will to action. In action, through creative power and passion, man achieves a kind of divinity. This romantic exaltation of action was appealing to artists but also to fanatics. Papini and his associate Giuseppe Prezzolini comprised the "magical" school of Pragmatism (in the sense of seeking "divinely-creative" power) in contrast to the "logical" school inspired by Peirce of G. Vailati and M. Calderoni.

Later tendencies. Certain extensions and applications of Pragmatism are to be found in current American philosophy. Sidney Hook has directed some of the critical techniques of analysis against a number of ideologies. In the tradition of the scientifically oriented Pragmatisms of Peirce and Dewey, he has explored the relation between the logic of experimental inquiry and the ethic of democracy. A converging of Pragmatism and Logical Positivism resulted in the movement of "logical empiricism" which, in addition to Dewey and Lewis, included the top-rank philosophers of science P.W. Bridgman, Rudolf Carnap, and Ernest Nagel, and the philosophical semanticist Charles W. Morris, all of whom were responsive to the Pragmatisms of Peirce, James, and Dewey. More recent and detailed studies of the structure of science, the nature of theories and explanation, by Carnap, Lewis, and Nagel, and a new interest in Instrumentalism deriving from the work of F.P. Ramsey, Ludwig Wittgenstein, W.H. Watson, and Gilbert Ryle (as well, of course, as from Dewey) exhibit a further continuation of Pragmatism.

Positivists and Nominalists

Maintaining a "more thorough Pragmatism" than that of Carnap and Lewis, who viewed choices made within scientific frameworks as pragmatic decisions, the prominent logician W.V. Quine has argued against the alleged boundary between analytic and synthetic truths. No portion of the conceptual scheme of science is exempt (as analytic truths were supposed to be) from possible revision in the flux of experience and in the light of pragmatic interests in efficacy and comprehension in predicting future experience. In important respects Quine's view of the evolution, organization, and function of the conceptual structure of science is close to that of Schiller, James, and Dewey.

Quine has also defended a methodological Pragmatism and relativism on ontological questions on the nature of being or reality. This position, also taken by another contemporary philosopher of science and language, Nelson Goodman, recalls the earlier Pragmatists' notion of the "plastic" character of reality, now seen as conceptually plastic in the sense of being expressible in a variety of systems of symbols and languages. It is conceptually misguided to seek *the* nature of objects, since what there *is* is not describable *in abstracto* from the particular language in which an ontological question has been put. Objects declared to be real, be they classes, numbers, atoms, or stones, may differ widely. But differences between the "theoretical" and the "factual" entities are basically differences of degree and purpose in evolving conceptualization. Hence reality is anything that it is truly said to be—in any of many linguistic and symbolic systems; and where differences in ways of speaking about objects call for choices, the ensuing adjudications will be pragmatic.

EVALUATION OF PRAGMATISM

Pragmatism has been vulnerable to certain criticisms. It has often been portrayed as a rationalization of the American business ethos—a portraiture perhaps inspired, but not by any scrutiny of the writings of the philosophers themselves. Similarly, the Pragmatic theory of truth has been assailed. Concerning an idea or belief, James had held that one can say: " 'It is useful because it is true' or

that 'it is true because it is useful.' " "Both phrases," he added, "mean the same thing." Most scholars, however, have denied this equivalence. His position may seem, moreover, to allow for an idea to be true (*i.e.,* useful or expedient) for one person and false (inexpedient) for others. Finally, James was accused of reducing truth to a subjective play of opinions that one happens to relish or find useful to believe. To these charges James replied that "what immediately feels most 'good' is not always most 'true' when measured by the verdict of the rest of experience." He also warned: "Woe to him whose beliefs play fast and loose with the order which realities follow in his experience."

As a single movement, Pragmatism is no longer extant; but as a body of ideas it contributes a heritage that is destined for future analysis and development. Chief among these are the interpretation of thought and meaning as forms of purposive behaviour, of knowledge as evaluative procedure in which normative and descriptive materials are integrally related, and of the logic of scientific inquiry as a norm of intelligent conduct in the affairs of men. Finally, Pragmatism has succeeded in its critical reaction to the 19th-century philosophy from which it emerged. It has influenced the current conception of philosophy as a critical method of investigating problems and clarifying communication rather than as a universal synthesis of knowledge. Pragmatism thus has certain affinities with the critical philosophizing of G.E. Moore and Bertrand Russell, as well as with the thought of the French intuitionist and vitalist Henri Bergson and his disciple Édouard Le Roy, of Blondel, of the early Positivists Mach and Duhem, of the fictionalist Hans Vaihinger, of the Vienna Circle and the philosopher of logic and language Ludwig Wittgenstein, and also of the founder of Phenomenology, Edmund Husserl, and some of the continuing forms of Phenomenology and Existentialism. It has recognized the relative, contingent, and fallible (yet still authentic) character of human reason, rather than perpetuating the dubious ideal of philosophy as a system of eternal truths. In so doing, and in thus altering the philosophical scene, Pragmatism has become vitally implicated in the practices of current intellectual life; and in the light of this fact, a more pragmatic justification of Pragmatism is difficult to imagine. (H.S.T.)

Rationalism

Rationalism is the philosophical view that regards reason as the chief source and test of knowledge. Holding that reality itself has an inherently logical structure, the Rationalist asserts that a class of truths exists that the intellect can grasp directly. There are, according to the Rationalists, certain rational principles—especially in logic and mathematics, and even in ethics and metaphysics—that are so fundamental that to deny them is to fall into contradiction. The Rationalist's confidence in reason and proof tends, therefore, to detract from his respect for other ways of knowing. Rationalism has long been the rival of Empiricism, the doctrine that all knowledge comes from, and must be tested by, sense experience. As against this doctrine, Rationalism holds reason to be a faculty that can lay hold of truths beyond the reach of sense perception, both in certainty and generality. In stressing the existence of a "natural light," Rationalism has also been the rival of systems claiming esoteric knowledge, whether from mystical experience, revelation, or intuition, and has been opposed to various irrationalisms that tend to stress the biological, the emotional or volitional, the unconscious, or the existential at the expense of the rational.

TYPES AND EXPRESSIONS OF RATIONALISM

Rationalism has somewhat different meanings in different fields, depending upon the kind of theory to which it is opposed.

Nativism in psychology and linguistics

In the psychology of perception, for example, Rationalism is in a sense opposed to the genetic psychology of the Swiss scholar Jean Piaget, who, exploring the development of thought and behaviour in the infant, argued that the categories of the mind develop only through the infant's experience in concourse with the world. Similarly, Rationalism is opposed to Transactionalism, a point of view in psychology according to which man's perceptual skills are achievements, accomplished through actions performed in response to an active environment. On this view, the experimental claim is made that perception is conditioned by probability judgments formed on the basis of earlier actions performed in similar situations. As a corrective to these sweeping claims, the Rationalist defends a nativism, which holds that certain perceptual and conceptual capacities are innate—as suggested in the case of depth perception by experiments with "the visual cliff," which, though platformed over with firm glass, the infant perceives as hazardous—though these native capacities may, at times, lie dormant until the appropriate conditions for their emergence arise.

In the comparative study of languages, a similar nativism was developed in the 1950s by the innovating syntactician Noam Chomsky, who, acknowledging a debt to Descartes, explicitly accepted the rationalistic doctrine of "innate ideas." Though the 4,000 languages spoken in the world differ greatly in sounds and symbols, they sufficiently resemble each other in syntax to suggest that there is "a schema of universal grammar" determined by "deep structures" or "innate presettings" in the human mind itself. These presettings, which have their basis in the brain, set the pattern for all experience, fix the rules for the formation of meaningful sentences, and explain why languages are readily translatable into one another. It should be added that what Rationalists have held about innate ideas is not that some ideas are full-fledged at birth but only that the grasp of certain connections and self-evident principles, when it comes, is due to inborn powers of insight rather than to learning by experience.

Metaphysical and epistemological Rationalism

Common to all forms of speculative Rationalism is the belief that the world is a rationally ordered whole, the parts of which are linked by logical necessity and the structure of which is therefore intelligible. Thus in metaphysics it is opposed to the view that reality is a disjointed aggregate of incoherent bits and is thus opaque to reason. In particular, it is opposed to the logical atomisms of such thinkers as David Hume and Ludwig Wittgenstein, who held that facts are so disconnected that any fact might well have been different from what it is without entailing a change in any other fact. Rationalists have differed, however, with regard to the closeness and completeness with which the facts are bound together. At the lowest level, they have all believed that the law of contradiction "A and not-A cannot coexist" holds for the real world, which means that every truth is consistent with every other; at the highest level, they have held that all facts go beyond consistency to a positive coherence; *i.e.,* they are so bound up with each other that none could be different without all being different.

In the field where its claims are clearest—in epistemology, or theory of knowledge—Rationalism holds that some, at least, of man's knowledge is gained through a priori (prior to experience), or rational, insight as distinct from sense experience, which too often provides a confused and merely tentative approach. In the debate between Empiricism and Rationalism, Empiricists hold the simpler and more sweeping position, the Humean claim that all knowledge of fact stems from perception. Rationalists, on the contrary, urge that some, though not all, knowledge arises through direct apprehension by the intellect. What the intellectual faculty apprehends is objects that transcend sense experience—universals and their relations. A universal is an abstraction, a characteristic that may reappear in various instances: the number three, for example, or the triangularity that all triangles have in common. Though these cannot be seen, heard, or felt, Rationalists point out that man can plainly think about them and about their relations. This kind of knowledge, which includes the whole of logic and mathematics as well as fragmentary insights in many other fields, is, in the Rationalist view, the most important and certain knowledge that the mind can achieve. Such a priori knowledge is both necessary (*i.e.,* it cannot be conceived as otherwise) and universal, in the sense that it admits of no exceptions. In critical

philosophy, epistemological Rationalism finds expression in the claim that the mind imposes its own inherent categories or forms upon incipient experience (see below *Kant and Hegel*).

Ethical and religious Rationalism

In ethics Rationalism holds the position that reason, rather than feeling, custom, or authority, is the ultimate court of appeal in judging good and bad, right and wrong. Among major thinkers, the most notable representative of rational ethics is Immanuel Kant, who held that the way to judge an act is to check its self-consistency as apprehended by the intellect: to note, first, what it *is* essentially, or in principle—a lie, for example, or a theft—and then to ask if one can consistently will that the principle be made universal. Is theft, then, right? The answer must be "No," because, if theft were generally approved, no one's property would be his own as opposed to anyone else's and theft would then become meaningless; the notion, if universalized, would thus destroy itself, as reason, by itself, is sufficient to show.

In religion Rationalism commonly means that all of man's knowledge comes through the use of his natural faculties, without the aid of supernatural revelation. "Reason" is here used in a broader sense, referring to man's cognitive powers generally, as opposed to supernatural grace or faith—though it is also in sharp contrast to so-called existential approaches to truth. Reason, for the Rationalist, thus stands opposed to many of the religions of the world, including Christianity, which have held that the divine has revealed itself through inspired persons or writings and which have required, at times, that its claims be accepted as infallible, even when they do not accord with natural knowledge. Religious Rationalists hold, on the other hand, that if the clear insights of man's reason must be set aside in favour of alleged revelation, then human thought is everywhere rendered suspect—even in the reasonings of the theologians themselves. There cannot be two ultimately different ways of warranting truth, they assert; hence Rationalism urges that reason, with its standard of consistency, must be the final court of appeal. Religious Rationalism can reflect either a traditional piety, when endeavouring to display the alleged sweet reasonableness of religion, or an anti-authoritarian temper, when aiming to supplant religion with the "goddess of reason."

HISTORY OF RATIONALISM

Pre-Socratics

Epistemological Rationalism in ancient philosophies. The first Western philosopher to stress rationalist insight was Pythagoras, a shadowy figure of the 6th century BC. Noticing that, for a right triangle, a square built on its hypotenuse equals the sum of those on its sides and that the pitches of notes sounded on a lute bear a mathematical relation to the lengths of the strings, Pythagoras held that these harmonies reflected the ultimate nature of reality. He summed up the implied metaphysical Rationalism in the words "All is number." It is probable that he had caught the Rationalist's vision, later seen by Galileo, of a world governed throughout by mathematically formulable laws.

The difficulty in this view, however, is that, working with universals and their relations, which, like the multiplication table, are timeless and changeless, it assumes a static world and ignores the particular, changing things of daily life. The difficulty was met boldly by the Rationalist Parmenides, who insisted that the world really is a static whole and that the realm of change and motion is an illusion, or even a self-contradiction. His disciple Zeno of Elea further argued that anything thought to be moving is confronted with a row of points infinite in number, all of which it must traverse; hence it can never reach its goal, nor indeed move at all. Of course, perception tells us that we do move; but Zeno, compelled to choose between perception and reason, clung to reason.

Plato and Aristotle

The exalting of rational insight above perception was also prominent in Plato (*c.* 427–347 BC). In the *Meno,* Socrates dramatized the innateness of knowledge by calling upon an illiterate slave boy and, drawing a square in the sand, proceeding to elicit from him, step by step, the proof of a theorem in geometry of which the boy could never have heard (to double the size of a square, draw a square on the diagonal). Such knowledge, Rationalists insist, is certain, universal, and completely unlearned.

Plato so greatly admired the rigorous reasoning of geometry that he is alleged to have inscribed over the door of his Academy "Let no one unacquainted with geometry enter here." His famous "ideas" are accessible only to reason, not to sense. But how are they related to sensible things? His answers differed. Sometimes he viewed the ideas as distilling those common properties of a class in virtue of which one identifies anything as a member of it. Thus what makes anything a triangle is its having three straight sides; this is its essence. At other times, Plato held that the idea is an ideal, a non-sensible goal to which the sensible thing approximates; the geometer's perfect triangle "never was on sea or land," though all actual triangles more or less embody it. He conceived the ideas as more real than the sensible things that are their shadows and saw that the philosopher must penetrate to these invisible essences and see with the eye of his mind how they are linked together. For Plato they formed an orderly system that was at once eternal, intelligible, and good.

Plato's successor Aristotle (384–322 BC) conceived of the work of reason in much the same way, though he did not view the ideas as independent. His chief contribution to Rationalism lay in his syllogistic logic, regarded as the chief instrument of rational explanation. Man explains particular facts by bringing them under general principles. Why does one think Socrates will die? Because he is a man, and man as such is mortal. Why should one accept the general principle itself that all men are mortal? In human experience such principles have so far held without exception. But the mind cannot finally rest in this sort of explanation. Man never wholly understands a fact or event until he can bring it under a principle that is self-evident and necessary; and he then has the clearest explanation possible. On this central thesis of Rationalism, the three great Greeks were in accord.

Nothing comparable in importance to their thought appeared in Rationalistic philosophy in the next 1,800 years, though the work of Thomas Aquinas in the 13th century was an impressive attempt to blend Greek Rationalism and Christian revelation into a single harmonious system.

Rationalism in non-Western thought

Thoroughgoing Rationalism is to be found only in the philosophical tradition that has come down from Greece; the mysticism of India and the practicality of China have offered a less congenial soil for it. The nearest parallels to it in Eastern thought are found in the work of the Indian philosopher Śaṅkara, who flourished about AD 800, and in that of the Chinese sage Chu-Hsi (1130–1200). Both were commentators on the ancient scriptures of their lands; and in ordering the scattered insights of these sources into intelligible systems, they did for their respective peoples something like what Aquinas did for the West in his harmonizing of Greek with Christian thought. Śaṅkara held, as Sir Sarvepalli Radhakrishnan has expressed it, that "the Absolute is the unattainable goal towards which the finite intellect strives." Perception is confined to what is transient and fragmentary; reason rises to truth that is timeless and universal; but even reason falls short of full understanding, which is achieved, if at all, only through mystical vision. Chu Hsi, who has influenced Chinese thought for the past six centuries, was a disciple of Confucius, though he had a stronger speculative interest than his master. He held that in all human minds a single reason was at work, which he called "the Way." All things were in some degree manifestations of it, and hence the understanding of the world lay in more complete identification with it.

Continental Rationalists

Epistemological Rationalism in modern philosophies. The first modern Rationalist was René Descartes (1596–1650), who was an original mathematician whose ambition was to introduce into philosophy the rigour and clearness that delighted him in mathematics. He set out to doubt everything in the hope of arriving in the end at something indubitable. This he reached in his famous *cogito ergo sum,* "I think, therefore I am"; for to doubt one's own doubting would be absurd. Here then was a fact of absolute certainty, rendered such by the clearness and distinctness with which it presented itself to his reason. His task was to build on this as a foundation, to deduce from it a

series of other propositions, each following with the same self-evidence. He hoped thus to produce a philosophical system on which men could agree as completely as they do on the geometry of Euclid. The main cause of error, he held, lay in the impulsive desire to believe before the mind is clear. The clearness and distinctness upon which he insisted was not that of perception but of conception, the clearness with which the intellect grasps an abstract idea, such as the number three, or its being greater than two.

His method was adopted in essentials by both Benedict Spinoza (1632–77) and G.W. Leibniz (1646–1716), who agreed that the framework of things could be known by a priori thinking. They differed from him, however, in their starting points. What was most undeniable to Spinoza was not the existence of his self but that of the universe, called by him Substance. From the idea of Substance, and with the aid of a few definitions and axioms, he derived his entire system, which he set forth in his *Ethics* in a formal fashion patterned after Euclid's geometry. Still, for both Spinoza and Leibniz much in nature remained stubbornly opaque. Leibniz distinguished necessary truths, those of which the opposite is impossible (as in mathematics), from contingent truths, the opposite of which is possible, such as "snow is white." But was this an ultimate distinction? At times Leibniz said boldly that if only man knew enough, he would see that every true proposition was necessarily true—that there are no contingent truths, that snow must be white.

Kant and Hegel

How, then, does reason operate and how is it possible to have knowledge that goes beyond experience? A new answer was given by Immanuel Kant (1724–1804) in his *Critique of Pure Reason,* which, as he said, involved a Copernican revolution in philosophy. The reason man can be certain that his logic and mathematics will remain valid for all experience is simply that their framework lies within his own mind; they are forms of arrangement imposed from within upon the raw materials of sensation. Man will always find things arranged in certain patterns because it is he who has unwittingly so arranged them. Kant held, however, that these certainties were bought at a heavy price. Just because a priori insights are the reflection of man's own mind, he cannot trust them as a reflection of the world outside himself. Whether the rational order in which man arranges his sensation—the order, for example, of time, space, and causality—represents an order holding among things-in-themselves (German *Dinge-an-sich*) he cannot hope to know. Kant's Rationalism was thus the counterpart of a profound Skepticism.

Rationalism in non-Western thought

G.W.F. Hegel (1770–1831), the most thoroughgoing of Rationalist thinkers, attempted to break out of this Skepticism. He argued that to think of an unknowable is already to bring it within the sphere of what is known and that it is meaningless to talk of a region in which logic is invalid. Further, to raise the question "Why?" is to presume that there is an intelligible answer to it; indeed the faith of the philosopher must be that the real is the rational and the rational real, for this faith is implicit in the philosophic enterprise itself. As an attempt to understand and explain the world, philosophy is a process of placing something in a context that reveals it as necessary. But this necessity is not, as earlier Rationalists had supposed, an all-or-nothing affair issuing in a self-evident finality. Understanding is a matter of degree. What alone would wholly satisfy thought is a system that is at once all-inclusive and so ordered that its parts entail each other. Hegel believed that the universe constitutes such a whole and, as an idealist, held that it is a single, absolute mind. To the degree that the philosopher embodies and realizes this mind, his own mind will achieve both truth and reality. Indeed, the advance of civilization reflects the enlarging presence and control of such a system in the human spirit. Broadly similar Rationalistic systems were developed in England by F.H. Bradley (1846–1924) and Bernard Bosanquet (1848–1923) and in America by Josiah Royce (1855–1916).

Ethical Rationalism. The views of Kant were presented above as typical of this position (see above, *Types and expressions of Rationalism*). But few moralists have held to ethical Rationalism in this simple and sweeping form. Many have held, however, that the main rules of conduct are truths as self-evident as those of logic or mathematics. Lists of such rules were drawn up by Ralph Cudworth and Henry More among the Cambridge Platonists of the 17th century, who were noted for holding that moral principles were intrinsic to reality; and in the 18th century Samuel Clarke and Richard Price, defenders of "natural law" ethics, and the "common sense" moralist Thomas Reid also presented such lists. A 20th-century revision of this Rationalism has been offered by the Rational Intuitionists H.A. Prichard and Sir David Ross of Oxford under the name of deontology (Greek *deon,* "duty"), which respects duty more than consequences. Ross provides a list of propositions regarding fidelity to promises, reparation for injuries, and other duties, of which he says: "In our confidence that these propositions are true there is involved the same trust in our reason that is involved in our trust in mathematics." What is taken as self-evident, however, is not specific rules of conduct, but *prima facie* duties—the claims that some types of action have on men because of their nature. If a man is considering whether to repay a debt or to give the money to charity, each act has a self-evident claim on him; and their comparative strengths must be settled by a rational intuition.

Legalistic versus utilitarian ethics

The most influential variety of 20th-century ethical Rationalism has probably been the Ideal Utilitarianism of the British moralists Hastings Rashdall (1858–1924) and G.E. Moore (1873–1958). Both were teleologists (Greek *telos,* "end") inasmuch as they held that what makes an act objectively right is its results (or end) in intrinsic goods or evils. To determine what is right, reason is required in two senses: firstly, the inference to the consequences is an act of inductive reasoning; secondly, the judgment that one consequence is intrinsically better than another is a priori and self-evident. Moore thought that there is a single rule for all conduct—one should so act as to produce the greatest good—and that this is also a principle self-evident to reason.

Religious Rationalism. Stirrings of religious Rationalism were already felt in the Middle Ages regarding the Christian revelation. Thus the skeptical mind of Abelard (1079–1142) raised doubts by showing in his *Sic et Non* ("Yes and No") many contradictions among beliefs handed down as revealed truths by the Church Fathers. The greatest of the Medieval thinkers, Thomas Aquinas (1225–74), was a Rationalist in the sense of believing that the larger part of revealed truth was intelligible to and demonstrable by reason, though he thought that a number of dogmas opaque to reason must be accepted on authority alone.

Expansion of religious Rationalism. Religious Rationalism did not come into its own, however, until the 16th and 17th centuries, when it took two chief forms: the scientific and the philosophic.

Galileo and Descartes

Galileo (1564–1642) was a pioneer in astronomy and the founder of modern dynamics. He conceived of nature as governed throughout by laws statable with mathematical precision; the book of nature, he said, is "written in mathematical form." This notion not only ruled out the occasional appeal to miracle; it also collided with dogmas regarding the permanent structure of the world—in particular with that which viewed the Earth as the motionless centre of the universe. When Galileo's demonstration that the Earth moves around the Sun was confirmed by the work of Newton and others, a battle was won that marked a turning point in the history of Rationalism, since it provided a decisive victory in a crucial case of conflict between reason and apparently revealed truth.

The Rationalism of Descartes, as already shown, was the outcome of philosophic doubt rather than of scientific inquiry. The self-evidence of the *cogito,* seen by his "natural light," he made the ideal for all other knowledge. The uneasiness that the church soon felt in the face of such a test was not unfounded, for Descartes was in effect exalting the natural light into the supreme court even in the field of religion. He argued that man's guarantee against the possibility that even this natural light might deceive him lay in the goodness of the Creator. But then to prove this Creator, he had to assume the prior validity of the natural light itself. Logically, therefore, the last word lay with rational insight, not with any outside divine warrant.

Descartes was inadvertently beginning a Copernican revolution in theology. Before his time, the truths regarded as most certain were those accepted from revelation; afterwards these truths were subject to the judgment of human reason, thus breaking the hold of authority on the European mind.

Deism

Four waves of religious Rationalism. The Rationalist attitude quickly spread, its advance forming several waves of general interest and influence. The first wave occurred in England in the form of Deism. Deists accepted the existence of God, but spurned supernatural revelation. The earliest member of this school, Lord Herbert of Cherbury (1583–1648), held that a just God would not reveal himself to a part of his creation only and that the true religion is thus a universal one, which achieves its knowledge of God through common reason. The Deistic philosopher John Toland (1670–1722), in his *Christianity Not Mysterious,* sought to show that "there is nothing in the Gospels contrary to reason, nor above it"; any doctrine that is really above reason would be meaningless to man. Attacking revelation, the freethinking polemicist Anthony Collins (1676–1729) maintained that the prophecies of the Old Testament failed of fulfillment; and the religious controversialist Thomas Woolston (1670–1733) urged that the New Testament miracles, as recorded, are incredible. Matthew Tindall (1657–1733), most learned of the English Deists, argued that the essential part of Christianity is its ethics, which, being clearly apparent to natural reason, leaves revelation superfluous. Thus the Deists, professing for the most part to be religious men themselves, did much to reconcile their public to the free play of ideas in religion.

Voltaire and Diderot

The second wave of religious Rationalism, less moderate in tone and consequences, was French. This wave, reflecting an engagement with the problem of natural evil, involved a decay in the natural theology of Deism such that it merged eventually with the stream that led to materialistic Atheism. Its moving spirit was Voltaire (1694–1778), who had been impressed by some of the Deists during a stay in England. Like them, he thought that a rational man would believe in God but not in supernatural inspiration. Hardly a profound philosopher, he was a brilliant journalist, clever and humorous in argument, devastating in satire, and warm in human sympathies. In his *Candide* and in many other writings, he poured irreverent ridicule on the Christian scheme of salvation as incoherent and on the church hierarchy as cruel and oppressive. In these attitudes he had the support of Diderot (1713–84), editor of the most widely read encyclopaedia that had appeared in Europe. The Rationalism of these men and their followers, directed against both the religious and the political traditions of their time, did much to prepare the ground for the explosive French Revolution.

Hegelianism

The next wave of religious Rationalism occurred in Germany under the influence of Hegel, who held that a religious creed is a halfway house on the road to a mature philosophy, the product of a reason that is still under the sway of feeling and imagination. This idea was taken up and applied with learning and acuteness to the origins of Christianity by David Friedrich Strauss (1808–74), who published in 1835, at the age of 27, a remarkable and influential three-volume work, *Das Leben Jesu* (*The Life of Jesus, Critically Examined,* 1846). Relying largely on internal inconsistencies in the Synoptic Gospels, Strauss undertook to prove these books to be unacceptable as revelation and unsatisfactory as history. He then sought to show how an imaginative people innocent of either history or science, convinced that a Messiah would appear, and deeply moved by a unique moral genius, inevitably wove myths about his birth and death, his miracles, and his divine communings.

Strauss's thought as it affected religion was continued by the philosophical historian Ernest Renan (1823–92) and as it affected philosophy by the humanist Ludwig Feuerbach (1804–72) of the Hegelian left. Renan's *Vie de Jésus* (1863; *Life of Jesus*) did for France what Strauss's book had done for Germany, though the two differed greatly in character. Whereas Strauss's work had been an intellectual exercise in destructive criticism, Renan's was an attempt to reconstruct the mind of Jesus as a wholly human person—a feat of imagination, performed with a disarming admiration and even reverence for its subject and with a felicity of style that gave it a large and lasting audience. Feuerbach's *Wesen des Christentums* (1841; Eng. trans. by George Eliot, *Essence of Christianity,* 1853) applied the myth theory even to belief in the existence of God, holding that "man makes God in his own image."

Evolution

The fourth wave occurred in Victorian England, following the publication in 1859 of Darwin's *Origin of Species.* This book was taken as a challenge to the authority of Scripture because there was a clear inconsistency between the Genesis account of creation and the biological account of man's slow emergence from lower forms of life. The battle raged with bitterness for several decades but died away as the theory of evolution gained more general acceptance.

STATUS OF RATIONALISM

Religious. With increasing freedom of thought and wider acceptance of scientific views, Rationalism in religion has lost its novelty and much of its controversial excitement. To the contemporary mind, it is too obvious to warrant debate that reason and revelation cannot both qualify as sources of ultimate truth for, were they to conflict, truth itself would become self-contradictory. Hence theologians have sought accommodation through new interpretative principles that discern different grades of authenticity within the Scriptures and through new views of religious truth, existential rather than cognitive, that turn from propositional dogmas to the explication of lived human existence. Criticism of supernaturalism, however, is still carried on by such societies as the Rationalist Press Association, in Great Britain, and the Humanist Association, in the United States.

Ethical. Rationalism in ethics has suffered its share of criticism. Regarding its lists of rules—on the keeping of promises, the return of loaned goods, etc.—it has been argued, for example, that if they were specific enough to be useful (as in the rule against lying or stealing), they would tend to have exceptions—which no rule laid down by reason ought to have. On the other hand, if without exceptions, they would often prove to be tautologies: the rule of justice, for example, that we should give everyone his due would then mean only that we should give him what is justly his. After enduring a period of eclipse, however, during which noncognitive theories of ethics (emotive and existential) and relativism had preempted the field, rationalistic views, which agree in holding that moral standards do not depend upon the varying attitudes of persons or peoples, were receiving renewed attention in the mid-20th century. Prominent among these developments has been the "good-reasons" approach taken by the broadly gauged scholar Stephen Toulmin, by Kurt Baier, and others, which examines the contexts of various moral situations and explores the kinds of justification appropriate for each.

Metaphysical. Typical of the ways of reasoning employed by Rationalists are two approaches taken to the metaphysical doctrine that all things are connected by internal relations: one a logical, the other a causal argument. An internal relation is one that could not be removed without affecting the terms themselves between which the relation holds. The argument runs: Everything is related to everything else at least by the relation "A *is different from* B." But difference is itself an internal relation, since the terms could not remain the same if it were removed. Hence everything is so connected with everything else that it could not be what it is unless they were what they are. The appeal to internal relations played an important part in the philosophies of Hegel, F.H. Bradley, and A.N. Whitehead (1861–1947).

The other line of argument is causal. Every event, it is maintained, is connected with every other, either directly or indirectly. Sir James Jeans argued that if the law of gravitation is valid, a man cannot crook his little finger without affecting the fixed stars. Here the causal relation is direct. It can also be shown that seemingly unrelated events are joined indirectly through their common con-

nection with some remote historical event, by a chain of events leading back, for example, to Columbus' discovery of America. But if this had been different, all of its consequences would presumably have been different; thus an indirect and internal relation proves to have been present.

Many Rationalists have held with Spinoza that the causal relation is really a logical one—that a causal law, if precisely stated, would reveal a connection in which the character of the cause logically necessitates that of its effect; and if this is true, they maintain, the facts and events of the world must thus compose a single rational and intelligible order.

Rationalism and quantum physics

In the 20th century, such Rationalism met with a new and unexpected difficulty presented by quantum physics. According to the indeterminacy principle, formulated in 1927 by the German physicist Werner Heisenberg, it is impossible to discover with precision both the position and the velocity of a moving electron at the same time. This implies that definite causal laws for the behaviour of these particles can never be attained, but only statistical laws governing the behaviour of immense aggregates of them. Causality, and with it the possibility of rational understanding, seemed to be suspended in the subatomic world. Some interpreters of the new physics, however, notably Max Planck, Albert Einstein, and Bertrand Russell, sustained the hopes of the Rationalists by insisting that what was excluded by the indeterminacy principle was not the fact of causality in this realm, but only the precise knowledge of it.

Indeed, some leaders of 20th-century science took the new developments in physics as on the whole supporting Rationalism. Protons and electrons, they contended, though beyond the reach of the senses, can still be known; and their behaviour, at least in groups, is increasingly found to conform to mathematical law. In 1932, Sir James Jeans, an astrophysicist and popularizer of science, said with a curious echo of Galileo, "the universe appears to have been designed by a pure mathematician."

Challenges to epistemological Rationalism. At first glance the claim of Empiricism that knowledge must come from sense experience seems obvious: how else could one hope to make contact with the world around one? Consequently, Rationalism has been sharply challenged—in the 19th century by the Empiricism of John Stuart Mill (1806–73) and in the 20th by that of the Logical Positivists. Mill argued that all a priori certainties are illusory: why does man believe, for example, that two straight lines cannot enclose a space? Is it because he sees it as logically necessary? No; it is because he has experienced so long and so unbroken a row of instances of it—a new one whenever he sees the corner of a table or the bordering rays of a light beam—that he has formed the habit of thinking in this way and is now unable to break it. A priori propositions, Mill claimed, are merely empirical statements of very high generality.

This theory has now been abandoned by most Empiricists themselves. Its implication that such statements as "2 + 2 = 4" are only probably true and may have exceptions has proved quite unconvincing. The Rationalist's rejoinder is that one cannot, no matter how hard one tries, conceive 2 + 2 as making 5; for its equalling 4 is necessary. But a priori knowledge is also universal. Neither of these two characteristics can be accounted for by sense experience. That a crow is black can be perceived, but not that it must be black or that crows will always be black; no run of perceptions, however long, could assure us of such truths. On the other hand, a priori truths can be seen with certainty—that if a figure, for instance, is a plane triangle within a Euclidean space, its angles must and always will equal two right angles.

Perhaps the most formidable challenge to Rationalism has come in the 20th century from such Logical Positivists as the Oxford Empiricist A.J. Ayer (1910–) and Rudolf Carnap (1891–1970), who had been a central figure in the Vienna Circle, where this movement first arose. Unlike Mill, they accepted a priori knowledge as certain; but they laid down a new challenge—the denial of its philosophic importance. A priori propositions, they said, are (1) linguistic, (2) conventional, and (3) analytic: (1) They are statements primarily of how one proposes to use words; if one says that "a straight line is the shortest line between two points," this merely reports one's definition of "straight" and declares one's purpose to use it only of the shortest. (2) Being a definition, such a statement expresses a convention to which there are alternatives; it may be defined in terms of the paths of light rays if one chooses. (3) The statement is analytic in that it merely repeats in its predicate a part or the whole of the subject term and hence tells nothing new; it is not a statement about nature but about meanings only. And since Rationalistic systems depend throughout upon statements of this kind, their importance is illusory.

To this clear challenge some leading Rationalists have replied as follows: (1) Positivists have confused real with verbal definition. A verbal definition does indeed state what a word means; but a real definition states what an object is, and the thought of a straight line is the thought of an object, not of words. (2) The Positivists have confused conventions in thought with conventions in language. One is free to vary the language in which a proposition is expressed, but not the proposition itself. Start with the concept of a straight line, and there is no alternative to accepting it as the shortest. (3) Some a priori statements are admittedly analytic, but many are not. In "whatever is coloured is extended," colour and extension are two different concepts of which the first entails the second, but is not identical with it in whole or part. Contemporary Rationalists therefore hold that the a priori has emerged victorious from the Empiricists' efforts to discredit such knowledge and the Positivists' attempts to trivialize it.

(B.Bl.)

Utilitarianism

The fundamental principle of Utilitarianism, a tradition in ethics stemming from the late 18th- and 19th-century English philosophers and economists Jeremy Bentham and John Stuart Mill, is that an action is right if it tends to promote happiness and wrong if it tends to produce the reverse of happiness—not just the happiness of the performer of the action but that of everyone affected by it. Such a theory is in opposition to egoism, the view that a man should pursue his own self-interest, even at the expense of others, and to any ethical theory that regards some acts or types of acts as right or wrong independently of their consequences. Utilitarianism also differs from ethical theories that make the rightness or wrongness of an act dependent upon the motive of the agent; for, according to the Utilitarian, it is possible for the right thing to be done from a bad motive.

THE NATURE OF UTILITARIANISM

Utilitarianism is an effort to provide an answer to the practical question "What ought a man to do?" Its answer is that he ought to act so as to produce the best consequences possible.

Assessment of consequences

Basic concepts. In the notion of consequences the Utilitarian includes all of the good and bad produced by the act, whether arising after the act has been performed or during its performance. If the difference in the consequences of alternative acts is not great, some Utilitarians do not regard the choice between them as a moral issue. According to Mill, acts should be classified as morally right or wrong only if the consequences are of such significance that a person would wish to see the agent compelled, not merely persuaded and exhorted, to act in the preferred manner.

In assessing the consequences of actions, Utilitarianism relies upon some theory of intrinsic value: something is held to be good in itself, apart from further consequences, and all other values are believed to derive their worth from their relation to this intrinsic good as a means to an end. Bentham and Mill were hedonists; *i.e.*, they analyzed happiness as a balance of pleasure over pain and believed that these feelings alone are of intrinsic value and disvalue. Utilitarians also assume that it is possible to compare the intrinsic values produced by two alternative actions and to estimate which would have better consequences. Bentham believed that a hedonic calculus is theoretically

possible. A moralist, he maintained, could sum up the units of pleasure and the units of pain for everyone likely to be affected, immediately and in the future, and could take the balance as a measure of the overall good or evil tendency of an action. Such precise measurement as Bentham envisioned is perhaps not essential, but it is nonetheless necessary for the Utilitarian to make some interpersonal comparisons of the values of the effects of alternative courses of action.

Methodologies. As a normative system providing a standard by which an individual ought to act and by which the existing practices of society, including its moral code, ought to be evaluated and improved, Utilitarianism cannot be verified or confirmed in the way in which a descriptive theory can; but it is not regarded by its exponents as simply arbitrary. Bentham believed that only in terms of a Utilitarian interpretation do words such as "ought," "right," and "wrong" have meaning and that whenever anyone attempts to combat the principle of utility, he does so with reasons drawn from the principle itself. Bentham and Mill both believed that human actions are motivated entirely by pleasure and pain; and Mill saw that motivation as a basis for the argument that, since happiness is the sole end of human action, the promotion of happiness is the test by which to judge all human conduct.

Meaning, motivation, and commonsense theories

One of the leading Utilitarians of the late 19th century, a Cambridge philosopher, Henry Sidgwick, rejected their theories of motivation as well as Bentham's theory of the meaning of moral terms and sought to support Utilitarianism by showing that it follows from systematic reflection on the morality of "common sense." Most of the requirements of commonsense morality, he argued, could be based upon Utilitarian considerations. In addition, he reasoned that Utilitarianism could solve the difficulties and perplexities that arise from the vagueness and inconsistencies of commonsense doctrines.

Most opponents of Utilitarianism have held that it has implications contrary to their moral intuitions—that considerations of utility, for example, might sometimes sanction the breaking of a promise. Much of the defense of Utilitarian ethics has consisted in answering these objections, either by showing that Utilitarianism does not have the implications that they claim it has or by arguing against the moral intuitions of its opponents. Some Utilitarians, however, have sought to modify the Utilitarian theory to account for the objections.

Criticisms. One such criticism is that, although the widespread practice of lying and stealing would have bad consequences, resulting in a loss of trustworthiness and security, it is not certain that an occasional lie to avoid embarrassment or an occasional theft from a rich man would not have good consequences, and thus be permissible or even required by Utilitarianism. But the Utilitarian readily answers that the widespread practice of such acts would result in a loss of trustworthiness and security. To meet the objection to not permitting an occasional lie or theft, some philosophers have defended a modification labelled "rule" Utilitarianism. It permits a particular act on a particular occasion to be adjudged right or wrong according to whether it is in accordance with or in violation of a useful rule; and a rule is judged useful or not by the consequences of its general practice. Mill has sometimes been interpreted as a "rule" Utilitarian, whereas Bentham and Sidgwick were "act" Utilitarians.

"Rule," "act," and "ideal" Utilitarianisms

Another objection, often posed against the hedonistic value theory held by Bentham, holds that the value of life is more than a balance of pleasure over pain. Mill, in contrast to Bentham, discerned differences in the quality of pleasures that made some intrinsically preferable to others independently of intensity and duration (the quantitative dimensions recognized by Bentham). Some philosophers in the Utilitarian tradition have recognized certain wholly nonhedonistic values without losing their Utilitarian credentials. A British philosopher, G.E. Moore, a pioneer of 20th-century Analysis, regarded many kinds of consciousness—including love, knowledge, and the experience of beauty—as intrinsically valuable independently of pleasure, a position labelled "ideal" Utilitarianism. Even in limiting the recognition of intrinsic value and disvalue to happiness and unhappiness, some philosophers have argued that those feelings cannot adequately be further broken down into terms of pleasure and pain and have thus preferred to defend the theory in terms of maximizing happiness and minimizing unhappiness. It is important to note, however, that even for the hedonistic Utilitarians, pleasure and pain are not thought of in purely sensual terms; pleasure and pain for them can be components of experiences of all sorts. Their claim is that, if an experience is neither pleasurable nor painful, then it is a matter of indifference and has no intrinsic value.

Another objection to Utilitarianism is that the prevention or elimination of suffering should take precedence over any alternative act that would only increase the happiness of someone already happy. Some recent Utilitarians have modified their theory to require this focus or even to limit moral obligation to the prevention or elimination of suffering—a view labelled "negative" Utilitarianism.

HISTORICAL SURVEY

The ingredients of Utilitarianism are found in the history of thought long before Bentham.

Antecedents of Utilitarianism among the ancients. A hedonistic theory of the value of life is found in the early 5th century BC in the ethics of Aristippus of Cyrene, founder of the Cyrenaic school, and 100 years later in that of Epicurus, founder of an ethic of retirement, and their followers in ancient Greece. The seeds of ethical universalism are found in the doctrines of the rival ethical school of Stoicism and in Christianity.

Growth of classical English Utilitarianism. In the history of English philosophy, some historians have identified Bishop Richard Cumberland, a 17th-century moral philosopher, as the first to have a Utilitarian philosophy. A generation later, however, Francis Hutcheson, a British "moral sense" theorist, more clearly held a Utilitarian view. He not only analyzed that action as best that "procures the greatest happiness for the greatest numbers" but proposed a form of "moral arithmetic" for calculating the best consequences. The Skeptic David Hume, Scotland's foremost philosopher and historian, attempted to analyze the origin of the virtues in terms of their contribution to utility. Bentham himself said that he discovered the principle of utility in the 18th-century writings of various thinkers: of Joseph Priestley, a dissenting clergyman famous for his discovery of oxygen; of the Frenchman Claude-Adrien Helvétius, author of a philosophy of mere sensation; of Cesare Beccaria, an Italian legal theorist; and of Hume. Helvétius probably drew from Hume, and Beccaria from Helvétius.

From Hutcheson to Bentham

Another strand of Utilitarian thought took the form of a theological ethics. John Gay, a biblical scholar and philosopher, held the will of God to be the criterion of virtue; but from God's goodness he inferred that God willed that men promote human happiness.

Bentham, who apparently believed that an individual in governing his own actions would always seek to maximize his own pleasure and minimize his own pain, found in pleasure and pain both the cause of human action and the basis for a normative criterion of action. The art of governing one's own actions Bentham called "private ethics." The happiness of the agent is the determining factor; the happiness of others governs only to the extent that the agent is motivated by sympathy, benevolence, or interest in the good will and good opinion of others. For Bentham, the greatest happiness of the greatest number would play a role primarily in the art of legislation, in which the legislator would seek to maximize the happiness of the entire community by creating an identity of interests between each individual and his fellows. By laying down penalties for mischievous acts, the legislator would make it unprofitable for a man to harm his neighbour. Bentham's major philosophical work, *An Introduction to the Principles of Morals and Legislation* (1789), was designed as an introduction to a plan of a penal code.

With Bentham, Utilitarianism became the ideological foundation of a reform movement, later known as "philosophical radicalism," that would test all institutions and policies by the principle of utility. Bentham attracted as his

disciples a number of younger (earlier 19th-century) men. They included David Ricardo, who gave classical form to the science of economics; John Stuart Mill's father, James Mill; and John Austin, a legal theorist. James Mill argued for representative government and universal male suffrage on Utilitarian grounds; he and other followers of Bentham were advocates of parliamentary reform in England in the early 19th century. John Stuart Mill was a spokesman for women's suffrage, state-supported education for all, and other proposals that were considered radical in their day. He argued on Utilitarian grounds for freedom of speech and expression and for the noninterference of government or society in individual behaviour that did not harm anyone else. Mill's essay "Utilitarianism," published in *Fraser's Magazine* (1861), is an elegant defense of the general Utilitarian doctrine and perhaps remains the best introduction to the subject. In it Utilitarianism is viewed as an ethics for ordinary individual behaviour as well as for legislation.

James and J.S. Mill; Sidgewick; Moore

Late 19th- and 20th-century Utilitarianism. By the time Sidgwick wrote, Utilitarianism had become one of the foremost ethical theories of the day. His *Methods of Ethics* (1874), a comparative examination of egoism, the ethics of common sense, and Utilitarianism, contains the most careful discussion to be found of the implications of Utilitarianism as a principle of individual moral action.

The 20th century has seen the development of various modifications and complications of the Utilitarian theory. G.E. Moore argued for a set of ideals extending beyond hedonism by proposing that one imaginatively compare universes in which there are equal quantities of pleasure but different amounts of knowledge and other such combinations. He felt that he could not be indifferent toward such differences. The recognition of "act" Utilitarianism and "rule" Utilitarianism as explicit alternatives was stimulated by the analysis of moral reasoning in "rule" Utilitarian terms by Stephen Toulmin, a British philosopher of science and moralist, and by Patrick Nowell-Smith, a moralist of the Oxford linguistic school; by the interpretation of Mill as a "rule" Utilitarian by another Oxford Analyst, J.O. Urmson; and by the analysis by John Rawls, a Harvard moral philosopher, of the significance for Utilitarianism of two different conceptions of moral rules. "Act" Utilitarianism, on the other hand, has been defended by J.J.C. Smart, a British-Australian philosopher.

EFFECTS OF UTILITARIANISM IN OTHER FIELDS

The influence of Utilitarianism has been widespread, permeating the intellectual life of the last two centuries. Its significance in law, politics, and economics is especially notable.

Legal and political theory

The Utilitarian theory of the justification of punishment stands in opposition to the "retributive" theory, according to which punishment is intended to make the criminal "pay" for his crime. According to the Utilitarian, the rationale of punishment is entirely to prevent further crime by either reforming the criminal or protecting society from him and to deter others from crime through fear of punishment.

In its political philosophy Utilitarianism bases the authority of government and the sanctity of individual rights upon their utility, thus providing an alternative to theories of natural law, natural rights, or social contract. What kind of government is best thus becomes a question of what kind of government has the best consequences—an assessment that requires factual premises regarding human nature and behaviour.

Generally, Utilitarians have supported democracy as a way of making the interest of government coincide with the general interest; they have argued for the greatest individual liberty compatible with an equal liberty for others on the ground that each individual is generally the best judge of his own welfare; and they have believed in the possibility and the desirability of progressive social change through peaceful political processes.

With different factual assumptions, however, Utilitarian arguments can lead to different conclusions. If the inquirer assumes that a strong government is required to check man's basically selfish interests and that any change may threaten the stability of the political order, he may be led by Utilitarian arguments to an authoritarian or conservative position. On the other hand, William Godwin, an early 19th-century political philosopher, assumed the basic goodness of human nature and argued that the greatest happiness would follow from a radical alteration of society in the direction of anarchistic Communism.

Classical economics received some of its most important statements from Utilitarian writers, especially Ricardo and John Stuart Mill. Ironically, its theory of economic value was framed primarily in terms of the cost of labour in production rather than in terms of the use value, or utility, of commodities. Later developments more clearly reflected the Utilitarian philosophy. William Jevons, one of the founders of the marginal utility school of analysis, derived many of his ideas from Bentham; and "welfare economics," while substituting comparative preferences for comparative utilities, reflected the basic spirit of the Utilitarian philosophy. In economic policy, the early Utilitarians had tended to oppose governmental interference in trade and industry on the assumption that the economy would regulate itself for the greatest welfare if left alone; later Utilitarians, however, lost confidence in the social efficiency of private enterprise and were willing to see governmental power and administration used to correct its abuses.

Economic and social theory

As a movement for the reform of social institutions, 19th-century Utilitarianism was remarkably successful in the long run. Most of their recommendations have since been implemented unless abandoned by the reformers themselves; and, equally important, Utilitarian arguments are now commonly employed to advocate institutional or policy changes.

SUMMARY AND EVALUATION

As an abstract ethical doctrine, Utilitarianism has established itself as one of the small number of live options that must be taken into account and either refuted or accepted by any philosopher taking a position in normative ethics. In contemporary discussion it has been divorced from adventitious involvements with the analysis of ethical language and with the psychological theory with which it was presented by Bentham. Utilitarianism now appears in various modified and complicated formulations. Bentham's ideal of a hedonic calculus is usually considered a practical if not a theoretical impossibility. Present-day philosophers have noticed further problems in the Utilitarian procedures. One of them, for example, is with the process of identifying the consequences of an act—a process that raises conceptual as well as practical problems as to what are to be counted as consequences, even without precisely quantifying the value of those consequences. The question may arise whether the outcome of an election is a consequence of each and every vote cast for the winning candidate if he receives more than the number necessary for election; and in estimating the value of the consequences, one may ask whether the entire value or only a part of the value of the outcome of the election is to be assigned to each vote. There is also difficulty in the procedure of comparing alternative acts. If one act requires a longer period of time for its performance than another, one may ask whether they can be considered alternatives. Even what is to count as an act is not a matter of philosophical consensus.

Problems in Utilitarian procedures

These problems, however, are common to almost all normative ethical theories since most of them recognize the consequences—including the hedonic—of an act as being relevant ethical considerations. The central insight of Utilitarianism, that one ought to promote happiness and prevent unhappiness whenever possible, seems undeniable. The critical question, however, is whether the whole of normative ethics can be analyzed in terms of this simple formula. (H.R.W.)

BIBLIOGRAPHY. For bibliographies of Platonism and Aristotelianism, see the *Macropædia* articles PLATONISM and ARISTOTELIANISM, respectively.

Atomism: E. CANTORE, *Atomic Order: An Introduction to the Philosophy of Microphysics* (1969); F. COPLESTON, *A History of Philosophy*, 8 vol. (1950–66); E.J. DIJKSTERHUIS, *Die Mecha-*

nisierung des Weltbildes (1956; Eng. trans., *The Mechanization of the World Picture*, 1961), a history of science from antiquity to the 17th century; A.S. EDDINGTON, *The Philosophy of Physical Science* (1939); K. LASSWITZ, *Geschichte der Atomistik vom Mittelalter bis Newton*, 2 vol. (1890, reprinted 1963), a 19th-century classic; A.G.M. VAN MELSEN, *Van atomos naar atoom* (1949; Eng. trans., *From Atomos to Atom: The History of the Concept Atom*, 2nd ed. 1960), including references for the primary sources; L.K. NASH, *The Atomic-Molecular Theory* (1950), a discussion of the first phase of the chemical atomic theory; E.T. WHITTAKER, *History of the Theories of Aether and Electricity*, rev. ed., 2 vol. (1951–54), only for readers with a solid background in science; L.L. WHYTE, *Essay on Atomism: From Democritus to 1960* (1961), a brief introduction to the idea of Atomism and its history. See also MRINALKAUTI GANGOPADHYAYA, *Indian Atomism: History and Sources* (1980).

Eleaticism: The main edition of the fragments of the Eleatic, as of all the Pre-Socratic, philosophers is still HERMANN DIELS, *Die Fragmente der Vorsokratiker* (1903; 11th ed. rev. by WALTHER KRANZ, 3 vol., 1964), which is a critical edition of the Greek fragments with German translations; for English translations see KATHLEEN FREEMAN, *Ancilla to the Pre-Socratic Philosophers* (1948, reprinted 1962); PILO ALBERTELLI, *Gli Eleati: testimonianze e frammenti* (1939), is a good Italian translation with commentary. Much of the best material is in general works on ancient Greek or Pre-Socratic philosophy. The most comprehensive treatment is in EDUARD ZELLER, *Die Philosophie der Griechen*, 6–7th ed., 3 vol., ed. by W. NESTLE (1920–23); in the revised Italian trans. of Zeller, ed. with extensive additions by RODOLFO MONDOLFO, the portion on the Eleatic school by GIOVANNI REALE (1967) occupies the whole 3rd volume of the 1st section and contains a good selected bibliography; an English translation of Zeller is entitled *A History of Greek Philosophy, from the Earliest Period to the Time of Socrates*, 2 vol. (1881). See also JOHN BURNET, *Early Greek Philosophy*, 4th ed. (1930, reprinted 1963); GEOFFREY KIRK and JOHN E. RAVEN, *The Presocratic Philosophers*, pp. 263–306 (1957); W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 2 (1965); and GUIDO CALOGERO, *Storia della logica antica*, vol. 1, pp. 109–208 (1967).

Epicureanism: The first complete collection of the extant works and fragments of Epicurus is H. USENER (ed.), *Epicurea* (1887, reprinted 1966). A smaller selection, with Eng. trans. and commentary, is CYRIL BAILEY (ed.), *Epicurus: The Extant Remains* (1926), a very useful book that includes the "Vatican Fragments." All of the ethical fragments (and several other items) are published in CARLO DIANO (ed.), *Epicuri Ethica* (1946), with extensive Latin commentaries. CARLO DIANO (ed. and trans.), *Lettere di Epicuro e dei suoi* (1946), contains 14 letters of Epicurus and his friends taken from Pap. Herc. 1418. G. ARRIGHETTI (ed.), *Epicuro, Opere* 2nd ed. (1967), contains all of the works and fragments (including the *Peri Physeos*), with notes and an *index verborum;* to be used with caution. Lucretius can be read in the three volumes prepared, with introduction, translations, and comments, by CYRIL BAILEY, *Lucreti, De Rerum Natura* (1947).

English translations of Epicurus and Lucretius include WHITNEY J. OATES (ed.), *The Stoic and Epicurean Philosophers* (1940), which contains translations by CYRIL BAILEY and by H.A.J. MUNRO. See also *The Philosophy of Epicurus: Letters, Doctrines, and Parallel Passages from Lucretius*, trans. by GEORGE K. STRODACH (1963); DIOGENES LAERTIUS, *Lives of Eminent Philosophers*, bk. 10 on Epicurus, trans. by R.D. HICKS (1925); and LUCRETIUS, *On the Nature of the Universe*, trans. by R.E. LATHAM (1964).

Valuable for its breadth and richness of detail, CYRIL BAILEY, *The Greek Atomists and Epicurus* (1928, new ed., 1964), is a fundamental work. For Epicurus' psychology and his relation to Aristotle, see CARLO DIANO, "La psicologia d'Epicuro e la teoria delle passioni," in *Giornale critico della filosofia italiana* (1939–42). A book that stresses Epicurus' anti-Platonism is N.W. DEWITT, *Epicurus and His Philosophy* (1954). Also important for its perceptive study of Epicurus' religiosity and ethics is A.J. FESTUGIÈRE, *Épicure et ses dieux* (1946; Eng. trans., *Epicurus and His Gods*, 1956). See also the article "Epikur," by W. SCHMID in the *Reallexikon für Antike und Christentum*, 5:681–819 (1961); L.L. WHYTE, *Essay on Atomism: From Democritus to 1960* (1961); and BENJAMIN FARRINGTON, *The Faith of Epicurus* (1967). G.D. HADZSITS, *Lucretius and His Influence* (1935), traces the influence of Epicurean ideas, especially since Roman times, in a broad perspective. For further bibliography, see PHILLIP DE LACY, "Some Recent Publications on Epicurus and Epicureanism (1937–1954)," in *Classical Weekly*, 48:169–177 (1955). BERNARD FRISCHER, *The Sculpted Word: Epicureanism and Philosophical Recruitment in Ancient Greece* (1982), a discussion of the school's success.

Pythagoreanism: The collection of the fragments in HERMANN DIELS and WALTHER KRANZ, *Die Fragmente der Vorsokratiker*, 6th ed., vol. 1 (1951), is insufficient; additions are given in MARIA TIMPANARO CARDINI (ed.), *Pitagorici: Testimonianze e frammenti*, 3 vol. (1958–64); and in CORNELIA J. DE VOGEL, *Pythagoras and Early Pythagoreanism* (1966). For the pseudo-Pythagoreans, see HOLGER THESLEFF (ed.), *The Pythagorean Texts of the Hellenistic Period* (1965).

The best comprehensive introduction to Pythagoreanism is the long chapter "Pythagoras and the Pythagoreans," in W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 1, pp. 146–340 (1962). Somewhat different approaches have been taken by DE VOGEL (*op. cit.*); and JAMES A. PHILIP, *Pythagoras and Early Pythagoreanism* (1966), works that demand more active criticism by the reader. Fairly full references to the discussion of Pythagoreanism up to 1960 are in WALTER BURKERT, *Weisheit und Wissenschaft: Studien zu Pythagoras, Philolaos, und Platon* (1962; Eng. trans., *Lore and Science in Ancient Pythagoreanism*, 1972), a highly technical and at times rather overcritical work. Among later technical discussions, likely to become influential, are the articles "Pythagoras" and "Pythagoreer" in *Pauly-Wissowa Realencyclopädie*, vol. 47, (1963), and suppl. vol. 10 (1965)—of the contributors, KURT VON FRITZ and H. DORRIE arrive at less controversial conclusions than B.L. VAN DER WAERDEN.

Hellenistic Pythagoreanism is treated in HOLGER THESLEFF, *An Introduction to the Pythagorean Writings of the Hellenistic Period* (1961); additions and corrections in *Entretiens Fondation Hardt*, vol. 18 (1972). Neo-Pythagoreanism is treated in PHILIP MERLAN, *From Platonism to Neoplatonism* (1953).

For up-to-date bibliographical accession, see *L'Année philologique* (annual), under the subject heading "Pythagorica" and the various Pythagoreans.

Realism: There is no single monograph that covers comprehensively the whole topic of Realism. The following works will be helpful for the study of its particular phases.

For Medieval Realism, see M.H. CARRÉ, *Realists and Nominalists* (1946, reissued 1967). For Neo-realism and critical Realism, see E.B. HOLT *et al., The New Realism* (1912, reprinted 1970); DURANT DRAKE *et al., Essays in Critical Realism* (1920, reprinted 1968); RENÉ KREMER, *La Théorie de la connaissance chez les néo-réalistes anglais* (1928); T.E. HILL, "Realistic Theories," *Contemporary Theories of Knowledge*, pp. 77–205 (1961, reissued 1980), good general coverage of the American and English fields. Neoscholastic Realisms are treated in LÉON NOËL, *Le Réalisme immédiat* (1938); and works by ÉTIENNE GILSON: *Le Réalisme méthodique* (1936), which presents the author's own views; ch. 5, "Vade Mecum of a Young Realist," trans. by W.J. QUINN, in R. HOUDE and J. MULLALLY (eds.), *Philosophy of Knowledge* (1960); and *Réalisme Thomiste et critique de la connaissance* (1939), a critique of the leading Neoscholastic views. For the linguistic approach, see MARTIN LEAN, *Sense-Perception and Matter* (1953, reprinted 1973); J.L. AUSTIN, *Sense and Sensibilia* (1962).

Universals in contemporary thought are treated in I.M. BOCHENSKI, ALONZO CHURCH, and NELSON GOODMAN, *The Problem of Universals* (1956); FARHANG ZABEEH, *Universals* (1966); FANAYOT BUNCHVAROV, *Resemblance and Identity: An Examination of the Problem of Universals* (1966, reissued 1982); NICHOLAS WOLTERSTORFF, *On Universals: An Essay in Ontology* (1970).

For Scientific Realism, see BERTRAND RUSSELL, *Our Knowledge of the External World*, rev. ed. (1926, various printings); W.H. WERKMEISTER, *The Basis and Structure of Knowledge* (1948, reissued 1968), bibliography, pp. 420–438; MORITZ SCHLICK, "Are Natural Laws Conventions?" in H. FEIGL and MAY BRODBECK (eds.), *Readings in the Philosophy of Science*, pp. 181–188 (1953); ROMANO HARRÉ, *Theories and Things* (1961); GROVER MAXWELL, "The Ontological Status of Theoretical Entities," in *Minnesota Studies in the Philosophy of Science*, 3:3–27 (1962); J.J.C. SMART, *Philosophy and Scientific Realism* (1963); DAVID M. ARMSTRONG, *Universals and Scientific Realism*, 2 vol. (1980).

Other 20th-century Realisms are discussed in G.D. HICKS, *Critical Realism* (1938); JAMES FEIBLEMAN, *The Revival of Realism* (1946, reissued 1972); WILFRID SELLARS, *Science, Perception and Reality* (1963); R.M. CHISHOLM (ed.), *Realism and the Background of Phenomenology* (1961), convenient bibliography, pp. 290–304; J.D. WILD (ed.), *The Return to Reason: Essays in Realistic Philosophy* (1953); E.B. MCGILVARY, *Toward a Perspective Realism* (1956); D.M. ARMSTRONG, *A Materialist Theory of the Mind* (1968); INGEBORG WIRTH, *Realismus und Apriorismus in Nicolai Hartmanns Erkenntnistheorie* (1965); GUSTAV BERGMANN, *Realism: A Critique of Brentano and Meinong* (1967).

Scholasticism: For the latest and most detailed bibliography, see W. TOTOK, *Handbuch der Geschichte der Philosophie*, vol. 2, *Mittelalter und frühe Neuzeit* (1970). Among the most reliable, best grounded presentations of the whole period are: E. GILSON, *History of Christian Philosophy in the Middle Ages* (1955); F. COPLESTON, *A History of Philosophy*, vol. 2, *Mediaeval Philosophy*, vol. 3, *Late Mediaeval and Renaissance Philosophy* (1950, 1953; paperback edition, 1962, 1963); M. DE WULF, *Histoire de la philosophie médiévale*, 6th ed., 3 vol. (1934–47; Eng.

trans. of vol. 1, 1951); and E. BREHIER, *La Philosophie du moyen âge*, 2nd ed. (1949). Still indispensable, though obsolete in some details is: F. UEBERWEG, *Grundriss der Geschichte der Philosophie*, vol. 2, B. GEYER, *Die patristische und scholastische Philosophie*, 11th ed. (1928). Lucidly arranged and divided is the 13th volume of FLICHE-MARTIN, *Histoire de l'Église: Le Mouvement doctrinal de XIe au XIVe siècle* (1951), which includes contributions by A. Forest, F. van Steenberghen, and M. de Gandillac. M. GRABMANN, *Mittelalterliche Geschichte der scholastischen Methode*, 2 vol. (1909–11, reprinted 1956), covers only the time until the first years of the 13th century. For a first introduction for the general reader, see J. PIEPER, *Scholastik* (1960; Eng. trans., *Scholasticism*, 1960). Special problems concerning the continuing influence of medieval Scholasticism are treated in the following monographs: A. KOYRE, *Descartes und die Scholastik* (1923); A. TELLKAMP, *Das Verhältnis John Locke's zur Scholastik* (1927); and J.O. FLECKENSTEIN, *Scholastik, Barock, exakte Wissenschaften* (1949). The following are sources on Neoscholasticism: J.P. GOLINAS, *La Restauration du Thomisme sous Leon XIII et les philosophies nouvelles* (1959); GIOVANNI BOSSI, *Le origini del Neotomismo nell'ambiente di studio del Collegio Alberoni* (1957); A. VIEL, "Le Mouvement thomiste au XIXe siècle." *Revue Thomiste* (1909–10); E. BETTONI, *La Situation actuelle de la philosophie parmi les catholiques dans les divers pays* (1948), a survey of centres of study, institutes, and publications; and J.S. ZYBURA (ed.), *Present-Day Thinkers and the New Scholasticism: An International Symposium* (1926).

Skepticism: The basic statements and arguments of various forms of Skepticism are given in: (Academic Skepticism)—CICERO, *Academica* and *De natura deorum*, both with trans. by H. RACKHAM, Loeb Classical Library (1956). (Pyrrhonian Skepticism)—SEXTUS EMPIRICUS, *Adversus Mathematicos*, with trans. by R.G. BURY, Loeb Classical Library: vol. 1–2, *Against the Logicians* and *Outlines of Pyrrhonism* (1933–36); vol. 3, *Against the Physicists, Against the Ethicists* (1936); vol. 4, *Against the Professors* (1959–60); and *Scepticism, Man and God: Selections from the Major Writings of Sextus Empiricus*, ed. by P. HALLIE, trans. by S.G. ETHERIDGE (1964). (Renaissance Skepticism)—MICHEL DE MONTAIGNE, "L'Apologie de Raimond Sebond," in PIERRE VILLEY (ed.), *Les Essais de Michel de Montaigne*, new ed. (1922). (Skepticism and fideism)—BLAISE PASCAL, *Pensées*, ed. by L. BRUNSCHVICG (1951). (Skepticism in relation to modern philosophy)—PIERRE BAYLE, *Dictionnaire historique et critique*, esp. the articles "Pyrrho" and "Zeno of Elea," both of which appear in BAYLE'S *Historical and Critical Dictionary: Selections*, trans. and ed. by RICHARD H. POPKIN (1965); DAVID HUME, *Dialogues Concerning Natural Religion*, ed. by N. KEMP SMITH (1947); *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals*, 2nd ed. (1957), and *A Treatise of Human Nature*, both ed. by L.A. SELBY-BIGGE (1958).

The standard studies of ancient Skepticism are: EDWYN R. BEVAN, *Stoics and Sceptics* (1959); VICTOR BROCHARD, *Les Sceptiques grecs* (1887); NORMAN MACCOLL, *The Greek Sceptics from Pyrrho to Sextus* (1869); MARY MILLS PATRICK, *The Greek Sceptics* (1929); LEON ROBIN, *Pyrrhon et le scepticisme grec* (1944); and EDUARD ZELLER, *The Stoics, Epicureans and Sceptics*, trans. by O.J. REICHEL (1880). A recent fine study of the epistemological problems involved in ancient Skepticism is CHARLOTTE L. STOUGH, *Greek Skepticism* (1969). See also RAOUL RICHTER, *Der Skeptizismus in der Philosophie* (1904–08); RICHARD H. POPKIN, "Skepticism," *Encyclopedia of Philosophy*, vol. 7, pp. 449–461 (1967), which contains a bibliography on the subject; these examine Skepticism from ancient times to the 19th century. JOHN OWEN, *The Skeptics of the French Renaissance* (1893), is an interesting discussion of Renaissance and 17th-century Skepticism, though not particularly scholarly. DON CAMERON ALLEN, *Doubt's Boundless Sea: Skepticism and Faith in the Renaissance* (1964); and RICHARD H. POPKIN, *The History of Scepticism from Erasmus to Descartes*, rev. ed. (1968), which contains a lengthy bibliography, are studies of Renaissance Skepticism and its impact on philosophy and religion. RICHARD H. POPKIN, "Berkeley and Pyrrhonism," *Review of Metaphysics*, 5:223–246 (1951–52); "David Hume and the Pyrrhonian Controversy," *ibid.*, 6:65–81 (1952–53); "David Hume: His Pyrrhonism and His Critique of Pyrrhonism," *Philosophical Quarterly*, 1:385–407 (1950–51); "The High Road to Pyrrhonism," *American Philosophical Quarterly*, 2:1–15 (1965); "The Skeptical Crisis and the Rise of Modern Philosophy," *Review of Metaphysics*, 7:132–151, 307–322, 499–510 (1953–54); "The Skeptical Precursors of David Hume," *Philosophy and Phenomenological Research*, 16:61–71 (1955–56); and "Scepticism in the Enlightenment," in T. BESTERMANN (ed.), *Studies on Voltaire and the Eighteenth Century*, 26:1321–1345 (1963), are specialized studies on aspects of Skepticism in relation to modern philosophy. ARNE NAESS, *Scepticism* (1968), is a most interesting attempt to clarify Skepticism in relation to contemporary thought and to defend it as a viable outlook today. BENSON MATES, *Skeptical Essays* (1981), discusses epistemological problems and free will.

Sophists: Ancient sources and fragments are presented in R.K. SPRAGUE, *The Older Sophists, a Complete Translation* (1972), and in K. FREEMAN, *Ancilla to the Pre-Socratic Philosophers* (1948), to be used together with Freeman's *The Pre-Socratic Philosophers*, 2nd ed. (1949). The original Greek texts are in H. DIELS and W. KRANZ (eds.), *Die Fragmente der Vorsokratiker*, 6th ed., vol. 2 (1952); and are edited with an Italian translation and commentary in M. UNTERSTEINER, *Sofisti*, 4 vol. (1949–62). (*General discussions*): T. GOMPERZ, *Greek Thinkers*, Eng. trans. by L. MAGNUS, vol. 1, bk. 3, ch. 5–7 (1901); M. UNTERSTEINER, *The Sophists*, Eng. trans. by K. FREEMAN (1954); W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 3, *The Fifth-Century Enlightenment* (1969). G.B. KERFERD, *The Sophistic Movement* (1981), is a study of the intellectual contributions made by the movement.

The Second Sophistic period is treated in G.W. BOWERSOCK, *Greek Sophists in the Roman Empire* (1969).

Stoicism: HANS VON ARNIM, *Stoicorum Veterum Fragmenta*, 4 vol. (1905–24), is the standard text collection. The following works of classic Stoic authors and substantive issues in Stoicism in antiquity are most conveniently available in the "Loeb Classical Library": MARCUS AURELIUS, *Meditations*; EPICTETUS, *Discourses*; CICERO, *De fato, De finibus, De natura deorum*, and *Academica*; DIOGENES LAERTIUS, *Lives of Eminent Philosophers*, vol. 2, book 7; SENECA, *Epistulae Morales* and *Moral Essays*; SEXTUS EMPIRICUS, *Outlines of Pyrrhonism* and *Against the Dogmatists*, 4 vol.; GEORGE MURRAY, *The Stoic Philosophy* (1915), the classical statement of the grandeur of the Stoic philosophy; MAX POHLENZ, *Die Stoa*, 2 vol. (1948–49); and PAUL BARTH, *Die Stoa* (1903), are representative of scholarly studies of Stoic philosophy; JOHNNY CHRISTENSEN, *An Essay on the Unity of Stoic Philosophy* (1962), is a recent and comprehensive essay; LUDWIG EDELSTEIN, *The Meaning of Stoicism* (1966); and JOSIAH B. GOULD, *The Philosophy of Chrysippus* (1970), are among the best of recent studies of Greco-Roman Stoicism; EDUARD ZELLER, *Die Philosophie der Griechen in ihrer geschichtlichen Entwicklung*, 5 vol. (1856–68; in part translated as *Stoics, Epicureans, and Skeptics*, rev. ed., 1962); EDWYN R. BEVAN, *Stoics and Sceptics* (1913, reprinted 1959); and ROBERT D. HICKS, *Stoic and Epicurean* (1910, reprinted 1962), are illustrations of Stoic philosophy in the Hellenistic period. See also JASON L. SAUNDERS, *Greek and Roman Philosophy After Aristotle* (1966); and HARRY A. WOLFSON, *Philo: Foundations of Religious Philosophy in Judaism, Christianity and Islam*, 2 vol. (1962). EDWARD V. ARNOLD, *Roman Stoicism* (1911, reprinted 1958); and FREDERICK W. BUSSELL, *Marcus Aurelius and the Later Stoics* (1910), are excellent statements of Stoicism in the later Roman period. PIERRE DE LABRIOLLE, *Histoire de la littérature latine chrétienne* (1920; Eng. trans., *History and Literature of Christianity from Tertullian to Boethius*, 1924), an excellent presentation of the influence of Stoic views in late antiquity and the patristic period; ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955); HARRY A. WOLFSON, *The Philosophy of the Church Fathers*, 3rd ed. rev. (1970); MAURICE DE WULF, *Histoire de la philosophie médiéval*, 2 vol. (1925; Eng. trans., *History of Mediaeval Philosophy*, 2 vol., 1926); and ISAAC HUSIK, *A History of Medieval Jewish Philosophy* (1940) are careful introductions to Stoic influences in patristic and medieval times; JASON L. SAUNDERS, *Justus Lipsius: The Philosophy of Renaissance Stoicism* (1955); HAROLD HOFFDING, *A History of Modern Philosophy*, 2 vol. (1950); and FREDERICK C. COPLESTON, *A History of Philosophy*, rev. ed., vol. 3–6 (1962), trace the Stoic influence from its revival in the Renaissance into modern philosophy. JOHN M. RIST (ed.), *The Stoics* (1978), is an excellent collection of essays covering logic, cosmology, ethics, psychology, and aesthetics.

Stoic logic and physics are treated in BENSON MATES, *Stoic Logic* (1953); SAMUEL SAMBURSKY, *The Physical World of the Greeks*, 2nd ed. (1960; orig. pub. in Hebrew) and *Physics of the Stoics* (1959).

Analytic and Linguistic philosophy: G.J. WARNOCK, *English Philosophy Since 1900* (1958), is a very readable account covering much of the history. J.O. URMSON, *Philosophical Analysis* (1956), is especially good on logical atomism and its demise. BERTRAND RUSSELL, *The Problems of Philosophy* (1959), is a nontechnical account of his views just prior to the period of logical atomism. A collection of Russell's papers, *Logic and Knowledge*, ed. by R.C. MARSH (1956), contains his "Lectures on Logical Atomism," together with more technical papers. G.E. MOORE, *Some Main Problems of Philosophy* (1953), gives a good introduction to his methods and philosophical concerns. Two key papers by Moore are "A Defence of Common Sense," in *Contemporary British Philosophy*, Second Series (1925), and "Proof of an External World," in *Proceedings of the British Academy*, 25:273–300 (1939). Among anthologies are those of A.J. AYER (ed.), *Logical Positivism* (1959), which also includes

papers critical of the movement and has an excellent bibliography; BERNARD WILLIAMS and ALAN MONTEFIORE (eds.), *British Analytical Philosophy* (1966); RICHARD RORTY (ed.), *The Linguistic Turn* (1967); J.A. FODOR and J.J. KATZ (eds.), *The Structure of Language* (1964); and LEONARD LINSKY (ed.), *Semantics and the Philosophy of Language* (1952). The 1961 translation of WITTGENSTEIN's *Tractatus Logico-Philosophicus* by D.F. PEARS and B.F. MCGUINNESS is superior to the original English version. Wittgenstein's *Philosophical Investigations* (1953) is the best known and historically the most influential of the posthumously published works from his later period. A.J. AYER, *Language, Truth and Logic*, 2nd ed. (1946), remains the best introduction to Logical Positivism. GILBERT RYLE, *The Concept of Mind* (1949), is the most famous example of what has been called "ordinary language" philosophy. J.L. AUSTIN, *How To Do Things with Words* (1962), contains the most discussed features of his views. W.V.O. QUINE, *From a Logical Point of View* (1953), is a collection of fairly nontechnical essays that provide good examples of the continuing influence of formal logic.

Empiricism: Classic texts include JOHN LOCKE, *An Essay Concerning Human Understanding*, 2 vol. (1690); DAVID HUME, *A Treatise of Human Nature*, book 1, pt. 1 (1739); IMMANUEL KANT, *Kritik der reinen Vernunft* (1781; Eng. trans., *Critique of Pure Reason*, 1929); JOHN STUART MILL, *A System of Logic, Ratiocinative and Inductive*, books 1 and 2 (1843).

W.H. WALSH, *Reason and Experience* (1947); and H.H. PRICE, *Thinking and Experience*, 2nd ed. (1969), are good general surveys. For a comprehensive selection of standard works from Locke to J.S. Mill, see A.J. AYER and R. WINCH (eds.), *The British Empirical Philosophers* (1952). Modern works in the Empiricist tradition include BERTRAND RUSSELL, *Human Knowledge* (1948); W.T. STACE, *Theory of Knowledge and Existence* (1932); RUDOLF CARNAP, *Der logische Aufbau der Welt* (1928; Eng. trans., *The Logical Structure of the World*, 1967); and A.J. AYER, *Language, Truth and Logic*, 2nd ed. (1946), a short exposition of the extreme Empiricist position. HAROLD MORICK (ed.), *Challenges to Empiricism* (1980), is a collection of essays.

Existentialism: Fundamental texts include SOREN KIERKEGAARD, *Philosophiske smuler* (1844; *Philosophical Fragments*, 1936); *Afsluttende uvidenskabelig efterskrift til de philosophiske smuler* (1846; *Concluding Unscientific Postscript*, 1941). FRIEDRICH NIETZSCHE, *Also sprach Zarathustra* (1883–84; *Thus Spoke Zarathustra*, trans. by W. KAUFMANN in *The Portable Nietzsche*, 1954); *Zur Genealogie der Moral* (1877; *Toward a Genealogy of Morals*, trans. by W. KAUFMANN in *Basic Writings of Nietzsche*, 1966). KARL JASPERS, *Psychologie der Weltanschauungen* (1919; the first work of contemporary Existentialism, announcing all of the fundamental theses; *Philosophie*, 3 vol. (1932; *Philosophy*, 3 vol., 1969–71); *Vernunft und Existenz* (1935; Eng. trans., *Reason and Existenz*, 1955). MARTIN HEIDEGGER, *Sein und Zeit* (1927; 10th ed., 1963; *Being and Time*, 1962); *Was ist Metaphysik?* (1929; "What Is Metaphysics?" trans. by R.F.C. HULL and A. CRICK in *Existence and Being*, 1949); *Einführung in die Metaphysik* (1953; *An Introduction to Metaphysics*, 1959). GABRIEL MARCEL, *Être et Avoir* (1935; *Being and Having*, 1949); *Le Mystère de l'Être*, 2 vol. (1950; *The Mystery of Being*, 1950–51); *The Philosophy of Existence*, (1949). NICOLA ABBAGNANO, *La struttura dell'esistenza* (1939); N. LANGIULLI (ed. and trans.), *Critical Existentialism* (1969). JEAN-PAUL SARTRE, *L'Être et le néant* (1943; *Being and Nothingness*, 1956); *L'Existentialisme est un humanisme* (1946; *Existentialism and Humanism*, 1948). MAURICE MERLEAU-PONTY, *La Structure du comportement* (1942; *The Structure of Behavior*, 1963); *Phénoménologie de la perception* (1945; *The Phenomenology of Perception*, 1962).

Important surveys and analyses include MARJORIE GRENE, *Dreadful Freedom* (1948); HELMUT KUHN, *Encounter with Nothingness: An Essay on Existentialism* (1951); MAURICE NATANSON, *A Critique of Jean-Paul Sartre's Ontology* (1951); JAMES COLLINS, *The Existentialists: A Critical Study* (1952); WILLIAM BARRETT, *What Is Existentialism?* (1964); ADOLPH LICHTIGFELD, *Jaspers' Metaphysics* (1954); MARJORIE GRENE, *Martin Heidegger* (1957); MARY WARNOCK, *Existentialism* (1970); and ROBERT D. CUMMING, *Starting Point* (1979).

Idealism: The two best books on Idealism in English are A.C. EWING, *Idealism: A Critical Survey* (1933); and R.F.A. HOERNLE, *Idealism as a Philosophy* (1927). A.C. EWING (ed.), *The Idealist Tradition from Berkeley to Blanshard* (1957), is a useful volume containing selections from the texts of the major Idealists and selected criticisms of Idealism. Other important collections are C. BARRETT (ed.), *Contemporary Idealism in America* (1932); J.H. MUIRHEAD (ed.), *Contemporary British Philosophy: Personal Statements*, First–Second Series, 2 vol. each (1924); and G.P. ADAMS and W.P. MONTAGUE (eds.), *Contemporary American Philosophy*, 2 vol. (1930). Works on the history and theory of the subject include: B. BLANSHARD, *Reason and Analysis* (1962), a careful critical examination of schools of philosophy opposed to Idealism; G.W. CUNNINGHAM, *The Idealistic Argument in Recent British and American Philosophy* (1933), a thorough and dependable treatise; S.N. DASGUPTA, *Indian Idealism* (1933); NICOLAI HARTMANN, *Die Philosophie des Deutsches Idealismus*, vol. 1, *Fichte, Schelling, und die Romantik* (1923), vol. 2, *Hegel* (1929), 2nd ed. (1 vol., 1960); A.J.M. MILNE, *The Social Philosophy of English Idealism* (1962); J.H. MUIRHEAD, *The Platonic Tradition in Anglo-Saxon Philosophy: Studies in the History of Idealism in England and America* (1931); P.T. RAJU, *Idealistic Thought of India* (1953); L.S. ROUNER (ed.), *Philosophy, Religion, and the Coming World Civilization: Essays in Honor of William Ernest Hocking* (1966); and A. STERN, *Philosophy of History and the Problem of Values* (1962), a good secondary source.

On the classical systems of Indian philosophy, see S. RADHAKRISHNAN, *Indian Philosophy*, 2 vol. (1923–27), an authoritative exposition; S.N. DASGUPTA, *A History of Indian Philosophy*, 5 vol. (1922–55), an erudite examination of the Sanskrit and Pāli texts; *Indian Idealism* (1933); SRI AUROBINDO, *The Life Divine* (1949), a reinterpretation of the heritage; and SWAMI VIVEKANANDA, *The Yogas and Other Works* (1953), a collection with biography. In *Pragmatism* (1907) and in *The Varieties of Religious Experience* (1902), WILLIAM JAMES quoted Vivekananda at length as a typical representative of spiritual monism. In the *Sacred Books of the East*, 50 vol., ed. by F. MAX MULLER (1879–1910), see especially vol. 1 and 15, *The Upanishads* and vol. 8, *The Bhagavadgītā*. In the *Harvard Oriental Series*, 47 vol., ed. by CHARLES ROCKWELL LANMAN (1890–1968), see especially vol. 3, *Buddhism in Translations*, and vol. 17, *The Yoga-System of Patañjali*.

Writings of Idealists are listed in their biographies. P.A. SCHILPP, editor of "The Library of Living Philosophers Series," has issued four important volumes on Idealists: *The Philosophy of Alfred North Whitehead*, 2nd ed. (1951); *The Philosophy of Ernst Cassirer* (1949); *The Philosophy of Sarvepalli Radhakrishnan* (1962); and *The Philosophy of Martin Buber* (1967); and a volume on Brand Blanshard is in preparation. Other works on Idealists include: J.H. COTTON, *Royce on the Human Self* (1954), one of the best books on Royce; D.S. ROBINSON, *Royce and Hocking: American Idealists* (1968); and GABRIEL MARCEL, *La Métaphysique de Royce* (1945; Eng. trans., 1956), which shows the relation of Marcel to Royce and Hocking. H.T. KIM, "Nishida and Royce," *Philosophy East and West*, 1:18–29 (1952), and "The Logic of the Illogical: Zen and Hegel," *ibid.*, 5:19–29 (1955), are informative accounts of Nishida's contributions to Idealism. The following volumes of the *Revue Internationale de Philosophie* contain valuable bibliographies: *Henri Bergson*, vol. 3 (1948); *Léon Brunschvicg*, vol. 5 (1951); *Hegel*, vol. 6 (1952); *George Berkeley*, vol. 7 (1953); *Benedetto Croce*, vol. 7 (1953); *Kant*, vol. 8 (1954); *Whitehead*, vol. 15 (1961); *Leibniz*, vol. 20 (1966); and *Josiah Royce*, vol. 21 (1967). GODFREY VESEY (ed.), *Idealism Past and Present* (1982), includes historical essays on idealistic thought of the past 2,500 years.

Materialism: For the period up to about 100 years ago, see F.A. LANGE, *Geschichte des Materialismus und Kritik seiner Bedeutung in der Gegenwart*, 2 vol. (1902; Eng. trans., *History of Materialism and Criticism of Its Present Importance*, 3rd ed., 3 vol. (1925). For more recent times, see JOHN PASSMORE, *A Hundred Years of Philosophy*, 2nd ed. (1966). There are excellent articles by KEITH CAMPBELL and H.B. ACTON on "Materialism" and "Dialectical Materialism" in the *Encyclopedia of Philosophy*, vol. 5, pp. 179–188 and vol. 2, pp. 389–397 (1967), which also contains excellent articles on nearly all of the noncontemporary philosophers here mentioned. Examples of work by most of the contemporary writers are given in JOHN O'CONNOR (ed.), *Modern Materialism: Readings on Mind-Body Identity* (1969); and C.V. BORST (ed.), *The Mind-Brain Identity Theory* (1970). See also HERBERT FEIGL, *The "Mental" and the "Physical": The Essay and a Postscript* (1967); J.J.C. SMART, *Philosophy and Scientific Realism* (1963); D.M. ARMSTRONG, *A Materialist Theory of the Mind* (1968); and WILFRID SELLARS, *Science, Perception and Reality* (1963), especially ch. 1. A rather difficult book defending Materialism from the difficulties about intentionality is D.C. DENNETT, *Content and Consciousness* (1969). For two very different styles of antimaterialist argument, see J.R. LUCAS, *The Freedom of the Will* (1970); and NORMAN MALCOLM, *Problems of Mind* (1971). Another interesting critique of Materialism is in JOHN BELOFF, *The Existence of Mind* (1962). A mainly mechanistic philosophy of biology is presented by the German biologist B. RENSCH in *Biophilosophie auf erkenntnistheoretischer Grundlage (Panpsychistischer Identismus)* (1968; Eng. trans., *Biophilosophy*, 1971), though Rensch's philosophy is also panpsychist.

Some classic Materialist works are: LUCRETIUS, *On the Nature of the Universe*, trans. by R.E. LATHAM (1951); THOMAS HOBBES, *Body, Mind and Citizen: Selections*, ed. by R.S. PETERS (1962); RENE DESCARTES, *Philosophical Writings*, trans. and ed. by ELIZABETH ANSCOMBE and P.T. GEACH (1954); and A. VARTANIAN,

La Mettrie's "L'Homme Machine": A Study in the Origins of an Idea (1960), which is a critical edition with introductory monograph and notes.

For epistemic Materialism, see RUDOLF CARNAP, "Psychology in Physical Language," in A.J. AYER (ed.), *Logical Positivism* (1959); Carnap's replies to Herbert Feigl and A.J. Ayer in P.A. SCHILPP (ed.), *The Philosophy of Rudolf Carnap* (1964); and H. REICHENBACH, *Experience and Prediction* (1938).

The most relevant and important works by Ryle and Wittgenstein are: GILBERT RYLE, *The Concept of Mind* (1949); and LUDWIG WITTGENSTEIN, *Philosophical Investigations* (1953). For a Materialist critique of Ryle, see BRIAN MEDLIN, "Ryle and the Mechanical Hypothesis," in C.F. PRESLEY (ed.), *The Identity Theory of Mind*, 2nd ed. (1971).

Phenomenology: Most of the classical works on Phenomenology were written by Husserl himself (see the bibliography of the HUSSERL article). In EDMUND HUSSERL, *Logische Untersuchungen*, 2nd ed., 2 vol. (1913–21; Eng. trans., *Logical Investigations*, 2 vol., 1970), one of the fundamental texts on Phenomenology, the phenomenological method is applied in the area of logic. The following works appeared in the journal *Jahrbuch für Philosophie und phänomenologische Forschung* (1913–30); Husserl's *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie* (1913; Eng. trans., *Ideas: General Introduction to Pure Phenomenology*, 1931, reprinted 1969), through which Phenomenology established itself as transcendental philosophy and received worldwide reaction; *Vorlesungen zur Phänomenologie des inneren Zeitbewusstseins* (1928; Eng. trans., *The Phenomenology of Internal Time-Consciousness*, 1964); and *Formale und transzendentale Logik* (1929; Eng. trans., *Formal and Transcendental Logic*, 1969); MAX SCHELER, *Der Formalismus in der Ethik* (1916); and MARTIN HEIDEGGER, *Sein und Zeit* (1927; Eng. trans., *Being and Time*, 1962). The following works appeared in the series *Husserliana* (Husserl's collected works): *Cartesianische Meditationen und Pariser Vorträge* (1950; Eng. trans., *The Paris Lectures*, 1964), which contains the text of the Paris lectures of 1929 and the subsequent elaboration; *Die Idee der Phänomenologie* (1950: Eng. trans., *The Idea of Phenomenology*, 1964), an introduction to Phenomenology in five lectures from 1907; *Ideen*, 3 vol. (1950–52); *Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie* (1954; Eng. trans., *The Crisis of European Sciences and Transcendental Phenomenology*, 1970), Husserl's later work (1934–37), significant for the problems regarding the life-world; *Erste Philosophie*, 2 vol. (1956–59), a critical history of ideas and a theory of reduction presented in a series of lectures, 1923–24; *Phänomenologische Psychologie* (1962), lectures from 1925, a different wording of the *Encyclopædia Britannica* article of 1927, with remarks by Heidegger and the Amsterdam addresses of 1925; *Zur Phänomenologie des inneren Zeitbewusstseins* (1966), text on the problem of time from 1893–1917 with lectures (in the middle) from 1905 on the inner time-consciousness (first edited by Heidegger); *Analysen zur passiven Synthesis* (1966), a phenomenological analysis having sensation as its subject matter; and *Philosophie der Arithmetik* (1970), early manuscripts from 1890 to 1901. His *The Basic Problems of Phenomenology*, trans. by ALBERT HOFSTADTER (1982; originally published in German., 1975), reproduces a course of lectures given in 1927. See also JOSEPH J. KOCKELMANS, *Edmund Husserl's Phenomenological Psychology* (1967); and RUDOLF BOEHM, *Vom Gesichtspunkt der Phänomenologie* (1968).

In the series "Phaenomenologica" works are published that are written from a Phenomenological perspective, HERBERT SPIEGELBERG, *The Phenomenological Movement*, 2nd ed., 2 vol. (1965), worthy of particular mention, also appeared in this series. Since 1904 MARVIN FARBER has edited the journal *Philosophy and Phenomenological Research* (by no means, however, exclusively dedicated to Phenomenology). In England *The Journal of the British Society for Phenomenology* first appeared in 1970. In the United States the following journals have appeared: *Journal of Phenomenological Psychology* (semi-annual); *Research in Phenomenology* (1971); and *Analecta Husserliana: The Yearbook of Phenomenological Research.*

Positivism and Logical Empiricism: Literature on Classical Positivism includes J. WATSON, *Comte, Mill and Spencer* (1895); W.M. SIMON, *European Positivism in the Nineteenth Century* (1963); JOHN STUART MILL, *Auguste Comte and Positivism* (1865); AUGUSTE COMTE, *Cours de philosophie positive*, 6 vol. (1830–42; Eng. trans. and cond. by H. MARTINEAU, *The Positive Philosophy of Auguste Comte*, 2 vol., 1853). For further references to the ethical views of the classical Positivists, see the bibliography of Utilitarianism in the *Encyclopedia of Philosophy.*

Many of the original classics of Logical Positivism and Logical Empiricism, both books and articles, are listed in the ample bibliography of A.J. AYER (ed.), *Logical Positivism* (1959), an anthology that contains, among other important essays, RUDOLF CARNAP'S "Psychology in Physical Language." The early history of Viennese Positivism is well told in VICTOR KRAFT, *Der Wiener Kreis: Der Ursprung des Neopositivismus* (1950. 2nd ed. 1968; Eng. trans., *The Vienna Circle*, 1953, reprinted 1969). Another important source is J. JOERGENSEN, *The Development of Logical Empiricism*, vol. 2, no. 9 of the *International Encyclopedia of Unified Science* (1951). For a brief account of the European movement of Logical Positivism and its migration and impact in the United States, see H. FEIGL, "The Wiener Kreis in America," in D. FLEMING and B. BAYLIN (eds.), *The Intellectual Migration: Europe and America 1930–1960* (1969). Books, mainly in the foundations of the sciences, but also in philosophy of language and epistemology, many by the leading Logical Empiricists, are listed in the ample *Bibliography and Index*, in HERBERT FEIGL and CHARLES MORRIS (eds.), *International Encyclopedia of Unified Science*, vol. 2, no. 10 (1969). Of direct relevance are the major works of R. Carnap, O. Neurath, M. Schlick, P. Frank, H. Reichenbach, E. Nagel, C.G. Hempel, R. von Mises, and Charles Morris. For criticisms, those of KARL R. POPPER may be used; and the intellectual autobiography of Carnap, the 26 descriptive and critical essays, and his replies, in P.A. SCHILPP (ed.), *The Philosophy of Rudolf Carnap* (1963). For more recent evaluations and reactions, see P. ACHINSTEIN and S.F. BARKER (eds.), *The Legacy of Logical Positivism: Studies in the Philosophy of Science* (1969); and *Minnesota Studies in the Philosophy of Science*, vol. 1–5 (1956–70).

Pragmatism: Classic works include CHARLES SANDERS PEIRCE, "The Fixation of Belief," "How to Make Our Ideas Clear," and "What Pragmatism Is," in *Collected Papers*, vol. 5, ed. by C. HARTSHORNE and P. WEISS (1934); WILLIAM JAMES, *Principles of Psychology*, 2 vol. (1890), *The Will to Believe and Other Essays in Popular Philosophy* (1897), *Pragmatism: A New Name for Some Old Ways of Thinking* (1907), and *The Meaning of Truth* (1909); JOHN DEWEY, *How We Think* (1910), *The Influence of Darwin on Philosophy* (1910), *Democracy and Education* (1916), *Essays in Experimental Logic* (1916), *Reconstruction in Philosophy* (1920, 1948), *Human Nature and Conduct* (1922), *Logic: The Theory of Inquiry* (1938), *Theory of Valuation* (1939), and *Problems of Men* (1946). On F.C.S. Schiller see R. ABEL, *The Pragmatic Humanism of F.C.S. Schiller* (1955), with a bibliography of Schiller's writings; on French and Italian pragmatists, H.S. THAYER, *Meaning and Action: A Critical History of Pragmatism*, part 3 (1968), with further bibliographical references.

For surveys of the movement, see H.S. THAYER, *Meaning and Action: A Critical History of Pragmatism* (1968), with bibliography; "Pragmatism," in D.J. O'CONNOR (ed.), *A Critical History of Western Philosophy*, pp. 437–462 (1964); and H.S. THAYER (ed.), *Pragmatism: The Classic Writings* (1970), the basic writings in the Pragmatism of Peirce, James, Dewey, Mead, and Lewis, and further bibliographical references; John Dewey, "The Development of American Pragmatism," in *Philosophy and Civilization*, pp. 13–35 (1931); and CHARLES W. MORRIS, *The Pragmatic Movement in American Philosophy* (1970). JACQUES BARZUN, *A Stroll with William James* (1983), is an excellent discussion of his ideas.

Rationalism: The classic Ancient Greek work is PLATO, *Meno;* essential modern works are DESCARTES, *Meditationes de Prima Philosophia . . .* (*Meditations on First Philosophy*); SPINOZA, *Ethics;* LEIBNIZ, *Monadologie* (*Monadology and Other Philosophical Writings*); KANT, *Kritik der reinen Vernunft* (*Critique of Pure Reason*); 19th century. Rationalism is epitomized in HEGEL, *Phänomenologie des Geistes* (*Phenomenology of Mind*); FRANCIS HERBERT BRADLEY, *Appearance and Reality.*

For Rationalism in the theory of knowledge, see BRAND BLANSHARD, *Reason and Analysis* (1962); GEORGE BOAS, *Rationalism in Greek Philosophy* (1961); ERNST CASSIRER, *Die Philosophie der Aufklärung* (1932; Eng. trans., *Philosophy of the Enlightenment*, 1951); M.R. COHEN, *Reason and Nature: An Essay on the Meaning of Scientific Method*, 2nd ed. (1953): A.C. EWING, *Idealism: A Critical Survey* (1934); H.H. JOACHIM, *The Nature of Truth* (1906); A.E. MURPHY, *The Uses of Reason* (1943); H.J. PATON, *In Defence of Reason* (1951); BERTRAND RUSSELL, *Problems of Philosophy* (1912); W.H. WALSH, *Reason and Experience* (1947).

For Rationalism in metaphysics, see the classics listed above. For two outstanding examples from the present century, see J.M.E. MCTAGGART, *The Nature of Existence*, 2 vol. (1921–27), together with the commentary of C.D. BROAD, *Examination of McTaggart's Philosophy*, 2 vol. (1933–38); and ALFRED NORTH WHITEHEAD, *Process and Reality* (1929). For Rationalism in ethics, see WILLIAM WOLLASTON, *The Religion of Nature Delineated* (1722); and KANT, *Die Metaphysik der Sitten* (1785; Eng. trans., *The Metaphysics of Morals*, 1799). For early forms of the appeal to self-evident rules, see RICHARD PRICE, *A Review of the Principal Questions in Morals* (1758). For later types of Rationalism, see G.E. MOORE, *Principia Ethica* (1903); W.D.

ROSS, *The Right and the Good* (1930), and *Foundations of Ethics* (1939); BRIAN ELLIS, *Rational Belief System* (1979).

For Rationalism in religion, excellent standard works are: W.E.H. LECKY, *History of the Rise and Influence of the Spirit of Rationalism in Europe,* 2 vol. (1865); A.D. WHITE, *History of the Warfare of Science with Theology in Christendom,* 2 vol. (1910); J.M. ROBERTSON, *A Short History of Freethought, Ancient and Modern,* 2nd ed., 2 vol. (1906); A.W. BENN, *History of English Rationalism in the Nineteenth Century,* 2 vol. (1906); J.B. BURY, *A History of Freedom of Thought* (1913). SIGMUND FREUD, *Die Zukunft einer Illusion* (1927; Eng. trans., *The Future of an Illusion,* 1928), offers a psychoanalytic study of religious belief.

Utilitarianism: The classical texts are JEREMY BENTHAM, *An Introduction to the Principles of Morals and Legislation* (1789; 2nd ed., 1823); JOHN STUART MILL, *Utilitarianism* (1861; 4th ed., 1871); HENRY SIDGWICK, *The Methods of Ethics* (1874; 7th ed., 1907); G.E. MOORE, *Principia Ethica* (1903) and *Ethics* (1912); D.D. RAPHAEL (ed.), *British Moralists: 1650–1800,* 2 vol. (1969), containing selected readings.

Useful anthologies include MARY PETER MACK (ed.), *A Bentham Reader* (1969); J.B. SCHNEEWIND (ed.), *Mill's Ethical Writings* (1965); SAMUEL GOROVITZ (ed.), *Utilitarianism: John Stuart Mill with Critical Essays* (1971).

Secondary, historical, and contemporary studies include DAVID LYONS, *Jeremy Bentham* (1972) ERNEST ALBEE, *A History of English Utilitarianism* (1902, reprinted 1957); LESLIE STEPHEN, *The English Utilitarians,* 3 vol. (1900, reprinted 1968); ELIE HALEVY, *La Formation du radicalisme philosophique,* 3 vol. (1901–04; Eng. trans., *The Growth of Philosophical Radicalism,* 1928); J.P. PLAMENATZ, *The English Utilitarians* (1949); J.B. SCHNEEWIND (ed.), *Mill: A Collection of Critical Essays* (1968); STEPHEN TOULMIN, *An Examination of the Place of Reason in Ethics* (1950); P.H. NOWELL-SMITH, *Ethics,* ch. 16 (1954); RICHARD BRANDT, *Ethical Theory,* ch. 12–15 (1959); DAVID LYONS, *Forms and Limits of Utilitarianism* (1965); JAN NARVESON, *Morality and Utility* (1967); MICHAEL D. BAYLES (ed.), *Contemporary Utilitarianism* (1968); J.J.C. SMART, *An Outline of a System of Utilitarian Ethics* (1961). DONALD REGAN, *Utilitarianism and Cooperation* (1980), a presentation of a new utilitarian theory with a good survey of disputes among utilitarians.

Philosophies of the Branches of Knowledge

It is no longer possible for any one individual to be a leader in all branches of learning. Perhaps Leibniz, who died in 1716, was the last universal scholar, and his life is a warning rather than an example. In such a situation the enterprise of philosophy is in a sense more difficult than ever before; no one can hope to contribute much to the clarification of ideas in logic who is not himself a logician, or to the clarification of ideas in physics who is not himself a physicist, and so on throughout the fields of specialized knowledge. It thus seems inevitable that there should have evolved philosophical disciplines corresponding to the various branches of study and providing the frameworks upon which they may be studied and evaluated in and of themselves. (Discussions of particular schools of philosophy mentioned in this article are found in PHILOSOPHICAL SCHOOLS AND DOCTRINES, WESTERN.) (Ed.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 10/52, and the *Index*.

This article is divided into the following sections:

Philosophy of science 652
  Nature, scope, and relations of the subject 652
  Historical development 653
  Elements of scientific enterprise 658
  Movements of scientific thought 662
  Philosophical status of scientific theory 665
  Interrelationships of science and culture 667
Philosophy of nature 669
  Philosophy of physics 669
  Philosophy of biology 675
Philosophy of history 679
  Speculative theories 679
  Analytical problems 682
Philosophy of religion 683
  Religion as a fact in human experience, culture, and history 683
  Views with transcendent references 684
  Views with anthropic references 685
  The view from within as privileged 685
  The dimension of religion for insiders 685
  Internal criticisms of religion 686
  The rejection of religion or religiousness 686
  The acknowledgment of religion or religiousness as valid 688
  History of the philosophy of religion 689
  Basic themes and problems in the philosophy of religion 692
  The present situation in the philosophy of religion 694
Philosophy of art 694
  Distinguishing characteristics 694
  The interpretation of art 695
  The mediums of art 697
  Art as imitation (representation) 698
  Art as expression 701
  Art as form 703
  Pragmatic theories of art 705
Philosophy of logic 710
  Logic as a discipline 710
  Issues and developments in the philosophy of logic 712
  Logic and other disciplines 713
The nature of mathematics 714
  Critical questions 715
  Definition of mathematics 716
  Applied mathematics 717
Western philosophy of law 718
  Problems of the philosophy of law 718
  Historical survey of legal theories 719
  The status of contemporary philosophy of law 724
Philosophy of education 725
  Common assumptions about human learning 725
  Differing conceptions of education 726
Bibliography 729

Philosophy of science

Taken broadly as the progressive improvement of man's understanding of nature, the intellectual enterprise of science originally formed an integral part of philosophy, and the two areas of inquiry have never finally separated. Little more than a hundred years ago, theoretical physics—concerned with the fundamental debate about physical nature—was still described as "natural philosophy," as distinguished from the two other chief divisions of abstract discussion, viz., moral philosophy and metaphysical philosophy—the latter including ontology, the study of the deepest nature of reality or being. In fact, only during the 20th century, following the professionalization and specialization of the natural sciences, did the philosophy of science become recognized as a separate discipline.

Methodological and epistemological issues—*i.e.*, issues regarding the investigator's manner of approach to nature—are treated in this section. Issues regarding the substantive character of nature as so revealed—*i.e.*, as it is in and of itself—are treated below in *Philosophy of nature*.

NATURE, SCOPE, AND RELATIONS OF THE SUBJECT

The philosophy of science attempts first to elucidate the elements involved in scientific inquiry—observational procedures, patterns of argument, methods of representation and calculation, metaphysical presuppositions—and then to evaluate the grounds of their validity from the points of view of formal logic, practical methodology, and metaphysics. The philosophy of science is thus a topic for explicit analysis just as are other subdivisions of philosophy. The boundaries between these subdivisions are, at certain points, somewhat arbitrary; it is not easy to separate completely, for example, the philosophical validation of scientific hypotheses from the formal study of inductive logic (which reasons from facts to general principles), or the debate about observation in philosophy of science from that in epistemology (the theory of knowledge).

Historically, the preoccupations of those who, if living today, would be called philosophers of science have been of two main kinds—ontological and epistemological, or epistemic. This division reflects a long-standing distinction between object and subject—*i.e.*, between nature, regarded as that about which man sets out to acquire scientific knowledge, and man himself, regarded as the creator and either discoverer or possessor of that knowledge. Since 1920, new directions within physics, particularly in quantum mechanics, have discredited any hard-and-fast distinction between the knower and the known or between the observer and his observation. Nevertheless, the distinction remains relevant on an everyday level and can be cautiously retained for the purposes of initial exposition.

Ontological versus epistemic concern

The ontological preoccupations of philosophers of science frequently overlap into the substantive areas of the sciences themselves, for they explore the general problem,

> What kinds of entities and elements or theoretical terms can properly figure in man's scientific theories? And what sort of existence, or other objective status, do such things possess?

In particular cases, this general problem has inevitably raised questions of substance as well as of intellectual method. An early 20th-century debate between two Austrian physicists, Ernst Mach and Ludwig Boltzmann, and a German physical chemist, Wilhelm Ostwald, about the existence and reality of atoms, for instance, involved both substantive issues of physics and chemistry and philosophical issues of a more strictly analytical kind.

Until recently the epistemic concerns of the subject have been more purely philosophical, though this autonomy is now being challenged by studies in psychology that explore cognitive processes and others in sociology that examine the conditioning of cognition through interpersonal and group relationships. Epistemologically, philosophers of science have analyzed and evaluated the concepts and methods employed in studying natural phenomena and human behaviour, whether individual or collective; this analysis has covered the general concepts and methods characteristic of all scientific inquiries and also the more particular ones that distinguish the subject matters and problems of different special sciences. In treating the epistemic issues that arise about science and scientific procedures, the emphasis in this article is placed upon their consideration in general terms; the concepts and methods peculiar to a discipline, say, sociology, are discussed elsewhere.

With its vast range of concerns, the philosophy of science has attracted the attention of men with very different professional backgrounds and interests. At one extreme, as in the writings of Ernest Haeckel, the German Darwinian evolutionist, it has merged into a sweeping kind of popular science; at an opposite extreme, as in 20th-century Logical Positivism or Logical Empiricism (see below), which holds that knowledge is only what is scientifically verifiable, it has been treated as an extension of formal logic and conceptual analysis. Between these extremes, as in the work of the British astrophysicist Arthur Eddington and the German quantum physicist Werner Heisenberg, the philosophy of science has moved to the frontiers of the sciences and has directly confronted problems about the existence, status, and validity of theoretical entities and concepts.

Correspondingly, the philosophy of science has been approached in very diverse spirits, ranging from the highly abstract and mathematical to the concrete and historical and from the severely positivistic to the frankly theological. From René Descartes in the 17th century to the Logical Positivist Otto Neurath in the 20th, the success of pure mathematics and logic has inspired the mathematically minded to cast the whole of natural science into a single formal system after the pattern of geometry. Their opponents—from John Locke, an 18th-century British Empiricist, to N.R. Hanson, a recent U.S. philosopher of science—have sought the proper basis of man's intellectual confidence in the nature of scientific investigation, regarded as a human activity. Equally, Positivists such as Hans Reichenbach, a 20th-century German-American philosopher, have looked to philosophy for proof that scientific inquiries alone can provide knowledge worthy of the name; while theists such as Pierre Duhem, a French theoretical physicist, have argued that the claims of science are inherently limited and so leave room for other, more embracing varieties of metaphysical and religious truth.

Relations with other disciplines

This diversity of concerns and approaches within the philosophy of science has affected its relation with other neighbouring disciplines. On a practical plane, for instance, different philosophical interpretations have implied different procedures for testing and assessing the strength of rival concepts and hypotheses. Thus, no clear dividing line can be drawn between the philosophical analysis of scientific theories and the statistical analysis of scientific procedures and experiments or, notably, between the philosophy of science and the history of scientific ideas. Recent debates indicate that the particular questions that a historian of science brings to his analysis of scientific change inescapably depend on his philosophical attitudes and commitments.

On a more general and abstract level, the philosophy of science has never been definitively separated from metaphysics and epistemology. Indeed, some 20th-century philosophers—for instance, the U.S. philosophical logician Willard V. Quine—effectively restrict the legitimate areas of metaphysics and epistemology to what have here been called the ontal and epistemic aspects of the philosophy of science. In Quine's view, the traditional ontological problem of what there is in the world as man knows it must be attacked by a logical analysis of the claims about what kinds of things exist that are implicit in alternative theoretical systems. Meanwhile, the work of such cognitive psychologists as Switzerland's Jean Piaget, who explored the processes involved in the genesis of knowledge, is eroding the barriers between the logical analysis of conceptual systems, the psychological investigation of thought processes, and the epistemological validation of intellectual procedures. Piaget, for instance, based his investigations into the acquisition of concepts on a philosophy akin to that of Kant, who held that all knowledge bears the imprint of the mind's own structure; and, though a psychologist himself, Piaget referred to certain aspects of his work as "genetic epistemology," a philosophical designation.

The present survey of the philosophy of science, therefore, contains no effort to prejudge the central question, whether the methods of logical analysis alone are legitimate or whether at certain points the subject legitimately overlaps into such neighbouring subjects as cognitive psychology, the history of science, and epistemology. Philosophers of science themselves have been sharply divided, some rejecting any alliance except with logic, others cultivating its wider historical and behavioral connections; and both points of view must thus be taken into account.

HISTORICAL DEVELOPMENT

The Pre-Socratics

Classical and medieval periods: the beginnings of a philosophy of nature. At the outset the problems of science were as much those of method as of substance and were inseparable from those of what has long been called natural philosophy. The first attempts to move beyond traditional mythologies to a rational account of nature, beginning with the Ionian and south Italian philosophers around 600 BC, involved the sorts of elements that any such account should comprise. Detailed empirical, or observational, considerations favouring one or another of the rival accounts were premature. Were, for example, the diverse phenomena of nature manifestations of one single enduring form of matter or of several elementary substances mixed together, and was the fundamental substance continuous and fluid-like, or discrete and atomic? Others asked whether the observed forms of phenomena were evidence, rather, of some universal, underlying mind or of a variety of coexisting kinds of spirit responsible for phenomena having different orders of complexity.

The Pre-Socratics based their answers at least as much on epistemic grounds—*i.e.,* on what type of account would be genuinely intelligible—as on ontal or empirical ones—*i.e.,* on what sorts of enduring entities could possibly have, or be found in experience to have, the required kind of existence. Their answers ranged from the ontal Realism of Parmenides, foremost philosopher of Eleaticism—a school of southern Italy according to which all changes are transitory appearances concealing the mutual relations of deeper, unchanging realities—to the critical skepticism of Heracleitus—the Ephesian philosopher according to whom nothing in nature as man knows it can ever have this Parmenidean reality, and everything empirical is in flux.

Plato and Aristotle

Though Plato and Aristotle displayed more precise concern for actual cases, their philosophies of science still rested on the same mixture of ontological, epistemological, and empirical considerations. Questions about nature discussed in Plato's *Timaeus* and Aristotle's *Physics,* for instance, were neither purely metaphysical nor purely empirical in character though they had a methodological aspect akin to that of modern philosophy of science. Moreover, Plato's construction of the fundamental theories of science around concepts and patterns borrowed from geometry have had a profound influence in the modern period—upon René Descartes, for example, in the 17th century and upon the founder of modern logic, the German mathematician Gottlob Frege, in the 1880s and after.

Mathematical entities alone, Plato argued, have the sort

of enduring intelligibility that Parmenides had rightly demanded of the ultimate constituents in a rational science of nature. Thus, only a physical theory built on a numerical and geometrical framework will reveal the truly permanent structures and relationships behind the evident flux of phenomena. Within such a theory, all inferences will be self-evidently valid at all times and so exempt from the mutability of empirical events; correspondingly, the numbers and figures of formal mathematics will have an immutability denied to familiar physical objects. Planetary astronomy and the theory of matter were, in Plato's view, scientific fields within which this mathematical methodology showed immediate promise; the movements of the planets must be explained by constructions drawn from three-dimensional geometry, and the physics of matter seemingly involved atoms with shapes reflecting the geometry of the five regular solids (the tetrahedron, dodecahedron, etc.). In either case the mathematical theories themselves would alone be fully exact and intelligible, whereas empirical objects and processes could be no more than transitory and approximate illustrations of the enduring entities and theoretical relations underlying them.

Given that Aristotle's scientific preoccupations were centred on marine biology rather than planetary movements, he naturally developed a very different scientific methodology. In his view, mathematical entities and relations were too completely general and too remote from actual experience to explain the qualitative details of empirical entities. So the ultimate elements of nature must be not Plato's abstract mathematical forms that supposedly existed apart from actual phenomena but rather certain more specific entities, recognizable within the familiar sequences of empirical experience. Instances of such basic essences could be discovered in the typical life cycles of different creatures; for example, the morphogenesis of a seed exemplifies the "coming into being" of the corresponding type of animal or plant, of which the mature specific form—as defined by its essence—is the natural destination of its development. Having recognized the natural destinations toward which natural processes of different kinds were directed, it was then possible to construct a comprehensive classification of essences in terms of which the whole natural world would, in principle, be intelligible. Explanations within such an all-embracing natural history might not be self-evidently general and immutable, as were those of Plato's geometry, but the theoretical inferences involved would be no less deductive or necessary. It would also account directly for the specific qualitative characters of different observed objects and processes.

Atomists and Stoics

The themes stated by Plato and Aristotle are still represented today by two rival approaches to the philosophy of science—one (Platonic) based in logic, the other (Aristotelian) based in the history of science. Between them they dominated the subject during the later period of Greek antiquity, otherwise notable only for the debate between the Atomist successors of Democritus and Epicurus and the Stoic philosophers, led by Zeno of Citium. This debate provided the first profound analysis of the strengths and weaknesses of atomistic explanation. Epicureans argued for a purely corpuscular view in which the individual units of matter moved quite independently, except when they were in actual contact. For the Stoics the empirical world was intelligible only in terms of interactions and stable patterns maintained by harmonies operative at a distance.

These two debates—between Platonists and Aristotelians and between Stoics and Epicureans—presented clearly and for the first time the chief alternative modes of explanation available to science and analyzed their possibilities and limitations in general terms. More than 2,000 years before the rise of modern thermodynamics and field theory, for instance, Aristotle had already recognized the difficulties of explaining changes in physical state (*e.g.*, melting and evaporation) within a purely atomistic theory of matter. Even earlier, Plato had demonstrated the possibility of a unified mathematical explanation of the differences between different kinds of material substance. In the 20th century the theoretical physicist Werner Heisenberg cited pre-Socratic arguments regarding the ultimate constitution of nature as relevant to contemporary problems.

Hellenistic through medieval times

By contrast with the period before Euclid the geometer (*i.e.*, to 300 BC), the ensuing Hellenistic, Islāmic, and medieval periods added little to the understanding of scientific methodology and explanation. From the Alexandrian astronomer Ptolemy, who detailed the geocentric theory, most natural philosophers deliberately restricted their intellectual claims in an instrumentalist manner—*i.e.*, by endeavouring merely to "save the phenomena" by devising successful mathematical procedures for predicting, for example, lunar eclipses and planetary motions. In this way they disregarded the mechanisms responsible for those phenomena, thus preserving the computational techniques of the sciences from the risk of conflict with theology for the following 1,250 years, until the time of Copernicus.

In the High Middle Ages, accordingly, the possibility that man could make himself the intellectual master of nature was largely neglected. Human understanding was dependent on God's illumination. The reliability of scientific knowledge lay not in the merits of its methodology but in the divine grace. Hence, man had no direct line of access to nature; the only road to knowledge was through the divine mind. Thus, all the central questions in the philosophy of science were restated as theological questions about the relationship between God's omniscience and the more limited knowledge of man. In this context the metaphor of "illumination" was taken so seriously in the 13th century that the subject of optics was cultivated by a distinguished Oxford scholar as much for its theological implications as for its physical content.

The Renaissance and after: from manifesto to critique. Although the intellectual Renaissance of the 16th and 17th centuries was accompanied by a secularization of learning, which shifted the centre of philosophical and scientific debate from monasteries to universities—and even to salons—the link between philosophy and theology was not abruptly snapped. Descartes, Newton, and Leibniz, the leading scholars of the time, were concerned to demonstrate that their positions were compatible with sound theology. Medieval controversies about human knowledge and divine grace found an echo in such arguments as Descartes's assertion that the rational methods of inquiry can be relied on only provided that God does not deliberately deceive us. Two new factors, however, combined to give the 17th-century debate about scientific methodology a new autonomy. First, philosophy now posed the central questions in the philosophy of science—both about the origins and functions of scientific concepts and about the structure and validity of scientific arguments—and faced them directly, instead of only as refracted through a theological prism. Second, these questions were acquiring an immediate relevance and significance, simply because men were then launching new, empirically based theories of nature with a seriousness unknown for some 1,200 years.

Between 1600 and 1800 the debate in the philosophy of science was barely separable from that within science itself. From Bacon and Galileo by way of Descartes and Leibniz to Laplace and Kant, the major participants in the philosophical debate played significant roles on the scientific stage as well. Thus, Bacon, author of the method of exhaustive induction (see below), and Descartes both attempted to formulate explicitly a new method for the improvement of the intellect—*i.e.*, codifying the rational procedures of science in a way that would free them from arbitrary and unfounded or superstitious assumptions (Bacon's idols) and ground them in a logically impregnable manner on the properties of "clear and distinct," or self-evidently valid, concepts (as distinguished by Descartes).

Francis Bacon and Descartes

To be sure, the two men offered different recipes for a rational science and described the outcome of a properly conducted scientific inquiry in quite different terms. On the one hand, Bacon was preoccupied with empirically observed facts as the starting point for all science and relied on theories only insofar as they were derived from those facts. Ideally, he held that the scientist should provide an exhaustive enumeration of all examples of the empirical phenomenon under investigation as a preliminary to identifying the natural "form" of which they were the manifestation. Though Bacon remained unclear about the exact character of the abstraction involved, he

is commonly assumed to have claimed that theoretical propositions in science are justified only if they have been deduced formally from such an enumeration.

In contrast to such so-called "Baconian induction," Descartes focussed upon the problem of constructing self-consistent and coherent deductive systems of theory, within which argument would proceed with the formal security familiar in Euclidean geometry. Whereas Bacon had reacted against the Scholastic reliance on Aristotle's authority by calling for a return to firsthand experience, Descartes reacted against the Skepticism of 16th-century humanists by pointing to mathematics as the pattern to which all genuinely certain knowledge about nature could aspire. Inasmuch as Euclid's axioms, definitions, and postulates had captured the intrinsic characteristics of spatial relations and provided a theoretical starting point from which the whole of geometry could be deductively inferred, the task for 17th-century physics was to extend Euclid's intellectual structure by adding further, equally self-evident axioms, definitions, and postulates. Only in this way could the theories of motion, magnetism, and heat—eventually those of physiology and cosmology, too—achieve the same necessary deductive authority. Descartes set out to show, in the four volumes of his *Principia Philosophiae* ("Principles of Philosophy"), how all of the familiar phenomena of physics could be accounted for by a single, fully comprehensive system of mathematical theory, based on Euclidean foundations and conforming to his own deductivist principles. The very possibility of so interpreting nature was so impressive to Descartes that it lent a "moral certainty" to his conclusions.

Newton's hypothetico-deductive method

The arguments of Bacon and Descartes were really manifestos; both offered intellectual programs for a natural science yet to be built, and, while it is true that during the next 150 years, Galileo, Newton, and many others actually constructed the new physical science for which the philosophers had been calling, it is also true that the form of the resulting theories was, nonetheless, not exactly what either man had foreseen. On the one hand, there was little Baconian induction in Newton's intellectual procedures. Those 17th-century scientists, such as Robert Boyle—one of the founders of modern chemistry—who seriously attempted to apply Bacon's maxims found his pedestrian advice to be more of a hindrance than a help in the formulation of illuminating theoretical concepts. (It was said, somewhat unkindly, that Bacon "philosophized like a Lord Chancellor.") On the other hand, though Newton was powerfully influenced by Descartes's mathematical example, he followed his methodological maxims only up to a point. Granted that the theory of motion and gravitation of Newton's *Principia* did indeed conform to Descartes's recipe—adding further dynamical axioms, definitions, and postulates to those of Euclid's geometry—Newton nonetheless made no pretense of proving, in advance of empirical evidence, that these additional assumptions were uniquely self-evident and valid. Instead, he treated them as working assumptions to be accepted hypothetically for just so long as their consequences threw light, in exact detail, on hitherto-unexplained phenomena. Inevitably, the epistemic claims to be made on behalf of such explanations fell short of Descartes's full "deductivist" ambitions. Newton knew of no phenomena, for instance, that evinced the mechanisms of gravitational attraction and saw no point in "feigning hypotheses" about them.

In this way, Newton devised in practice—almost inadvertently—what philosophers of science have since labelled the hypothetico-deductive method, in which, as theorized by Descartes, the proper form of a theory is seen as a mathematical system in which particular empirical phenomena are explained by relating them back deductively to a small number of general principles and definitions. The method, however, abandons the Cartesian claim that those principles and definitions can themselves be established, finally and conclusively, before inquiring what light their consequences throw on actual scientific problems and phenomena.

Empiricists, Rationalists, and Kant

From 1700 on, the terrain of debate in the philosophy of science shifted. At first, attacks on Newton's methods and assumptions by Leibniz, Berkeley, and the remaining Cartesians continued, from different points of view. But by 1740 the time for both manifestos and objections was past; the basic scientific soundness of Newton's concepts was no longer in doubt, and the philosophical question thus became retrospective, viz., How had Newton done it? Over this new question, 18th-century philosophers were divided into three camps: Empiricism, Rationalism, and Kantianism. There were those who believed, like the Scottish Skeptic David Hume, that Newton's philosophy conformed to the Empiricist maxims of Francis Bacon and John Locke. Again, there were those—like the Swiss mathematician Leonhard Euler, one of the founders of modern analysis, and Immanuel Kant in his younger days—who assumed that Newton's physical principles would eventually be put on a fully demonstrative or self-evident basis as required by Cartesian Rationalism. Neither of these positions proved entirely successful, as Kant himself came to recognize: the Empiricists failed to do justice to the deductive rigour of Newton's theoretical arguments; and the Rationalists could not rigorously demonstrate the mathematical uniqueness of Newton's system. As was already known, even Euclidean geometry, which involves the axiom of parallels (according to which one, and only one, line can be drawn through a given point parallel to another line) could no longer claim a formal uniqueness. It had been shown, in 1733 and again in 1766, that alternative geometrical systems can consistently be developed in which the axiom of parallels is replaced by other mathematically acceptable alternatives. Clearly, the authority claimed for Newton's concepts and methodology could no longer be sustained in the old Rationalist way; thus a third alternative, that of Kantianism, arose.

One of the prime goals of Kant's so-called critical philosophy, with its famous so-called transcendental method, in which knowledge reflects the categorial structure of the mind, was to provide an alternative philosophical justification of Newton's results. The system of concepts used in Euclidean geometry and Newtonian physics is uniquely relevant to man's actual experience, Kant argued, not because the empirical applicability of their principles is self-evident—no such self-evidence can tell the inquirer anything about external nature. Still less is it because their inductive support is so strong—no Baconian argument can yield the required kind of certainty. Rather, it is because the scientist can arrive at a coherent, rational system of empirically applicable explanations only by constructing his theories around just those (Euclidean and Newtonian) concepts. He could, in fact, go even further. Euclidean axioms are required, Kant claimed, not merely for science alone; they specify explicitly cognitive structures (of the mind) that are implicitly involved also—as so-called forms of intuition (specifically of space and time)—in the prescientific rational organization of sensory experience into a coherent, intelligible world of substantial objects seen as interacting by causal processes. A grasp of Kant's transcendental method would then enable a thinker to recognize (or so Kant hoped) how and in what respects the use of his established system of rational forms and categories is indispensable alike for any coherent understanding or even for any experience.

The assimilation of Kant

Through World War I: philosophy of classical physics. Kant's ambitious philosophical enterprise took a long time to digest. A century later, in the 1880s, philosophers of science as different in other ways as the Austrian phenomenalist Ernst Mach and Heinrich Hertz, pioneer in electromagnetic wave theory, were both pursuing questions opened up by Kant; and some of their implications were still being explored in the 1970s, as, for example, in cognitive psychology. In general terms, Kant's central thesis—*i.e.*, that man confers a structure on his knowledge through the concepts and categories that he brings to the formation and interpretation of experience—has proved extremely fertile: it has helped in the analysis of theory construction, and it has suggested in sensory psychology that man's very capacity for perception can yield effective knowledge only to the extent that his sensory inputs themselves have a cognitive or conceptual structure.

In one respect, to be sure, Kant seems in retrospect to have attempted too much. Including in man's framework

of sensory and intellectual organization all of Euclidean geometry and of fundamental Newtonian physics and the prescientific notions of substance and cause, Kant set out to demonstrate a priori that man's actual current framework is the one and only effective framework—a proof which, as is known today, was misguided, for its thesis is simply not the case. This is so, not merely because alternative systems of geometry and dynamics can be developed consistently in mathematical terms (for Kant himself was aware of that fact); rather, it is so because 20th-century astrophysics and quantum mechanics have succeeded in giving non-Euclidean and post-Newtonian concepts an entirely coherent empirical application in the scientific explanation of natural phenomena—and this was something that Kant was not prepared to contemplate. Pure mathematics aside, indeed, Kant and most of his immediate successors were convinced that Euclid and Newton between them had somehow hit on a uniquely adequate system of geometry and physics—if not on the final mathematical truth about nature.

For about 100 years, then, the epistemic foundations and ontal commitments of this so-called classical science were largely taken for granted. The 19th-century debate in philosophy of science, accordingly, concentrated on peripheral topics and skirted all issues that might have called into question the ascendency of Euclid and Newton. The validity of the classical system having been assumed, the questions remaining for debate involved only its interpretation and implications; and the resulting positions can be classified, with slight oversimplification, under the headings of mechanistic (or Materialist) and Idealist doctrines, respectively.

Idealism versus mechanistic Materialism

The Idealists took to heart Kant's thesis that the cognitive structure of experience is imposed upon nature rather than discovered in it and sought to explore its broader consequences. The psychology of sensory perception, for example, previously barred from direct scientific study by Descartes's absolute separation of mind from matter, was now opened up for exploration; thus, by the mid-19th century, Hermann von Helmholtz, pioneer in a broad range of scientific studies, embarked on the remarkable investigations into the production of man's sensory experiences or ideas set out in his monumental *Handbuch der physiologischen Optik* (1856–67; Eng. trans., *Physiological Optics,* 1921–25).

For most of Kant's successors, however, the Idealist road led away from philosophy of science into other areas—particularly, into political ideology, philosophy of history, and sociology. Thus, it was not until well into the 20th century that the distinguished joint relativity–quantum theorist Sir Arthur Eddington, in his *Fundamental Theory* (1946), once again took up seriously the basic task of Kantian Idealism, viz., that of demonstrating, on a priori epistemological principles, that man's physical interpretation of nature embodies certain necessary structures imposed on physics by the character of his theoretical procedures themselves.

Meanwhile, the 19th-century mechanistic Materialists were disregarding Kant's central insights and concentrating instead on the apparent implications of the Newtonian system for other branches of science. A vigorous philosophical debate resulted, particularly in those fields that were just developing effective explanatory methods and theoretical concepts of their own. One good example of such a field was physiology, in which the work of a pioneering experimental physiologist, Claude Bernard, with its striking theoretical analysis of the vasomotor system and other regulatory mechanisms in the body, which anticipated 20th-century ideas about feedback systems, finally broke a long-standing deadlock between two opposed groups of scientists—the extreme mechanists, who recognized no difference at all between organic or physiological processes and the physicochemical phenomena of the inorganic world, and the outright vitalists, who insisted that the two kinds of phenomena were absolutely different. The debate also encouraged an epiphenomenal view of experience—as a kind of subjective mental froth without causal influence on the underlying physical mechanisms—and so sharpened the apparent threat to all claims about human "free will." Today it seems, indeed, that the sweeping conclusions of such scientific popularizers as the German evolutionist Ernst Haeckel, who wedded the ideas of classical physics with the new Darwinian history of nature to form a comprehensive Materialistic cosmology, or "anti-theology," carried more weight at the time than they seem in retrospect to have deserved.

Whewell versus J.S. Mill

There was one promising exchange about the central epistemic issues in the philosophy of science in mid-century, that which took place between William Whewell, a British philosopher and historian of science noted for his work on the theory of induction, and the political essayist, John Stuart Mill; but this was abortive. The debate ended in cross-purposes, largely on account of differences in temperament and preoccupations. Whewell's knowledge—not merely of contemporary physical science but of its whole historical background—was both broad and detailed. The mathematical necessity of arguments such as those in Newtonian dynamics impressed him quite as much as it had impressed Kant; but he gave a less grandiose account of the reasons for this necessity. Whewell's philosophy, a Kantian variation on Newton's hypothetico-deductive method, was historicized: it was only by a progressive approach that physicists arrived at the most coherent and comprehensive systems of what Whewell called "consilient" hypotheses—or separately derived, yet concordant, sets of laws—that were compatible with the empirical knowledge then at their disposal. Mill, on the other hand, principally concerned with the methodology of the social sciences, concentrated on the observational basis of science to the neglect of its theoretical organization and so emphasized the contingent, or unnecessitated, nature of all genuine empirical knowledge. In due course, certain of his doctrines, such as his account of arithmetical formulas as being a variety of empirical generalization, exposed him to some ridicule; but, for the time being, while the sheer bulk and learning of Whewell's writings muffled the force of his arguments, Mill's more fluent and less technical style of exposition captured the popular attention.

Empiriocriticism; Realism; conventionalism

In consequence, it was not until the 1890s and early 1900s that serious doubts grew up about the finality of the Newtonian synthesis; and the writings of Ernst Mach, Heinrich Hertz, the eminent German physicist Max Planck, Pierre Duhem, and others inaugurated the new phase of far-reaching critical reanalysis characteristic of 20th-century philosophy of science. In one way or another, all of these men stood back and looked at the Euclidean and Newtonian systems with fresh and less committed eyes. They had learned Kant's lesson about the constructive character of formal theories, without sharing his belief in the unique rationality of the classical synthesis; and, as a result, the central topics of their discussions turned on the best ways of restating the Kantian problem. Granting that the intellectual activity of theory construction has the effect of building a physical necessity into man's theoretical arguments, they asked, what then follows, ontologically, about the reality or conventionality of the resulting atoms, forces, electrons, etc., and what can be said, epistemologically, about the cognitive status and logical validity of its theoretical principles?

At one extreme, an Austrian physicist and philosopher Ernst Mach, and Richard Avenarius, the author of a philosophy known as empiriocriticism, expounded a sensationalist form of Empiricism reminiscent of David Hume, who had insisted that all ideas be traceable to "impressions" (sensations). On their view, theoretical concepts are intellectual fictions, introduced to achieve economy in the intellectual organization of sensory impressions, or observations, for which alone ontal primacy can be claimed. Correspondingly, all claims to scientific knowledge had epistemic validity for them only insofar as they could be grounded in such sense impressions. As against this instrumentalist or reductionist position, Max Planck, author of the quantum theory, defended a qualified Realism, which, at the least, expressed the ideal toward which all conceptual development in physics proceeds; for, without a belief in the enduring reality of external nature, he argued, all motive for theoretical improvement in science would vanish. Between these two extremes, Henri Poincaré, equally

distinguished in mathematics and the philosophy of science, and Pierre Duhem, a French theoretical physicist, occupied a range of intermediate so-called conventionalist positions, which attempted to do justice to the arbitrary elements in theory construction while avoiding the sort of radical doubt about the ontal status of theoretical entities that led Mach into lifelong Skepticism about the reality of atoms.

The 20th-century debate: Positivists versus historians. In the mid-20th century, debate in the philosophy of science became notably detailed, elaborate, and critical; those 50 years, in fact, have seen the subject finally achieving the status of a well-established professional discipline. Not least among the causes of this development have been the profound changes that have taken place since 1900 within theoretical physics and other fundamental branches of natural science. So long as the classical synthesis of Euclid and Newton retained its unquestioned authority, there had been little occasion to probe its ontological and epistemological bases at all deeply; but relativity theory—which qualified earlier geometries and laws in terms of new insights into the tie-ins between space and time—and quantum mechanics—which qualified them in terms of a statistical and indeterministic formulation—posed a frontal challenge to that synthesis and inevitably provoked critical and philosophical questions about the validity of the methods and assumptions on which it had relied. Consequently, between 1920 and 1940 there arose a renewed interaction between theoretical physicists and philosophers of science—especially between the Viennese Positivists and the authors of the new quantum mechanics.

Relativity and quantum mechanics

The main themes of the subsequent debate were largely those introduced into discussion in the period around 1900. Mach's critical reductionism, based on his *Beiträge zur Analyse der Empfindungen* (1886; Eng. trans., *Contributions to the Analysis of the Sensations,* 1897), in which he tried to reduce all knowledge to statements about sensations, was a prime source both of the Positivism and Logical Empiricism of the Vienna Circle—a group of eminent philosophers and scientists that met regularly in Vienna during the 1920s and 1930s—and also of the epistemological theories about sense-data and logical constructions developed in Britain about the same time by Bertrand Russell, perhaps the foremost logician and philosopher then in England; by G.E. Moore, a meticulous pioneer in Linguistic Analysis; and by others. Meanwhile, the qualified Realism of Planck and Hertz was carried further by such men as Norman Campbell, an English physicist known for his sharpening of the distinction between laws and theories, and Karl Popper, an Austro-English philosopher recognized for his theory of falsifiability, both of whose views reflect the explicit methodology of many working scientists today. A notable exception would be the Positivistic followers of Niels Bohr in the Copenhagen school of theoretical physics. Finally, there has continued to be substantial support for intermediate, conventionalist compromises, with Kantian overtones, along the general lines developed by Poincaré and Duhem.

Logical Positivism versus Neo-Kantianism

From the rich complexity of recent philosophy of science, two main strands may be selected for special mention here. The first is the strand of neo-Humean Positivism, which first developed in the Vienna Circle and has flourished more recently in the United States and has been fundamentally preoccupied with epistemological issues. While largely abandoning Mach's belief that sensations are the sole ultimate ground of knowledge, its proponents have continued, with Mach, to regard theoretical entities as fictions or logical constructs, the validity of which depends entirely on the capacity to give them a basis in empirical observations. This neo-Humean position has derived much encouragement, if not formal confirmation, both from Einstein's emphasis on the essential role of the observer in relativity physics and from the attack from the side of quantum theory made by the German physicist Werner Heisenberg on any sharp distinction, at the subatomic level, between the observer, his observation, and the system observed (see below *Philosophy of nature*).

The Logical Positivists and Empiricists harnessed to these epistemic arguments a formal apparatus taken over from the philosophy of mathematics—specifically, from Russell and Whitehead's *Principia Mathematica* (1910–13). In their view, the activity of theory construction is logically equivalent to the creation of propositional systems, in which groups of propositions are ideally set out in axiomatic form. So interpreted, the hypothetico-deductive method becomes a recipe for devising a succession of progressively more comprehensive axiom systems, based on alternative sets of general postulates (or primitive propositions) posited without proof, from which particular, empirical propositions can be inferred. As in the case of the special theory of relativity, these particular propositions—for instance, that the axis of Mercury's orbital ellipse will precess (or turn) at a certain rate—can then be used to validate the general postulates by comparing them with actual experience thus—directly or indirectly—substantiating the more general primitive propositions as well. The subsequent debate within the Viennese school has been concerned, very largely, with the exact character and force of this substantiation—whether it be verification, confirmation, or corroboration and/or falsification. At its most ambitious extreme, the Viennese school aimed at constructing a single system of unified science, by which the entire corpus of positive knowledge would be embraced in a single, all-embracing axiom system to be constructed around Russell's abstract symbolic logic. According to this program, all truly scientific knowledge must, first of all, be validated by appeal to neutral empirical observations, on pain of being dismissed as meaningless; and it must then be incorporated into the larger scheme of unified science.

The strongest opposition to this Empiricist or Positivist strand has come, correspondingly, from a Neo-Kantian school that has questioned the very possibility of identifying the pool of theoretically neutral observations necessary for substantiating or discrediting alternative theories in a strictly logical manner. This Neo-Kantian strand in 20th-century philosophy of science was inaugurated by the pioneering thinkers Heinrich Hertz (in electromagnetic wave theory) and Ludwig Wittgenstein (in philosophy of language). Rejecting Mach's central epistemological questions about sensations and ideas, these men have started instead from Kantian questions about the use of representations or models in the explanation of phenomena. Hertz's treatise on *The Principles of Mechanics* (1894), for instance, expounded Newtonian dynamics as a formal representation that logically entailed empirical conclusions only insofar as the phenomena concerned were already describable in terms drawn from the theory itself; and Wittgenstein's *Tractatus Logico-Philosophicus* (1922) extended Hertz's analysis to provide a general philosophical theory of language as an instrument for the representation of facts. The implications of this approach for the philosophical analysis and methodology of science have been explored further by some of Wittgenstein's pupils and successors, who have shifted the focus of discussion away from the verification of scientific propositions to the establishment of scientific concepts and theories; by highlighting the problem of conceptual change, they have revived interest in the philosophical significance of the history of scientific ideas. For, from this point of view, logical questions about the structure of propositional systems must be joined by other, equally fundamental rational questions about the manner in which different theoretical systems come to succeed one another (see below *Conceptual change and the development of science*).

Controversy among scientists

During this same period, the remarkable changes taking place within such sciences as theoretical physics, biochemistry, and psychology have been provoking philosophical discussions among scientists themselves. For instance, the displacement of classical Newtonian physics by Heisenberg's quantum mechanics has stimulated a new round of arguments about causality and determinism, with some people hailing Heisenberg's Principle of Indeterminacy—which holds that the location of a particle is intrinsically imprecise in the measure that its momentum is precise (and vice versa)—as giving human free will the toehold that rigorous 19th-century determinism did not seem to allow. Progress in cellular and subcellular physiology, moreover, has given rise to further rounds of debate in

the philosophy of biology. Claude Bernard, the foremost experimentalist of 19th-century medicine, had never managed to extend his analysis of regulatory mechanisms, such as those of the nerves that control the size of blood vessels, to cover the processes of embryology and morphogenesis (development of organic forms); and, on this finer level, there had been a renewed deadlock, around the years 1900–20, between the vitalism of Hans Driesch, which posited an almost "soul-like" reality that guides development, and the mechanism of Jacques Loeb, both experimental biologists with philosophical concerns. Once again, supporters of neither extreme position could make out their case entirely; instead, biologists have tended toward the mediating systemic conceptions first introduced by Paul A. Weiss, a distinguished developmental biologist, in the mid-1920s, and subsequently developed in detail as applications of the new theories of cybernetics and feedback, which conduct comparative studies of automatic control systems in the nervous system and in electromechanical engineering. More recently still, the development of molecular biology has compelled scientists to reformulate the problem of morphogenesis yet again—this time as the problem of seeing how the structural patterns of nucleic acid macromolecules in the hereditary genetical material find a structural expression in the developing body as a result of interacting with the environment. At present, this question is still very largely unanswered.

Controversy in the behavioral sciences

Methodologically, since 1940 one new centre of philosophical debate has developed, this time in the behavioral sciences. Ever since Descartes and Hobbes, there has been sharp disagreement about the legitimacy of extending the methods and categories of physical science to the sphere of the higher, distinctively human mental processes; and, even in the 1970s, theoretical psychologists were still far from agreed in their explanations of human behaviour. Some psychologists insist that human actions are subject to laws and mechanisms of the same kind as physical processes; others deny that any direct analogy exists between rules of conduct and laws of nature. Currently, this dispute is most lively in the psychology of language. Behaviorists follow B.F. Skinner, an American psychologist, in rejecting any distinctive class of mental laws and processes, whereas cognitive psychologists and generative grammarians, led by Noam Chomsky, argue that linguistic activities are creative and rule conforming in respects that no behaviorist can explain. In sociology and anthropology, equally, the 20th century has been a period of methodological controversy. Here the unresolved conflict has to do with the significance of history in the explanation of collective human behaviour. On one side, there is, in sociology, a school of so-called structuralists or functionalists that follows another American scholar, Talcott Parsons; or, in anthropology, there are the British ethnologists A.R. Radcliffe-Brown and Bronisław Malinowski, students of primitive mentality and behaviour, who regard all of the cultural practices and social institutions that function within a given community at any time as related together systematically within an overall structure: to explain any one of those practices or institutions, they hold, it is enough to show how it connects up with all of the other contemporaneous aspects of the culture. On the other side, a more historically minded school, notably the German "critical Marxists" (such as Jürgen Habermas), emphasizes the dynamic, developing character of social structures and relationships. Here again, the methodological debate is still in progress, and its eventual outcome cannot yet be clearly foreseen.

ELEMENTS OF SCIENTIFIC ENTERPRISE

It is appropriate at this point to define the recurrent problems that have played a central part in the philosphical debate about science and the crucial elements that any adequate philosophy of science must include in its account. From the beginning, scientists themselves have been interested not merely in cataloging and describing the world of nature as they find it but in making the workings of nature intelligible with the help of compact and organized theories. Correspondingly, philosophers of science are obliged to consider not merely nature in isolation—as a mere assemblage of empirical facts, mutely waiting to be discovered by man—but also the manner in which man himself perceives and interprets those facts when bringing them within the grasp of an intelligible theory and the respects in which the validity of the resulting theoretical ideas (or concepts) is affected by that processing of the empirical data.

Historically speaking, the problems posed by this interaction of man and nature have been complex and confused. Though philosophers of science face, even today, many of the same questions that were already being debated in classical Athens, the range and relevance of those questions has been greatly clarified in the meanwhile. For instance, when philosophers in the 17th century analyzed the nature and possible scope of a mathematical and experimental account of nature, they helped to clear the ground for Newton to develop the intellectual program and methodology of modern theoretical physics; while the subsequent philosophical debate about natural and artificial classification similarly cleared the ground for the scientific taxonomy of the Swedish systematist Carolus Linnaeus and the theory of natural selection of Charles Darwin. Methodological clarification in the philosophy of science has, in this way, repeatedly led to creative advance in science itself and so given rise, in turn, to new experience on which philosophers can draw in taking their methodological analyses further.

It is easy enough to list the chief elements that must find a place in any philosophy of science, but problems arise in mapping the relations between them.

Empirical data and their theoretical interpretation. First are the empirical elements. The task of science is to explain actual events, processes, or phenomena in nature; and no system of theoretical ideas, technical terms, and mathematical procedures—or mathematical procedures alone—qualifies as scientific unless it comes to grips with those empirical facts at some point and in some way and helps to make them more intelligible. On the one hand, the facts in question may be discovered by using observational methods—*i.e.,* by recording them as and when they occur naturally, without employing any special contrivances affecting their occurrence. This situation is, of course, the normal case in astronomy, in which the objects of study cannot be influenced or controlled. Alternatively, they may be discovered by using experimental methods—*i.e.,* by devising special equipment or apparatus with the help of which those processes or phenomena are caused to occur on demand and under specially controlled conditions. In that case, the scientist can attack scientific problems—to use Kant's vivid metaphor—by "putting Nature to the question," as in much of physics and fundamental biology. Either way, a philosophical difficulty at once arises about the results of the scientist's empirical studies: for he must ask how such raw empirical facts can be sifted, stated, and described in a way that throws light on the scientist's own theoretical problems. Do all empirical facts whatever serve as raw material for science? Or is this true only of those that have been preselected for their theoretical relevance—or even, to some extent, reshaped to ensure it? Is a scientist concerned with every particular empirical event, as such, or only with general phenomena or regularities recognizable in those events? Different schools of philosophers treat this raw material in very different ways.

Empirical elements

In the second place, there are conceptual elements. Every science employs its own characteristic abstractions, terminology, and techniques of interpretation and explanation, which can be of very different kinds. They may be ideal types, as in gas theory and parts of sociology; conservation principles, as in dynamics and energetics; taxa, as in biological systematics; particles or constituents, as in genetics and subatomic physics; models or flow diagrams, as in econometric analysis. Such conceptual elements are the intellectual keys by which phenomena are made intelligible, and a most active philosophical debate has turned around the part they play in the interpretation of phenomena. If, for instance, the idea of particles or ultimate constituents of matter is regarded as a concept created by scientists for the purpose of their own theoretical analysis, can an independent existence then be claimed for such

Conceptual and formal elements

theoretical entities in the world of nature itself? Or must all such ideas be regarded as fictions or constructs for which the claim to reality goes no further than the paper on which the scientific explanations are written? Similarly, if the theoretical descriptions of nature arrived at in science are unavoidably idealized and abstract, does this imply that the necessity attaching to arguments in, say, theoretical physics is itself only an artifact, or by-product, of scientists' own procedures for interpreting phenomena? Or can one, after all, speak of natural events themselves as happening "of necessity"?

Finally, every natural science includes also formal and mathematical elements or mathematical elements alone. These may be mathematical algorithms, or procedures of calculation, like those used in computational astronomy since Babylonian days, or like the computer programs that are their 20th-century counterparts; or geometrical constructions, as in certain branches of optics; or methods of graphical analysis, such as those used in handling statistical data; or the axiomatic systems by which, from classical times on, geometry and physics have been organized into formal schemata of propositions bound together by logical relations. Philosophers in the Platonic tradition give such formal elements special consideration, viewing as authentically intelligible only those theories the content of which can be presented explicitly in formal, and preferably in mathematical, systems of propositions. Theories of this kind alone are capable—as the seminal German logician Gottlob Frege expressed it—of employing "concepts in their pure form." Thus, 20th-century philosophers of science have devoted much time and effort to the question: How far, and on what conditions, can other branches of natural science (*e.g.,* quantum mechanics or genetics) be cast in the same definitive, axiomatic form as classical mechanics and electrical theory? Or is this formal construction itself merely a human convenience, adopted to simplify the handling of the empirical data, which reveals nothing more about the underlying structure of nature itself?

Relation of theory to observation

Each of these three groups of elements poses problems about which philosophers of science are still in deep disagreement; and these differences of view can usefully be illustrated by indicating the various approaches adopted by members of rival schools when discussing each of the groups. At one extreme can be cited philosophers of a radically Empiricist frame of mind, who regard it as important, above all, to emphasize the empirical foundations of scientific knowledge; for them, the raw facts of experience are primary and entitled to absolute respect. On this view, general theoretical principles have authentic scientific content only when interpreted as empirical generalizations about directly grasped empirical data; and, correspondingly, abstract theoretical entities must be understood as logical constructions from more fundamental elements that can be directly identified in empirical experience. (This belief, of course, was the basis of Mach's conclusion that submicroscopic atoms were merely intellectual fictions and derived their scientific meaning entirely from the macroscopic sense experiences that they were used to explain.)

At the other extreme, philosophers of a fully Rationalist, or Cartesian, bent can be cited who reject the idea that raw empirical facts, in and of themselves, display any intelligible or law-governed relationships whatsoever—and still less any necessary ones. For them, as for Plato, the scientist's bare experience of nature is a disorganized aggregate, or flux, unless and until he is able to discover some rational structure or principles relating these disconnected facts to a larger, more intelligible whole. Rather than allow equal significance and authority to every passing occurrence, the scientist, on this view, must be highly selective in the observations to which he pays attention; indeed, the very function of a well-designed experiment is now to create phenomena that can illustrate the intelligible relationships that are the true concern of science and so deserve the status of scientifically authenticated facts.

Both of these approaches, the Empiricist and the Rationalist, emphasize valid and important points; but, in their extreme forms, they give rise to difficulties that are probably insuperable. As to the Empiricist approach, the credentials of any scientific concept or theory certainly depend to a substantial extent on its basis in empirical experience. Indeed, much has been learned about statistics, the calculus of probabilities, and the design of scientific experiments from careful analysis of the procedures by which empirical data are actually handled, even before questions of theoretical interpretation were directly raised. Yet it is questionable whether sense impressions alone could ever serve as evidence for any scientific position, as Mach and the sense-data philosophers assumed. All genuine scientific observations, as Kant expressed it, have the form of judgments—*i.e.,* are expressed in statements answering questions formulated beforehand. It is probably an exaggeration to insist that all legitimate theoretical statements in science must be related in a strictly deductive manner to the everyday empirical observations that they are used to explain; and it is a caricature to treat the explanatory power of theoretical laws and principles, as for example in physics, as no different in kind from that of such an elementary generalization as "All robins' eggs are greenish blue."

As to the Rationalist approach, one of the chief tasks for philosophers of science is certainly to account for the rational interconnections that give scientific explanations their characteristic intelligibility. In this respect, such men as Descartes, Kant, and Hertz have deepened the philosopher's understanding of the scientific enterprise by obliging him to recognize the ways in which the intellectual organization of scientific theories rests on the scientist's own constructive activities, rather than on the specific facts. Yet, it would again be misleading to use this fact as an excuse for regarding physical theories—to echo a phrase of Einstein's—as entirely "free creations of the human mind." While the step from observations to theories does not rest on formal entailments alone, it would be an equally serious counter-exaggeration to suggest that theory construction is totally arbitrary or unconstrained by the imperative demands of the specific problems to be solved.

Science as a multiform enterprise

The outstanding task for most philosophers of science is, accordingly, to find an acceptable middle way between the Rationalist and Empiricist extremes and thus to do justice both to the empirical foundations of theories and to their internal organization. The different emphases of philosophers commonly reflect, at most, differences in their substantive preoccupations. Those who are interested (as was Mill) in possible methods for developing the human or social sciences naturally place most stress on the empirical basis of scientific knowledge. Those who are familiar (as was Whewell) with the actual outcome of theory construction in established sciences, such as physics, naturally underscore the systematic coherence and structure of scientific understanding. Those who are concerned with the nature and validity of historical understanding (as Giambattista Vico was) likewise end by giving a very different account of certainty and necessity from those (like Descartes) whose ideal of scientific knowledge is a formal, mathematical one.

If the philosopher comes to grips with the full complexity of the scientific enterprise, this approach can lead him to a more exact understanding of the varied intellectual problems of the natural and human sciences. Once it is recognized how different are the kinds of questions arising within such diverse fields as quantum electrodynamics and developmental biology, clinical neurology and historical sociology, the goal of formulating a single scientific method—with a universally appliable set of procedures and criteria for judging new theories or ideas in all fields of science—may come to appear a mirage. Yet the philosopher's legitimate insistence on generality has already helped to promote important extensions and integrations of man's scientific understanding. So again, he must now avoid taking too dogmatic a stand, either for or against complete generality, bearing in mind Kant's warning that the reason can hope to map its own proper boundaries only at the price of occasionally overstepping them.

The empirical procedures of science. Along with the three groups of elements already discussed, each phase in the scientific enterprise—empirical, formal, and con-

ceptual, or interpretative—involves its own characteristic procedures. On the level of empirical observation and description, three topics may be briefly touched upon, all of which are discussed at greater length in other articles (see MEASUREMENT THEORY; CLASSIFICATION THEORY; and articles on the basic sciences).

Measurement theory

First, there are the procedures of measurement through which scientists arrive at quantitative estimates of the variables and magnitudes considered in their theories. By now, there is a well-developed body of knowledge upon which scholars are agreed about many of the techniques and precautions to be employed in practice in the measurement of empirical quantities, in the calculation of probable errors or significant deviations, and so on. About the deeper significance of measuring procedures and their outcomes, however, there are still unresolved philosophical disputes. These disagreements reflect the same differences of approach already noted. Thus, some philosophers regard any scientific theory concerned with measurable (or quantifiable) magnitudes as intrinsically superior to a qualitative (or, as they would say, an impressionistic) one, however rich and well organized the latter may be. Others, by contrast, would argue that any insistence on employing numerical measures at all costs, even in such a science as, for example, systematic biology, can only lead the investigator to misconceive the true nature of the problems involved. Again, philosophers of an extremely Empiricist or Positivist persuasion have sometimes interpreted the experimental procedures for measuring theoretical magnitudes in, for example, physics as providing implicit definitions of the associated technical terms—the so-called operational definitions—and have thus felt able to claim that the logical entailments that the scientist is seeking between observations and theories are established by linguistic fiat (see below *Status of scientific propositions and concepts or entities*).

Design of experiments

Secondly, there are statistical analytical procedures for the design of scientific experiments. The mathematical techniques employed for this purpose are, in fact, closely related to those involved in the theories of measurement, probable error, statistical significance, and others. In this area, the connection between philosophical discussions of inductive logic and the practical procedures of working scientists is at its closest. Whereas a religious scientist (Blaise Pascal), a Nonconformist minister (Thomas Bayes), and an astronomer (Pierre, marquis de Laplace), all mathematicians as well, analyzed the philosophical foundations of the modern calculus of probabilities in the 17th and 18th centuries, 20th-century mathematicians and inductive logicians have similarly explored the intellectual basis for the design and interpretation of significant experiments; and, by now, the relevant procedures form a full-fledged branch of mathematical statistics with many valuable applications, particularly in fields such as sociology and economics, in which large numbers of variables are involved.

Problems of experimental design, however, can be stated clearly and unambiguously only in situations in which questions of fundamental theoretical interpretations are not actively at issue; *e.g.*, in experiments to determine which of two antibiotics is the more effective against a given infection or to learn whether significant correlation exists between two known physical variables. As soon as more fundamental questions of theory arise, however, the problems of experimental design go beyond the scope of any purely statistical analysis. Moreover, the same is true of computer-programming procedures: the numerical data obtained from a straightforward scientific experiment can in many cases be fed into a computer programmed to select the graph or formula in best statistical accord with the data from among those hypotheses conforming to a predetermined set of conceptual or interpretative requirements; thus, in this sense, a computer can be used to perform inductive inferences. Devising brand-new styles of conceptual or theoretical interpretation, by contrast, involves extending or modifying present explanatory procedures to satisfy novel intellectual requirements; and these tasks call for something more than formal statistical or programming techniques.

Classification theory

Finally, the initial handling of the scientist's empirical data requires him to employ procedures of systematic classification. The nature and validity of scientific classification procedures and of the species, genera, families, and so on into which scientists divide their empirical subject matter have been the subject of a long and contentious debate. A long-standing philosophical deadlock between supporters of natural and artificial classification systems was largely broken—in zoological taxonomy, at any rate—through the success of Darwin's theory of natural selection. As Darwin showed, the species to which organic evolution gives rise are neither eternally unchanging natural entities nor mere fictions of the zoologist's arbitrary creation; considered as coherent, self-isolating populations, they have a genuine though temporary reality that is preserved by the contrary processes of variation and selection perpetuation. In some other areas of thought, however, the preliminary identification and classification of the empirical material still raises contentious philosophical questions. When the sociologist theorizes about social groups or systems in the human sciences, for instance, he has to decide what collections of men and institutions do, or do not, fall under those general headings. Can objective tests be found for identifying natural units of sociological analysis? Or, is this choice of units merely set up for the sociologist's own convenience? This uncertainty about the very subject matter of sociology is itself an obstacle to the creation of an agreed-upon body of social theory; and comparable difficulties can arise in anthropology, linguistics, and psychology as well. It is, therefore, not surprising that some critics have even questioned whether these disciplines can truly be called sciences.

The character of these continuing difficulties underscores one point of general significance about the relation of empirical evidence to scientific theories: though philosophers may find it necessary to distinguish the empirical phases, elements, and procedures of science from the theoretical ones analytically, it does not follow that they can be kept wholly separate in actual practice. Satisfactory measuring procedures, experimental designs, and systematic classification principles are, no doubt, necessary preconditions for effective theorizing; but they themselves are subject, in turn, to revision and refinement in the light of subsequent theoretical considerations. In arriving at his dynamical theories, for example, Newton had to begin by relying on older commonsensical notions of effort, weight, and amount of movement; but he soon replaced these by the more exact, theoretically defined concepts of force, mass, and momentum, and this change reacted back onto the empirical procedures of physics also. Likewise in other fields of science, the decision as to whether or not the outcome of any empirical procedure is scientifically relevant or significant soon ceases to be a purely empirical question, as theoretical changes react back onto those empirical procedures and compel the scientist to modify his manner of collecting and describing the supposedly raw data of science. In this way, the empirical evidence by which his scientific conclusions are justified rapidly loses its pure and theoretically neutral character.

The formal structures of science. In this section and the next, those aspects of the scientific enterprise will be considered that have dominated recent debate in the philosophy of science, viz., the formal structures of scientific theory and the processes of conceptual change. It will soon be clear that the philosophical problems to which these two aspects, respectively, give rise are correlative and complementary—the one being static, the other being dynamic.

Ideal of a formalized science

Since 1920, most analytical philosophers of science have explicitly based their program on a presupposition inherited from Descartes and Plato, viz., that the intellectual content of any natural science can be expressed in a formal propositional system, having a definite, essential logical structure—what a leading American philosopher of science, Ernest Nagel, concisely called "the structure of science" in his book of that title (1961). One immediate inspiration of this program was the work of David Hilbert, a late 19th-century mathematician. To make the methods of mathematical proof more explicit and more perspicuous and thus more rigorous, Hilbert employed the

techniques of formalization, a reduction to relations while disregarding the nature of the relata, and axiomatization, a tracing of entailments back to accepted axioms.

The same techniques were taken over into the philosophy of mathematics by a pioneer German logician, Gottlob Frege, and into symbolic logic by Bertrand Russell and his collaborator Alfred North Whitehead; and, from 1920 on, the Viennese Positivists and their successors attempted to employ them in the philosophy of science also, hoping to demonstrate the validity of formal patterns of scientific inference by the straightforward extension of methods already familiar in deductive logic.

According to the resulting program, the primary task for the philosophy of science was to repeat in quite general terms the kind of analysis by which, in the science of mechanics, Heinrich Hertz, the formulator of electromagnetic wave theory, had already sorted out the formal aspects of science from its empirical aspects. The program was founded on the expectation that it would be possible, first, to demonstrate the existence of formal structures that were essential to any science, properly so-called, and second, to identify the nature of scientific laws, principles, hypotheses, and observations by their characteristic logical functions. Once this had been done, rigorous formal definitions could then be given of validity, probability, degree of confirmation, and all of the other evidential relations involved in the judgment of scientific arguments.

The actual working out of this program has involved complex and highly technical investigations, in the course of which great ingenuity has been displayed—as, for instance, by an outstanding philosophical semanticist and analyst, Rudolf Carnap, in his system of inductive logic, for the criticism of arguments in support of empirical generalizations, and by Hans Reichenbach, an eminent German-American Positivist, in his analysis of probabilistic arguments. So far, however, the program has yielded substantial results only as applied to arguments expressed in an idealized formal symbolism modelled on the lower functional calculus of mathematical logic. By contrast, little has been done to show how one might extend the resulting formal procedures to arguments expressed in the practical terminologies of working science. That extension, in fact, raises difficulties and ambiguities that are so far unresolved and may prove unresolvable. The goal of a purely formal analysis of scientific inference has generated difficulties, for instance, by tempting logicians to play down important differences between mere descriptive generalizations about natural phenomena and the explanatory theories (laws, principles, etc.) that a scientist develops to make those phenomena intelligible; and succumbing to this temptation creates problems both within inductive logic and in its applications.

Problem of generalizations versus laws

One distinguished supporter of this program, Carl Hempel, originally a member of the Berlin group (allied with the Vienna Circle), has discussed what he calls the theoretician's dilemma: if the task of explaining natural phenomena requires a proof that the character of those phenomena is formally entailed by the conditions of their occurrence, taken together with certain straightforward generalizations based on previous empirical experience, and if those empirical generalizations include references to hypothetical entities, then the theorist is faced with an invidious choice: for, in that case, either his generalizations (his laws) do in fact provide a logical link between the conditions of the phenomena and their actual occurrence and the assumption of hypothetical entities is formally superfluous; or else they do not succeed in doing so, and that assumption will not have strictly explained the phenomena. Clearly, this dilemma can be evaded only by challenging the identification of laws with generalizations and insisting that any appeal to laws of nature always involves the scientist in reinterpreting natural phenomena, not in merely generalizing about them.

Looking beyond the internal structure of inductive logic, the dubious equation of scientific laws with empirical generalizations has also been criticized on the ground that it treats the content of those laws as matters of happenstance, far more accidental or contingent than those expressed in any genuine law of nature. In the opposing view, the explanatory force of, say, the physicist's law of inertia is totally different from that of such a generalizing statement as "All swans are white"; and one can learn nothing about the validity of actual physical arguments unless his philosophical analysis respects that crucial difference. It has not proved easy, however, to analyze the formal structure of the sciences in any less abstract manner than that of the Viennese Positivists or to give a true representation of the working language and arguments of science. In his *Essay on Metaphysics* (1940), R.G. Collingwood, a British philosopher and historian, made one striking attempt, in which the formal structure of intellectual systems was explained in terms not of direct entailments between more or less universal propositions but rather of mutual presuppositions between more or less general concepts. In this account, the principle of inertia was not the most universally true assertion in dynamics but was, rather, the most generally applicable presupposition, or principle of interpretation. Such an account has the merit of explaining why, within a particular science, certain formal patterns of argument carry the apparent necessity that they do; but at the same time it lays itself open to the charge of yielding too much to relativism and so of destroying the objectivity of scientific knowledge by giving the impression that the conceptual structures of science are imposed on phenomena by the arbitrary choice of the scientific theorist himself.

Doubts about formalization

The late 1960s, accordingly, saw a renewal of questioning about the original assumption, viz., that the entire intellectual content of a science can be captured in a propositional or presuppositional system. Certain of the doubts about this thesis revive criticisms put forward at the turn of the century by a U.S. Pragmatist, Charles Sanders Peirce, who argued that the logical status of the theoretical terms and statements in a science is—in the nature of the case—subject to historical change as the conceptual organization of the science develops. (This same insight has been explored more recently by a U.S. logician, Willard Quine, who rejects any attempt to classify statements within scientific theories using the traditional hard-and-fast dichotomies—contingent–necessary and synthetic–analytic—as fallacious and dogmatic.) Other criticisms of the thesis go deeper. By focussing his philosophical attention exclusively on the static formal structure of propositional systems and, so, on the intellectual content of the sciences at particular temporal cross sections in their development—they point out—the philosopher is distracted from the complementary questions about the manner in which the conceptual organization of a science changes and, so, from the traditional claims of natural science to be a rational as well as a logical enterprise. At this point in the debate, therefore, the spotlight shifts away from the static problem of analyzing a science in static logical terms to the historical problem of analyzing the dynamic processes of intellectual and conceptual change.

Conceptual change and the development of science. The problem of conceptual change has recently come back to the fore. The crucial question it poses is: "What is a concept?" In the heyday of Logical Empiricism, that question had largely been disregarded. Following the example of Frege, the Viennese Positivists had condemned any tendency to regard the philosophy of science as concerned with scientific thinking—which was in their view a matter for psychologists—and had restricted themselves to the formal analysis of scientific arguments. This preoccupation with logic was also reflected in their view of concepts. To interpret a concept such as force as referring either to a feeling of effort or to a mental image could lead, they argued, only to confusion. Instead, the philosopher must equate concepts with the terms and variables appearing in the propositional systems of science and define them, in part by reference to their roles in the formal structures of those propositional systems—thus fixing their systematic import—and in part by reference to the specific events and phenomena they are used to explain—thus fixing their empirical import. In the 1920s and 1930s, accordingly, all substantive philosophical questions about the concepts of science were dealt with summarily: they were simply translated into logical or linguistic questions about the

formal roles and empirical references of technical terms and mathematical variables.

Styles of explanatory procedure

Once the philosophy of science is approached more historically, however, those substantive questions must be faced afresh in their own right. Rival scientific theories will now be distinguished not merely as so many alternative formal systems, based on different primitive terms and axioms, but also as alternative ways of organizing the knowledge of nature, based on different explanatory techniques and modes of representation. The distinctive features of different scientific concepts will lie, as a result, not in their respective formal roles and empirical references but in the styles of explanatory procedure involved in their application. Those procedures may be of many different kinds: *e.g.,* physical conservation calculations, optical-ray diagrams, functional analyses, taxonomic classifications, historico-evolutionary reconstructions, or dynamical axiom systems. Correspondingly, they provide occasions for employing mathematical formulas, or intuitively intelligible models, or genealogical trees, or other styles of representation. In each case, however, the philosopher can describe the conceptual organization of the resulting explanations in terms neither of intuitive models nor of mathematical formulas and variables taken alone: what he must now consider is the entire pattern of theoretical interpretation—models, mathematics, and all.

Viewed from this alternative standpoint, the philosophy of science will begin by identifying the different styles of explanation characteristic of different sciences or of different stages in a given science and will recognize how those differences in explanatory style reflect the characteristic problems of different scientific fields and periods. So considered, empirical generalizations and descriptive classifications will serve to organize the empirical data of science in a preliminary way; but serious theoretical interpretation can begin only after that point. The central philosophical task now is to analyze, clearly and explicitly, (1) the standards by appeal to which scientists have to decide whether or not some interpretation is legitimate, justified, and conclusively established and (2) the considerations that justify giving up one currently accepted interpretation in favour of an alternative, novel one.

The first of these questions is one that the Logical Empiricists set out to answer in their own manner. They treated the empirical data and the theoretical principles of science as being connected by purely logical relations and attempted to define the required standards in terms of a formal theory of confirmation, corroboration, or falsification. The second question is one that they never seriously tackled. Instead, they assumed that one could, first, work out a quantitative index of acceptability for individual theories taken separately and, afterward, use this as a scale for measuring and comparing the merits of rival theoretical interpretations. By now, however, it is evident that, when biophysicists, say, abandon one theoretical approach in favour of another—as being more fruitful from the standpoint of biophysics—the considerations that lead them to do so are by no means analyzable in formal terms alone. On the contrary, the ability of a biochemist, say, to judge whether or not such a change in approach will effectively help to solve his theoretical problems is one of the most severe assessments of his substantive grasp of what biochemistry is about.

Reinterpretation of concepts and principles

In this way, the shift of attention from the propositions of science to its concepts is making philosophers more aware of the extent to which theoretical understanding involves the reinterpretation of empirical results, not merely their formal transformation. Similarly, the problem of conceptual change is raising questions about the processes by which theoretical interpretations succeed one another and about the procedures of conceptual judgment that are applied in the rational development of a science. These questions are currently under active discussion, and several lines of attack are being considered, none of which has finally established itself.

At one extreme, there are some who still regard theoretical concepts and principles as organized into compact, logical systems and who attempt to define the alternative standpoints of different sciences as the consequences of different basic premises or presuppositions. Having adopted this systematic approach, the investigator then discovers that conceptual change at a fundamental level finds adequate scope only through the replacement of one complete formal system by another, distinct and separate successor system. As a result, fundamental theoretical change is, in this view, intelligible only as the outcome of thoroughgoing intellectual revolutions, in which one entire theoretical system—axioms, principles, criteria of relevance, standards of judgment, and all—is swept aside in favour of another.

Alternatively, there are those who distinguish two different kinds of fundamental principles in a science—marking off the basic theoretical assertions such as "Matter consists of atoms combining into molecules" from its methodological maxims and standards of judgment, such as "All physical phenomena are to be explained in mechanical terms"—and who recognize fundamental conceptual changes in the science as legitimate, just so long as they respect the methodological maxims that are definitive of the science in question. In this second view, conceptual changes of any depth in the intellectual substance of a science will continue to be intelligible, provided only that the new views are still governed by the established program and framework concepts of the science in question. There will then be revolutions in science only when some entire intellectual approach is discredited—*e.g.,* that of 16th- and 17th-century iatrochemistry, which studied chemistry as a means of treating disease—or when some entirely new science is created, with its own complete system of interpretation—*e.g.,* molecular biology.

At the other extreme, there are some who doubt whether any sharp distinction can be drawn between substantive theoretical assertions and maxims of methodological procedure and who argue that all aspects of a natural science are alike open to historical reconsideration and modification. The more specific the theoretical doctrines and concepts being considered, the more risky they will be, and the more readily they will be modified or abandoned or both. From this third point of view, however, it is questionable whether any change, however drastic—even the hybridization of crystallography, viral genetics, and biochemistry, which led to the inauguration of molecular biology—is ever as discontinuous or revolutionary as the two former views imply. Instead, the attempt may be made to account for the procedures and processes involved in the historical development of scientific concepts by using the same general form of theory on every level—explaining innovation as requiring a selective choice between intellectual variants of different kinds; and, in this way, the theory of conceptual development can be brought into line with other historically based theories of natural and cultural change.

Developmental nature of science

Whichever alternative is adopted, one point must be kept in mind: the moment that problems about the changing theoretical organization of science begin to be treated in an authentically developmental manner, philosophical inquiries are given a quite new direction. This step compels one to view all questions about the logical structure and propositional systems of science against a broader historical background. In this new context the natural sciences are seen not as static formal structures but as rational enterprises characterized by certain typical intellectual procedures or movements. These basic procedures of intellectual development in science are the topic to which attention is directed in the next section.

MOVEMENTS OF SCIENTIFIC THOUGHT

Discovery and rationality. In analyzing the natural sciences for philosophical purposes as historically developing enterprises, the question "What is it that makes the sciences rational?" is raised in a new form: do the intellectual procedures that scientists actually employ to investigate and explain natural phenomena have definite and objective intellectual merits that make their adoption rationally prudent, wise, and obligatory? In answering this question, philosophical opinion has tended to polarize in recent years toward two extreme positions: on the one hand, a

formalist or positivist extreme, on the other, a romantic or irrationalist one.

Given their mathematical inspiration and preoccupations, both the Viennese Empiricists and their successors in Britain and the United States have interpreted the rationality of scientific procedures as depending solely on the formal validity, or logicality, of scientific arguments. In their view, questions of rationality can be raised about the scientist's work only at the final stage in his inquiries—*i.e.,* when he sets out, as the final outcome of his work, the explicit explanatory arguments in support of his novel theories or interpretations—only then, they declare, will there be anything about science that is capable of being criticized in logical or philosophical terms.

Formalist versus irrationalist

It is therefore a commonplace of recent Empiricist analysis in the philosophy of science that one must distinguish at the very outset between discovery and justification. The term discovery refers to all the stages in a scientific inquiry preceding the formulation of the new explanatory arguments that are its final outcome. The term justification refers, by contrast, to the demonstration that the formal validity or explanatory power of those arguments justifies the scientist in accepting their conclusions as scientifically validated or established. In this view, the rational concerns of the philosopher of science are restricted solely to this final phase of justification. All questions about the earlier stages—*i.e.,* about discovery—are matters of mere psychology, not of serious philosophy. As one widely accepted epigram expresses it, "There is no logic of discovery"; and this distinction—given the equation of rationality within logicality—seemingly invalidates all questions about the rationality of the preliminary steps by which a scientist arrives at a discovery.

At the opposite extreme, there are those, such as Michael Polanyi, a Hungarian-born scientist and philosopher, and Arthur Koestler, a novelist and journalist, who emphasize the parts played by intuition, guesswork, and chance in scientific investigation, citing these as evidence that theoretical achievement calls into play an intellectual creativity superior to mere rationality. According to this anti-Positivist argument, the modern scientist is a sleepwalker whose creative insight guides him to intellectual destinations that he could never clearly see or state beforehand: any excessive preoccupation with the rationality of scientific procedures, by contrast, springs from a pedestrian desire to clip the wings of imagination and to confine the scientist to stereotyped procedures, thus destroying the creative fertility of science. Rather than subjecting scientific intuition to the barren intellectual accountancy of the Positivists, the conclusion runs, one should embrace a romantic anti-rationalism.

In each of these extreme cases, however, the initial equation of rationality with logicality demands closer examination. Certainly, the activity of investigation and discovery can be examined with advantage from a psychological point of view as it has been, in fact, by a French mathematician, Jacques Hadamard, as well as from a philosophical point of view. Yet the possibility of such psychological inquiries does not obviously prove, entirely by itself, that procedures of intellectual investigation in science and mathematics are essentially nonrational. Chance, for instance, may help to bring relevant material to a scientist's attention. But chance—as has often been remarked—favours the prepared mind, and it is fair to ask how far the scientist acted rationally, after all, in picking out the items he did as being relevant to his particular problems. Similarly, in the case of creative intuition and the rest: once again, the man with the best trained mind can afford to give the freest rein to his intellectual imagination because he will be best qualified to appraise the rational context of his current problems and to recognize significant clues, promising new lines of analysis, or possible answers to his questions, as they come to mind.

Problematics of scientific inquiry

Neither denigrating the early phases of scientific inquiry as of merely psychological interest nor overpraising them as exercises of creative imagination disposes therefore of the philosophical problem that is here involved, viz., that of showing what makes certain procedures of investigation more rational than others. To find a middle way between formalism and irrationalism, it is necessary to look more closely at the nature of the problems of scientific inquiry. If the improvement of scientific concepts and theories depends on the development of more powerful explanatory procedures, the philosophical analysis of discovery then requires that one show what is essentially involved in devising such procedures, testing them out, and determining the range of their application. This problem must be dealt with, furthermore, not by a formal analysis of the resulting arguments alone but first and foremost by establishing what tasks any novel explanatory procedure in science can be required to perform, what demands its performance can properly be asked to satisfy, and so what intellectual goals a scientist is expected to be aiming at in all the phases of his investigations. Posed in these alternative terms, the problem of scientific rationality becomes a problem of showing how conceptual changes in science result in the introduction of novel ideas, which are—in a phrase coined by Mach as early as 1910—"better adapted, both to the facts and to one another." It is rational for older scientific theories to be displaced by newer ones that are functionally superior; and the task for philosophers of science is to demonstrate explicitly in what such functional adaptedness consists.

At the present time, many younger philosophers of science are actively analyzing the nature of the problems of science in these terms. Significantly, most of these men have had their own primary training within the natural sciences proper rather than in formal logic or pure mathematics, for the task requires a much more detailed analysis of the processes of intellectual innovation than has been customary hitherto. In place of the simple dichotomy between discovery and justification, for instance, it calls for a subdivision of the innovation process into a more complex sequence of distinct stages; and at each stage both rational and causal considerations are relevant. Thus, at the initial stage in any inquiry, a scientist must decide which among all of the philosophically conceivable variants from the current repertory of explanatory methods are to be taken seriously at all; which, that is, are genuine possibilities. This preliminary sorting of initially plausible from implausible innovations must be dealt with—and dealt with in the most rational manner possible—long before any question of justification arises.

This initial sorting procedure is one about which scientists themselves also speak cogently and eloquently. Far from deciding what novel suggestions are genuinely possible or plausible on a purely psychological basis or by the exercise of some mysterious, nonrational intuition, scientists will commonly explain their reasons for accepting one set of conceptual variants rather than another as deserving serious consideration. At the same time, such microanalyses of scientific innovation must certainly leave room for causal as well as rational questions. During certain periods in the historical development of science, for instance, scientists have notoriously disregarded novel possibilities that later turned out to hold a key to the solution of crucial theoretical difficulties. Looking back at such periods, it is possible to reconstruct with care the rational considerations that might have been advanced at the time to explain this neglect; but even so, one is occasionally forced to conclude that the men involved were prejudiced against those possibilities by factors external to their sciences; *e.g.,* by influences originating in the wider social, cultural, or political framework of their time. Thus, Newton was particularly afraid that his theory of material particles might be accused of supporting Epicureanism, whereas Darwin concealed his private speculations about the cerebral basis of mental activities because of public objections to Materialism.

Diverse approaches in different fields

In analyzing the microstructure of scientific problem solving, it is necessary, accordingly, to resist any temptation to generalize prematurely. Scientific investigators working in different fields, or at different times, apparently face theoretical difficulties of quite different kinds. One must therefore begin by studying the specific needs and tasks of each particular science, at one or another stage in its evolution, separately—seeking to recognize, in each individual case, the particular intellectual demands to be met by any new

concept or theory if it is to be successful. Eventually, the accumulated results of specific microanalyses may bring the investigator to a point at which he can again afford to generalize about all of the assorted theoretical problems confronting, say, physics and about the broader intellectual demands to be met by successful theoretical changes in a variety of scientific situations. At the present stage, however, though philosophers of science still cannot afford to beg these questions, they are compelled to conduct their analyses in a more piecemeal way—building up their picture of scientific innovation and discovery by considering a wide range of sample cases and working their way only gradually toward a more comprehensive account of the problematics of the scientific enterprise.

Validation and justification. If this situation is true of the earlier stages in discovery, it is no less true in the case of justification itself. Here again, from 1920 on, the debate in the philosophy of science focussed predominantly on two sharply opposed positions, both of which appear in retrospect to be excessively narrow. On the one hand, Empiricist philosophers argued for a view that made prediction the crucial test of scientific validity; on the other hand, philosophers of a more Rationalist temperament saw coherence and scope as the crucial requirements.

Empiricist versus constructivist positions

For Empiricists, the fundamental presupposition is that the facts justifying changes in scientific ideas are both intellectually prior to the theories that are, in due course, developed to explain them and also capable of being recognized independently and in advance of all theory construction. Given this presupposition, they regard prediction and validation as the crucial and distinctive steps in scientific procedure, arguing that, to establish the validity of any general scientific proposition, it is necessary to show that the theoretical generalization of which the validity is in question entails particular factual statements that are borne out by independent empirical observations. This validation process then involves two essential steps: (1) the formal step of inferring novel predictions from the theory and (2) the empirical step of comparing those predictions with the facts and so confirming the theory or proving it false.

On closer inspection, both steps in the received Empiricist procedure face serious difficulties, and these have lent strength—by reaction—to the alternative, constructivist position. As to step (1), there appears to be no objection to the idea of deducing particular factual predictions directly from theoretical hypotheses, so long as one accepts the Empiricist interpretation of laws of nature as universal empirical assertions on the same logical level as "All polar bears are white." Once that interpretation is questioned, however, it is less clear that direct deductive inferences from theory to fact are always practicable. On the contrary, if theoretical laws and purely empirical reports are, in the nature of the case, framed in terms of distinct and diverse sets of concepts, no general procedure can be available for passing deductively from one to the other. For the theory will then be a reinterpretation of the facts, not a mere generalization from them. Similarly, with step (2), an empirical confrontation of theories and facts gives rise to a more complex range of choices than those implied by the Empiricist account. When faced with discrepancies between prediction and observation, scientists certainly have to modify their theoretical explanations; but this modification can normally be made in any of several alternative ways. For instance, the theoretical relevance of a particular observation may be questioned; or some alternative theoretical interpretation may be put forward; or further refinements may be made within the structure of the theory concerned—and all of this can be done before any question arises of a direct and necessary conflict between the discordant observation and the general theoretical doctrine under investigation.

The rival, constructivist position derives its attractions from such objections as these. This position follows lines of thought already sketched by the French theoretical physicist Pierre Duhem at the turn of the century. On this account, the essential test of a science is that it should provide coherent, consistent, and wide-ranging theoretical organizations. Empirical facts will then be recognized as scientifically relevant only to the extent that they exemplify these interpretations and make them more discriminating. Thus, no single factual observation can ever serve as a logically crucial experiment and confirm or refute any one specific doctrine conclusively, taken apart from a whole complex of theory and interpretation. What is at risk in any experiment or observation, therefore, is the whole body of the theory, together with the current conventions governing its empirical application; and the more comprehensive a theory is, the more are scientists free to vary the details of their specific applications of it, rather than to accept any single counter-example as a challenge to its general validity.

Complementarity of the modes of justification

If these two philosophical approaches are reconsidered today against a broader and more historical background, however, they no longer appear to be either as exhaustive or as contradictory as they did in the 1920s and 1930s. By choosing suitable illustrations, of course, one can make each position highly attractive and plausible since, in one situation or another, the rational considerations that carry genuine weight in the actual justification of novel scientific theories include both predictive success and conceptual coherence. But the "Book of Nature," as Galileo called it, is like Holy Scripture: it offers texts to suit all occasions and purposes. And, on second thought, it can be argued that both Empiricist and constructivist philosophers oversimplify the justification process in science and the criteria by which scientists judge the validity of novel concepts and theories. Far from there being any single or simple test of validity, the question whether predictive success or coherence, simplicity, historical authenticity, or mechanical intelligibility is the key consideration—and in what sense of each ambiguous phrase—must be considered afresh from case to case, with an eye to the specific demands of each new scientific problem situation.

Within the historically developing enterprise of science, intellectual problems arise of many different types, depending both upon the kinds of subject matter under investigation and upon the stage of development of the science concerned. In one science and at one stage, particular weight may attach to a single unexpectedly successful prediction: as when the wave theory of light led to the totally unexpected discovery that a perfectly circular obstacle placed in front of a point source of light produces a circular shadow having a bright spot at its centre. In another science or at another time, however, it may be neither practicable nor relevant to infer such specific predictions, and new theories and concepts may be validated by considerations of quite other kinds. Even within a single science such as physics, indeed, scientists are not faced at every stage by problems and judgments of a single, uniform type. Instead, the historical evolution of physics—down the centuries from Nicole Oresme, Galileo, and Newton to Maxwell, Rutherford, and Heisenberg—has generated an entire genealogy of varied problems; and the considerations bearing on the theoretical difficulties facing physicists at different stages have themselves changed, quite legitimately, along with the substantive concepts and theories of the science. So, within the more complex framework of a developing rational enterprise, the philosopher's task is no longer to impose any single or simple criterion of intellectual choice upon scientific judgments of all kinds. Rather, his task is to recognize how the rational considerations and criteria of validity relevant to particular judgments vary with the theoretical problem situations that provide their historical contexts.

The unity of science philosophy

Unification, pluralism, and reductionism. As one notable illustration of the tug-of-war between logical and pragmatic issues in the philosophy of science, the "unity of science" movement may be cited. Under the vigorous leadership of Otto Neurath, a polymath sociologist and philosopher, this movement represented the high point in the ambitions of Viennese Positivism between World Wars I and II; for the general philosophical aims that motivated the search for a unified science are in striking contrast with the specific problem-solving considerations that lead working physicists to unify or integrate their theoretical concepts and explanatory procedures in actual scientific practice.

Aside from the primary test of predictive success, the Positivists of the Vienna Circle also did allow—on their own terms—for the further theoretical virtues of coherence and comprehensiveness. Their logico-mathematical approach to the propositional structure of scientific theories, however, led them to interpret this demand for coherent and comprehensive theories in a formal sense. On their interpretation, a totally unified body of scientific ideas would be a comprehensive, quasi-Euclidean system of scientific theorems, based on a single set of general axioms, postulates, and primitive propositions and applicable to natural phenomena of all kinds. Given sufficiently all-embracing empirical generalizations as the starting points of such a unified science, it would then be possible, in their view, to deduce particular statements about all the phenomena covered by the varied special sciences unified within its axiomatic scope. Taking the symbolic logic of Russell and Whitehead as their formal core, philosophical advocates of the unity of science then set out to construct, on a single axiomatic pattern, a fully comprehensive account of nature capable of explaining (*i.e.,* entailing) all natural phenomena whatsoever.

At first glance, this ambition seemed laudable and legitimate, but once again the Empiricist program subsequently encountered unforeseen entanglements. The reasons for this situation were not merely the discovery that the theoretical ideas employed within different branches of a science (*e.g.,* of mathematical physics) are more resistent to conceptual integration than had originally been hoped (the task of constructing a self-consistent relativistic theory of quantum electrodynamics, for instance, is one that still defeats the physicists); but, what is worse, it has now become apparent that several well-founded and properly respected branches of scientific theory do not lend themselves to exposition in a formal mathematical manner at all. Any satisfactory theory of organic evolution, for instance, has an irreducibly historical dimension; and there is no possibility of putting historical zoology on the sort of predictive basis that Empiricists have demanded, still less of incorporating it into Neurath's larger unified axiom system. Faced with this particular example, indeed, one distinguished Empiricist philosopher, Carl Hempel, has drawn a somewhat extreme conclusion, viz., that the theory of natural selection is not really an explanation of organic evolution at all—not even a bad one—but is merely an elaborate redescription of the historical episodes concerned. Yet this is simply a roundabout way of conceding that neither the historical problems nor the theoretical ambitions of evolutionary zoologists conform to the quasi-mathematical pattern that the Logical Empiricists have set out to impose on all of the natural sciences alike in the interests of a longer term axiomatic unification.

Unification in scientific praxis: reductionism

If, on the other hand, the demand for integration or unification is considered as a practical problem of methodology, it will then be found that the scientists are facing problems of a different and more pragmatic kind. The science of physiology poses an interesting example because, within this field, the problem of reductionism—*i.e.,* of whether all phenomena whatsoever can be reduced to physico-chemical terms alone—has repeatedly drawn active debate. Since the time of Antoine Lavoisier, who first explained correctly the process of combustion—*i.e.,* since the late 18th century and even before—there has been a methodological division of opinion, involving, on the one hand, those chemists and physiologists who dreamed of equating physiological functions with chemical reactions and planned their program for biochemistry around that ambition and, on the other hand, those clinical scientists and functionally minded physiologists who questioned the legitimacy of this so-called physicalist program and insisted that physiological phenomena displayed certain features or aspects inexplicable in physio-chemical terms alone. The scientific issues in debate in this case have never been concerned with formal matters of axiomatization and logical integration alone: once again, they have involved substantive questions of interpretation. Correspondingly, the provisional resolution of this dispute, accomplished by Claude Bernard in the mid-19th century, was not arrived at by constructing a single, unified axiom system of biochemistry-cum-physiology. Rather, Bernard distinguished the proper questions and concerns of the two sciences and demonstrated the substantive character—and limits—of their mutual relevance.

Regarded as specific, localizable processes within the main organs of the body, he argued, all physiological phenomena do, indeed, come within the scope of the same general physico-chemical laws and concepts as govern similar processes in inorganic systems. Within the special micro-environments of the body, however, those same general types of phenomena serve certain unique physiological functions, having no inorganic counterparts; and, to this extent, special problems and questions arise within physiology that cannot be exhaustively translated into the language of inorganic physics and chemistry. Though biochemistry and physiology in no sense conflict, there accordingly remains an essential plurality in the explanatory aims of the two sciences; and this plurality gives rise, in turn, to a corresponding plurality of methods and concepts.

Yet even this example does not yield conclusions from which one can safely generalize. Though in certain respects the explanatory aims of physiology and biochemistry will, most probably, always be distinct and separate, in other cases matters have gone the other way. When, in 1873, the Scottish physicist James Clerk Maxwell, for instance, integrated the previously independent sciences of electricity, magnetism, and optics into the unified physics of electromagnetism, there was no comparable division of opinion and no such methodological peace treaty was needed. In this case it remained possible, after Maxwell's work as before, to distinguish between straightforwardly electrical, magnetic, and optical phenomena on the empirical level; but on a more general, theoretical level such distinctions lost their earlier significance, and it ceased to be necessary to keep the problems, methods, and explanatory categories of the three earlier sciences separated.

Limited scope of unification

To sum up: in the methodological drive toward the unification of the sciences, as in the earlier phases of discovery and validation, the intellectual temptation to generalize prematurely exposes the philosopher to certain real dangers. In practice, the case for unifying the theories and concepts of two or more sciences has to be considered afresh in every instance, and it can rarely be decided in advance whether or not such a unification will achieve anything useful for the sciences. Instead, one has to analyze the practical demands of the current problems in the different fields and see how far those requirements can be met by developing a unified explanatory treatment for all of the special sciences in question. The integration of theoretical concepts achieved in the process will not consist solely in the formal running together of different propositional systems: more typically, it will require the development of a whole new pattern of theoretical interpretation. And, though it may be possible, in certain cases, to expound the resulting theory in axiomatic form, it must be established, in each case separately, whether or not this can be done. In this sense, conceptual and methodological unification represents a genuine movement in the development of scientific thought; but the logical form of the unified science towards which the philosopher is working is not something that he can lay down definitively before the event.

PHILOSOPHICAL STATUS OF SCIENTIFIC THEORY

Status of scientific propositions and concepts or entities. The section of this article entitled *Conceptualization and methodology of science* examined, first, the raw material (or elements) with which scientists have to work in developing their theories about the operations of the natural world and, second, the intellectual steps (or movements) by which they arrive at a scientific understanding of nature. By way of summary, it is appropriate to consider finally the main points of view about the intellectual status of the scientific concepts and doctrines embodying the understanding of nature that have emerged from the philosophical debate about science. Beginning with the epistemic status of theoretical propositions in science, it is well to consider the different claims that are made about the objectivity of their applications or their truth or both.

Then, turning to the ontological status of the scientific concepts or entities, it is likewise necessary to consider the claims that are made about the objectivity of their reference or of their meaning or both. In either case, the purpose of a philosophical critique of science is to establish just how far the content and reference of scientific knowledge can be regarded as a true report about the actual structure and operations of nature and just how far they represent, on the contrary, intellectual constructs or artifacts in terms of which men have chanced, chosen, or found it desirable to organize their thoughts about the structure and operations of nature.

Epistemic aspects

Starting with the epistemic status of scientific theories, three main views can be distinguished: At one extreme is a strict Realist position, which underscores the factual basis of all scientific knowledge and emphasizes the logical contingency that this basis implies for all substantive propositions in science. In this view, all but the most purely formal statements in science make assertions about how the world of nature is constituted and operates in fact—as contrasted with all of those alternative states of affairs that are clearly intelligible and so possible but which turn out not to be true of the actual world. Seen from this Realist standpoint, every proposition in science, from the most particular observational report to the most general theoretical principle, simply reports a more or less comprehensive empirical set of facts about nature and aspires to be an accurate, objective mirror of the more or less universal facts about which it speaks.

At the opposite extreme, there is a strict conventionalist position, which underscores the constructive role of the scientist's own theory articulation and emphasizes the logical necessity that is thereby built into the resulting conceptual structure. In this view, all but the most purely observational statements in science reflect the patterns by which the scientist shapes his conceptual picture of the world of nature—the patterns in terms of which all states of affairs clearly conceivable on the basis of current ideas have necessarily to be formulated. Seen from this conventionalist standpoint, theoretical thermodynamics, say, determines the character of all possible worlds consistent with the principles of energy conservation and entropy (or randomness) increase: a world to which thermodynamics is not applicable will then be not so much factually false as inconceivable in present terms.

Finally, a wide range of intermediate views seeks to evade the central opposition between Realists and conventionalists. One representative view of this kind, first made popular by Mach toward the end of the 19th century, invoked Kant's attack on things-in-themselves, viewing the attack as providing grounds for dismissing all debates about reality and objectivity as inescapably barren and empty. In its most developed form, this so-called operationalist position encourages the philosopher to regard theoretical propositions in science as meaningful only insofar as scientific practice includes specific operations—either manual measuring operations, or computational pencil-and-paper operations—in terms of which those propositions are given operational meaning. Nothing is then to be read into scientific knowledge beyond its operational meaning; in particular, scientists are not to be understood as claiming or disclaiming anything about the reality or conventionality of the states of affairs that they report. The idea of nature as a thing-in-itself is thus eliminated, as being an intellectual superstition and an obstacle to better scientific understanding, which survives from an earlier metaphysical era.

Ontological aspects

Three main views should here be distinguished. The central question is, now, whether the nouns and noun phrases used as technical terms in the theoretical propositions of science rely for meaning on any claim that they refer to objective, external entities; and current approaches to this question parallel existing views about the epistemological issues, viz., whether the propositions themselves rely for their truth on a claim to be mirroring or reporting objective, external facts. Here, too, the Realist interprets all of the chief technical terms of scientific theory as the names of objective entities existing in nature independently of all human theories and interpretations. In this view, entropy, say—a measure of the increase in randomness that every total system undergoes—is a genuine, objective magnitude that has, at all times, played a crucial part in the operations of nature even though physicists have only recently had the wit to discover it; and it just happens, correspondingly, to be the case—in those parts of the cosmos that can be observed—that the total entropy of an isolated system nowhere decreases.

The instrumentalist, for his part, regards all theoretical notions such as entropy as intellectual fictions or artifacts created by the scientist's own theory construction and quite distinct from the natural world of objects, systems, and phenomena that scientific theories have to explain. No doubt, scientific theory and the external reality of nature do come into contact on the everyday or empirical level of tables and chairs, rocks and flowers. Given the intellectual tasks of scientific theorizing, however, the resulting concepts are essentially abstract; and any grasping after real entities, as the objective external reference of the theoretical terms, reflects a plain misunderstanding of this theoretical enterprise.

Meanwhile, the phenomenalist repeats, in the case of the technical terms of science, the same agnostic criticism as that offered by the operationalist in the case of its theoretical propositions. In this view, it is simply a meaningless waste of time for scientists to debate the existence of enduring theoretical entities, regarded as external, objective things-in-themselves; just as it is similarly wasteful for them to interpret scientific theories as making, or denying, similar claims about the existence of objective, external states of affairs. Instead, the terms and concepts of science are all to be understood as the product of so many logical, or semantic, operations or constructions, and questions about their real existence are to be swept aside as damaging metaphysical superstitions.

Philosophical analysis and scientific practice. The arguments about these rival ontological and epistemological views cannot be safely left or judged without first looking more closely at the complex relationship between the general analytical interests of philosophers and the more specific intellectual concerns of working scientists themselves. For the degree to which each view about the reality of scientific entities and facts can carry conviction depends substantially on what branches of science are at issue. As the focus of philosophical attention has shifted historically from one scientific terrain to another, so, too, have the relative degrees of plausibility of these rival positions varied.

Quantum mechanics; taxonomy; perception

Since the 1920s, for instance, there has been a marked revival of philosophical discussion among scientists working in several specialized fields—particularly, among physicists concerned with the structure and development of quantum mechanics. In epistemic terms, the statistical character of quantum-mechanical explanations has prompted some fundamental questions about the status and limitations of human knowledge. Clearly, the extent and accuracy of human knowledge about nature are limited by the modes of operation of scientific instruments. Is it not also possible, however, that the significance of this statistical character lies at a deeper level? Perhaps the relevant objective relationships and states of affairs in nature itself are governed intrinsically by a merely probabilistic causality and so are essentially indeterminate. Or is there a point to be reached on the microphysical level at which any such distinction between subjective human knowledge and the objective state of affairs has finally broken down? The ontal implications of quantum mechanics have been as puzzling as the epistemic. Is an electron, say, a discrete particle that just happens to elude man's exact observation; is it an essentially blurred wave bundle having no precise dynamical characteristics; is it a concentration of probability, a mere theoretical symbol, or what? Or must one set all these ontological questions aside as lacking any significance for physics and as standing in the way of the physicist's proper task, that of extending the direct explanatory power of quantum-mechanical explanation itself?

Elsewhere, the philosophical debate about science has taken on other specific forms. Just as in Aristotle's natural philosophy the metaphysical controversy about Ideas and

essences was reflected in Aristotle's own methodological approach to biology and to the study of the natural relations and classification of organisms, so once again 20th-century reappraisals of traditional taxonomy—in the light of evolution theory, genetics, and population dynamics—have been an occasion for renewed philosophical debate. As a result, earlier disagreements about natural and artificial classifications have been reformulated and have generated a new dispute, about the possibility of basing taxonomy on a mathematical science of phenetics—in which the defining properties of different species, genera, etc. are all given quantitative numbers or measures—and so harnessing the technical resources of modern computers to its purposes.

Similarly, in the psychology of perception and related fields, the extension of understanding in recent years at last has permitted the framing of authentically empirical questions about perception and cognition, which lend themselves to direct investigation instead of being restricted to general a priori speculations. The result has been a theoretical debate, the final outcome of which will have profound effects on both philosophical epistemology and natural science. In areas of this debate where even Mach was content to pose entirely general questions, in the philosophical tradition of David Hume, about the role of sense impressions as the raw material of all cognition and perception whatsoever, it is now clear that many preliminary differences and complexities must be unravelled before one can hope to recognize the truly operative questions in this field. Far from all modes of knowledge and perception conforming to a single common pattern, man's sensory and practical dealings with the world call into play a variety of perceptual systems of which the operations justify no simple epistemological formula about impressions and ideas, sense-data and logical constructions, or intuitions and schemata. Thus, at the present time, the investigations of some physiologists, psychologists, and cyberneticists are bringing man's sensory and cognitive activities within the scope of natural science while at the same time preserving a feeling for the more general philosophical problems and insights of such philosophers as Locke and Leibniz, Hume and Kant, Helmholtz and Mach.

Pluralism and complementarity

At this point, the alliance between science and philosophy is simply carrying over into fields of science that are areas of methodological perplexity today the same interactions that were fruitful in earlier centuries within sciences having methods by now well understood. These interactions are unlikely to vindicate finally any one of the rival positions in the philosophy of science, whether ontological (Realist, instrumentalist, or phenomenalist) or epistemological (Realist, conventionalist, or operationalist). Probably such a vindication was, in any event, too much to expect. For in all the different special sciences—both natural and social—historical development eventually brings the investigator to a point at which he is ready to operate with a variety of technical terms or entities having very different logical characters and functions and at which his most general theoretical propositions or principles display corresponding differences in their logical status and implications.

So long as philosophical discussion is confined within the limits of an artificial, ideal language or propositional system, it is possible, perhaps, to continue posing purely abstract, general dilemmas about, say, theoretical entities or confirmation theory. But the bearing of such formal dilemmas on the actual content of contemporary scientific thought is becoming increasingly unclear. In debating the ontal status of theoretical entities, for instance, the question must at some stage be faced whether that phrase is intended to cover such notions as gene or pi-meson, species or cold front, momentum or superego, social class or economic market. (Certainly, not all of these terms have identical characters and functions.) In debating the epistemic status of scientific theories, likewise, it must be made clear whether one has in mind, say, the mathematical schema of quantum-mechanical field theory, the populational analysis of natural selection, the microstructures and mechanisms of molecular biology, the developmental sequences of cognitive psychology, the labour theory of economic value, the general regularities of terrestrial meteorology, or what. (Once again, not all of these theories have identical kinds of status or implications.)

Philosophical doctrines and approaches that carry great conviction when applied to the theories and ideas of one science may—not surprisingly—lose all of their plausibility when extended to other fields. Thus, an Empiricist analysis may apply quite straightforwardly to meteorology, yet entirely misrepresent the structure and implications of electromagnetic theory; while, in return, a Neo-Kantian account of theoretical physics may lack any direct relevance, say, to ideas about animal behaviour. Today as in classical Athens, analytical clarification in the philosophy of science goes, in this respect, hand in hand with methodological refinements in the sciences themselves. In retrospect, the methodological insights of Aristotle the marine biologist and of Plato the theoretical astrophysicist can be seen to have been complementary, rather than incompatible. Similarly, today, the philosopher must look at rival positions in the philosophy of science not merely as contradictory answers to technical questions within philosophy itself but equally as complementary contributions to the methodological improvement of theoretical understanding over the whole varied range of different scientific fields.

INTERRELATIONSHIPS OF SCIENCE AND CULTURE

Turn to broader human concerns

This survey has been concerned, almost exclusively, with philosophical problems and arguments about the sciences regarded as sources of theoretical knowledge. In pitting Realism against instrumentalism, mechanistic ideas against organicist ones, divine knowledge against human fallibility, or Platonic Ideas against Aristotelian essences, the philosopher is in each case concerned with the intellectual status, implications, and validity of certain general scientific concepts, methods, or entities. To confine oneself entirely to these intellectual aspects, however, would mean accepting a total abstraction of theory from practice and of scientific ideas from their behavioral expression. Thus, along with the present-day shift of emphasis from the physical to the human and social sciences, one finds that all such abstract approaches are coming once again under criticism, as over-intellectualizing the nature and implications of science.

Some of these attacks come from the neo-Marxist direction and reflect a traditional Marxian insistence on the unity of theory and action. (It was not for nothing that Lenin picked on Ernst Mach as a special target for scorn.) Analogous criticisms, however, are also coming from men with very different intellectual loyalties—*e.g.*, from the urban sociologist Lewis Mumford and from many contemporary Existentialists. In conclusion, therefore, a concise discussion is here given of some of the views about the relations between science and the rest of culture; *i.e.*, about the relevance of scientific knowledge to other spheres of experience and concern and, conversely, about the significance of broader, practical considerations for man's understanding of scientific theory itself.

The variety of these views has always been very great. Their exponents have ranged all the way from those who, like the energeticist Wilhelm Ostwald and the evolutionist Julian Huxley—both of whom rooted ethics in nature—present scientific ideas and procedures as rational panaceas for intellectual and practical problems of all kinds to those who, like Pierre Duhem and Carl von Weizsäcker, physicist and philosopher of nature, both of whom are theists, deliberately limit the claims of science so as to preserve a freedom of manoeuvre for ethics, for example, or theology. At each stage, most advocates of extreme claims for science have been ontological Realists; and, in strengthening their ontal and epistemic claims, they have also staked a claim to overriding intellectual priority on behalf of scientific knowledge, in contrast to other forms of experience. Similarly, those who would restrict the broader cultural claims of science have tended to be phenomenalists; and, in weakening their philosophical claims, they have also attempted to limit the authority of science to its own intellectual concerns as narrowly defined.

Science, technology, and ethical values

Whatever one's general philosophical position with re-

spect to the reality of scientific knowledge and entities may be, however, there are other more practical questions to be faced, questions about the specific implications of different scientific ideas and beliefs for parallel fields of human action and experience. On this point, one particular theme unites a wide range of radical critics of science, including both Lewis Mumford, U.S. social critic, and the Existentialists. Just as the Christian Dane Søren Kierkegaard, an early and seminal figure of Existentialism, condemned Kant's universalized system of ethics for ignoring the individuality of actual ethical problems and decisions, so today there is a widespread reaction against any tendency to treat social or practical decisions as technical matters, which can be left to the judgment of scientific or technological experts. The general methods of technology may, indeed, represent practical applications of the theoretical understanding arrived at by science; but all individual decisions about putting those general techniques to use—*e.g.*, in constructing an airport or power station—must be made not by appealing to any general formula or rule of thumb but by balancing a whole range of diverse considerations—economic and aesthetic, environmental and human, as well as merely technical.

According to another contemporary critique, the theoretical points of view adopted in natural science are general and abstract, but the practical demands of sociopolitical action and, *a fortiori*, of individual action, are concrete and particular; and, by itself, this contrast places an immediate restriction on the existential relevance of scientific ideas and engineering techniques. Such scholars as Thomas Huxley, a versatile scientist and defender of evolution, or Wilhelm Ostwald, a pioneer in electrochemistry, who viewed reality as essentially energy, might argue in general, abstract terms for interpreting ethical principles in evolutionary or thermodynamical terms if they pleased (so the critics continue); but such abstract speculative arguments have no bearing on the actual tasks of ethical decision and action. Here again, every ethical choice involves a unique constellation of considerations and demands; and this problem cannot be dealt with by appealing to any universal rule but must be appraised on an individual basis, as and when it arises.

Others take a more positive approach toward the contribution of science to an understanding of human values. Without necessarily claiming to transform ethics itself into a "science," they at any rate argue that the personal attitudes needed for effective work in science—adventurous skepticism and critical open-mindedness—have a wider relevance also to human conduct and social affairs. Supposing only that social and political discussion were conducted in this same tentative and critical spirit (they claim), its typical and deplorable passion and confusion could be replaced by the more rational consideration of the means required in order to achieve explicitly stated ends. While specific scientific ideas and doctrines may not be enough to direct social and political action by themselves, the scientific attitude may, nonetheless, have a profound significance for social policy and individual ethics alike.

Social significance of science

This contrast, between existentially minded critics of the claim that science is all-embracing and socially minded believers in the scientific attitude, may be epitomized by referring to contemporary discussions about the social significance of science itself. On the one hand, there has recently been a revival of explicitly anti-scientific views, which had been more or less dormant since the time of Blake, Johann Wolfgang von Goethe, and their successors in the Romantic movement. Supporters of this anti-science position point to the central role of military technology in the financial support of 20th-century scientific research and dismiss the average scientist's plea that he is not responsible for the uses to which his ethically neutral discoveries are put, as pallid and insincere. On the contrary (they argue), there is a long-standing and unholy alliance linking the collective institutions of the scientific and technological professions to the economic, industrial, and political powers that be. Faced with the fruits of this historical union (they conclude), it is time that scientists acknowledged their social responsibilities; and, failing better institutional controls, the outcome of this moral self-scrutiny may well prove to be a moratorium on further scientific research. Perhaps man already knows too much for his own good and needs to digest the significance of his existing stock of knowledge much further before adding to it and so widening yet again the gulf between theoretical knowledge and practical wisdom.

On the other hand, there are those who recognize science as playing a crucial role in modern society, but who go on to draw the opposite conclusion. Rather than putting a stop to science (these men would argue), its scope should be broadened; that is, scholars should be studying and understanding better the manner in which science serves as an element in the larger social order—perhaps by developing more adequate analyses of the social structure or perhaps by a large-scale extension of the methods of operations research. Aside from anything else (they point out), a moratorium on science is as impracticable as a moratorium on sin. It could be enforced only if political unanimity prevailed to an unimaginable degree among scientists. In the absence of such enforcement, liberal-minded countries will merely put themselves at a needless disadvantage—both economic and military—as compared with totalitarian states. Instead of pursuing this will-o'-the-wisp, scholars should put more effort into the task of understanding both the social preconditions of effective scientific development and the economic and political priorities involved in the practical application of scientific research.

Science and religion

As compared with the controversies of earlier centuries, the debate between science and religion is curiously muted today. There seems litle room nowadays for the theological passions that engulfed the discussion of Copernicus' new planetary theory, James Hutton's history of the Earth, or Darwin's theory of natural selection; and one would hesitate to speak any longer, as so many of our forefathers did, of warfare between science and religion as unavoidable. It is true that a few partisan writers can still find it a perplexing problem to decide such issues as whether the existence of life on other worlds would require a re-enactment there of the Christian fall and redemption or can insist—conversely—that the results of astronautical exploration refute any religious belief that God is an Old Man up in the sky. For most people, however, such questions have so far lost their earlier bite that they appear, by now, quite naïve.

What is the reason for this change? In earlier times, the term cosmology embraced not only the structure of the astronomical cosmos and the origins of the human species but also the religious significance of man's place in nature. Contemporary theologians, by contast, see physics and biology as having much less bearing on man's religious attitudes and preoccupations than their predecessors had supposed that they had. As a result, men's earlier ambition to construct a single, comprehensive world view, embracing the essential truths of both science and religion, no longer plays the active part in intellectual life that it formerly did. The only branches of science still capable of provoking vigorous theological debate, even now, are the human, rather than the natural sciences. The implications of Freudian psychology for the doctrine of grace and the use of psychedelic drugs for inducing quasi-mystical experiences are topics for live discussion today, in a way that evolution, astrophysics, and historical geology no longer are.

Limitations of science

This change of focus has been accompanied by a change in ideas about the intrinsic limits of science. It was formerly assumed that the boundaries between science and other aspects of human experience could be defined by marking off certain types of subject matter as essentially closed to scientific investigation. To one generation, the heart of this forbidden territory was the mind; to another, it was life; to a third, the creation. In this view, something in the essential nature of mental or vital activities, or in the origins of the present order of nature, made it impossible to treat these as phenomena open to study and explanation by the rational methods and intellectual procedures available to science. In fact, this view always had defects, from both the scientific and the theological points of view. To scientists, it seemed to impose an arbi-

trary restriction on their sphere of operations and so acted as a standing challenge and irritation. For theologians, it had the disadvantage of placing the essential claims of religion, so to speak, on a sandbank, where they risked being submerged in time by the rising tide of scientific knowledge. So, by tacit consent, the essential limits of science are now defined in quite different terms. These limits are now identified by recognizing that the character of scientific procedures themselves places restrictions on the relevance of their results. A scholar may choose to study whatever objects, systems, or processes he may please, but only certain of the questions that he asks about them will be answerable in the general, theoretical terms characteristic of science.

This change of approach may not have made the substantive problem—that of delimiting the frontiers of science exactly at all points—very much easier to deal with than it was before, but it has one genuine merit: it respects the crucial fact, to which attention has been drawn at several points in this present survey, that the distinctive features of science lie not in the types of object and event to which the scientist has access but in the intellectual procedures that his investigations employ and so in the kinds of problem that lend themselves to a scientific solution.

(S.E.T.)

Philosophy of nature

The discipline here entitled "the philosophy of nature" consists in the investigation of substantive issues regarding the actual features of nature as a reality and is divided into two parts: the philosophy of physics and the philosophy of biology. In this discipline, the most fundamental, broad, and seminal features of natural reality as such are explored and assessments are made of their implications for man's metaphysics, or theory of reality; for his *Weltanschauung,* or "world view"; for his anthropology, or doctrine of man; and for his ethics, or theory and manner of moral action. These implications are explored on the assumption that man's understanding of the natural setting in which his life is staged strongly conditions his beliefs and attitudes in many fields.

In its German form, *Naturphilosophie,* the term is chiefly identified with Friedrich Schelling and G.W.F. Hegel, early 19th-century German Idealists who opposed it to *Logik* and to the *Phänomenologie des Geistes* ("of the spirit or mind"). Employment of the term spread, in due time, beyond its narrower historical context in German Idealism and came to be used, particularly in Roman Catholic parlance, in the sense that it bears in this article (*e.g.,* the philosophies of physics and biology). Despite a notable decline in its usage in more recent years, the term is here employed, in the interest of the clear delineation of topics, as a complement to the philosophy of science, the discipline to which its subject matter has been allocated by recent philosophers. Thus in this work, the article on the philosophy of science is largely restricted to man's approach to nature, and thus to epistemological (theory of knowledge) and methodological issues, while that on the philosophy of nature encompasses the more substantive issues about nature as it is in itself.

PHILOSOPHY OF PHYSICS

Physics as a field of inquiry. *Essential features.* Physics is concerned with the simplest inorganic objects and processes in nature and with the measurement and mathematical description of them. Inasmuch as the binding forces of chemistry can now, at least in principle, be reduced to the well-known laws of physics, or calculated from quantum mechanics (the theory that all energy is radiated or absorbed in small unitary packets), chemistry can henceforth be considered as a part of physics in theory if not in practice. Moreover, it has become clear, through the general theory of relativity (which formulates nature's laws as viewed from various accelerating perspectives), that there is an aspect of geometry, too, that can be regarded as a part of physics. The fact that, over a wide range of circumstances, Euclidean, or ordinary uncurved, geometry presents a good approximation to reality is considered today not as a fact stipulated by a necessity of thought, nor a derivative from such a necessity, but as a fact to be established empirically; *i.e.,* by observation. In their application, the laws of Euclidean geometry refer to those experiences that arise with measurements of length and angle and optical sightings as well as with surface and volume measurements. The possibility—already extensively elucidated in antiquity—of deriving geometrical propositions by deduction from a few axioms, assumed without proof to be correct, had given rise in earlier philosophy to the opinion that the truth of these axioms must and could be guaranteed by a kind of knowledge that is independent of experience. The recognition of such a priori knowledge, however, has been superseded by the modern development of physics. While it is granted that a pure geometry is free to posit any axioms that it pleases, a geometry purporting to describe the real world must have true axioms. Today it is considered that, if Euclidean geometry is true of the world, this truth must be established empirically; the axioms would be true because the conclusions drawn from them correspond to experience. Actually, the world appears Euclidean, however, only when this experience is limited to cases in which the distances are not too great (not much greater than 10^9 light-years) and in which gravitational fields are not too strong (as they are in the vicinity of a neutron star).

Axiomatic versus empirical geometry

The possibility of deducing all known laws or regularities as logical inferences from a few axioms, which was discovered in Euclidean geometry, became a model for the construction also of another chapter in the history of physics. The classical physics of Newton, the 17th–18th-century father of modern physics, had employed Euclidean geometry as a foundation and had portrayed the solar system as a system of mass points subject to his mechanical axioms. The laws for falling bodies framed by the 16th–17th-century Italian physicist Galileo are the simplest logical consequences of Newton's axioms, and the laws framed by Johannes Kepler, a 16th–17th-century German astronomer, which precisely describe the motions of the planets, follow from them.

In addition to the laws of mechanics there are those of the broad sphere of electromagnetic phenomena as summarized in the equations of James Clerk Maxwell, a 19th-century Scottish physicist, which describe both the electric and magnetic fields and the laws of their mutual changes, equations that may thus be considered as the axioms of electrodynamics. Because they assume the mathematical form of partial differential equations—which express the rates at which differentials (small or infinitesimal distances or quantities) in several dimensions change with respect to their neighbours—electrodynamics is a local-action theory rather than an action-at-a-distance theory as in older formulations modelled after Newton's law of gravitation. The principle of local action states that the variations of electromagnetic magnitudes at a point in space can be influenced only by the electromagnetic conditions in the immediate vicinity of this point. The finite velocity of propagation for electromagnetic disturbances, which follows from this principle, leads on the one hand to the existence of electromagnetic wave events and on the other hand to conformity with the requirements of special relativity (a theory that formulates nature's laws as viewed from the perspectives of various velocities), which demand a maximum finite velocity for signals—the velocity of light in a vacuum.

Macroscopic versus microscopic physics

The most important division of physics today is one that replaces the traditional distinctions between mechanics, acoustics, and other classical branches of physics with that between macroscopic and microscopic physics, in which the latter investigates the conformity of atoms to law and their reactions in discrete quantum jumps, whereas the former extends from the level of ordinary human experience into astronomy to a total comprehension of the universe, attained through theoretical endeavours in the field of cosmology. Because it is now possible to observe especially bright objects (quasars) that are located perhaps 10^{10} light-years from the Earth, the possibility of empirically testing cosmological models is beginning to arise. In particular, the application of non-Euclidean, or curved,

geometries to the cosmos has suggested the conception of a finite, yet boundless, world space (positively curved), in which the maximum possible distance between two points would no longer be much greater than 10^{10} light-years.

Historical sketch. In the historical development of physics before the 17th century, geometry was the only field in which extensive advances were made; besides geometry, only the rudiments of statics (the laws of levers, the principle of hydrostatics of the 3rd-century BC scientist Archimedes) were clarified. After Galileo had discovered the laws of falling bodies, Kepler's laws describing the motions of the planets and Newton's reduction of them to a set of dynamical axioms established the science of classical mechanics, to which was annexed the investigation of electromagnetism. These developments culminated in the discovery of induction by Michael Faraday, an English physical scientist, the knowledge of local action by Faraday and Maxwell, and the discovery of electromagnetic waves by a German physicist, Heinrich Hertz. It was not until the 19th century that the law of the conservation of energy was first recognized as a general law of nature, through the work of Julius von Mayer in Germany and James Joule in England, and that the concept of entropy (see below *Entropy and the direction of time*) was formulated by Rudolf Clausius, a mathematical physicist. At the beginning of the 20th century, the German physicist Max Planck introduced the so-called quantum of action, $h = 6{,}626 \times 10^{-27}$ erg-seconds, which, when multiplied by the vibration frequency, symbolized by the Greek letter nu, ν, demarcates a basic packet of energy. Albert Einstein then extended the quantum theory to light. The real existence of atoms was proved by him and other investigators, and the science of microphysics thus arose. The researches of Niels Bohr on the quantum-theoretical significance of atomic spectra paved the way for broader search into the fine details of quantum laws, the final comprehension of which was introduced by Werner Heisenberg in 1924 and then systematically developed by Max Born, Heisenberg, and Pascual Jordan, of Germany, and by P.A.M. Dirac, of England. Moreover, Erwin Schrödinger, an Austrian physicist, pursuing a line of thought pointed out by Einstein and Louis de Broglie, arrived at results that were outwardly quite different from those of Heisenberg *et al.*, but were mathematically equivalent. The quantum mechanics, or wave mechanics, created by these men, which formulated quantum phenomena, were later extended to quantum electrodynamics.

General and special theories of relativity

Einstein's theory of relativity, first formulated in 1905, which was eventually extended from a special to a general formulation, brought about a revolutionary transformation in physics similar to that induced by quantum theory. The Newtonian mechanics of mass points turned out to have been merely an approximation to the more exact relativistic mechanics. The most important consequence of the special theory of relativity, the equivalence of mass (m) and energy (E),

$$E = mc^2, \tag{1}$$

in which c is the velocity of light, was formulated by Einstein himself.

After 1916 Einstein strove to extend the theory of relativity to the so-called general theory, a formulation that includes gravitation, which was still being expressed in the form imparted to it by Newton; *i.e.*, that of a theory of action at a distance. Einstein did succeed in the case of gravitation in reducing it to a local-action theory, but, in so doing, he increased the mathematical complexity considerably, as Maxwell, too, had done when he transformed electrodynamics from a theory of action at a distance to a local-action theory.

The great importance of physics for the technology that depends upon it—which has become a leading factor in the rapidly increasing development in the conditions of human existence—is shown historically in the close connection of decisive technical developments with basic advances in physical knowledge. Einstein's equivalence of mass and energy—to cite but one example—pointed to the atomic nucleus as an energy source that could be opened up through the study of nuclear physics. Moreover, the intellectual influence proceeding from physics and affecting the development of modern thought has become especially strong through the deepened grasp of the concept of causality that has followed from quantum theory (see below *Modalities of the natural order*).

Basic characteristics and parameters of the natural order. *Framework of the natural order.* Earlier mathematicians and particularly Richard Dedekind, a pre-World War I number theorist, have precisely defined the concept of real numbers, which include both rational numbers, such as $\frac{277}{931}$, expressible as ratios of any two whole numbers (integers), and irrational numbers, such as $\sqrt{27}$, π, or e, which lie between the rationals. By reference to these numbers, the Newtonian concept of space and time, which presupposes a Euclidean geometry of space, may be made precise: the values of the time t, ordered according to the ideas of earlier and later, can be made to correspond to the single real numbers, ordered according to those of smaller and larger. Also, the points on a straight line can be brought into correspondence with the real numbers in such a manner that the location of a point P between two other points P_1 and P_2 corresponds to a number assigned to P that lies between those assigned to P_1 and P_2.

Cartesian analytic geometry

Guided by the wish to find a method that allows the systematic proof of all philosophical truths, René Descartes, often called the founder of modern philosophy, established in the 17th century the analytic geometry of Euclidean planes. In it the points of a plane can be designated by two numbers x, y, their coordinates. One chooses two orthogonal coordinate axes, $x = 0$ and $y = 0$, like those of a graph, and, with any point P, associates its two projections, one upon each coordinate axis, which define the location of P. A curve in the $x - y$ plane is then expressed by an equation $f(x,y) = 0$, shorthand for any equation ("function") containing x's and y's. In the context of analytic geometry, every theorem of plane Euclidean geometry may be expressed by equations and thus be analytically proved.

This procedure can also be extended to three-dimensional Euclidean space by introducing three mutually perpendicular axes x,y,z. In this case, there are two different axis systems—either congruent or mirror reflections—analogous to right-handed and left-handed screws.

Metrology in modern physics

The simple space-time relationships of Newtonian physics have been changed in many ways by modern developments. The concept of simultaneity has been made relative by the special theory of relativity; every time measurement t is thus tied to a definite inertial system or moving frame of reference. It is accordingly appropriate to speak not primarily of points in time but of events, which are defined in each case by giving both a point in space and a point in time.

More specifically, an inertial system is a coordinate system that, relative to the fixed stars, is in uniform, straight-line motion (or at rest) with no rotation. In all inertial systems, Newton's principle of inertia, which states that all mass points not acted upon by some force persist in uniform motion with a constant velocity, is valid.

Hyperspherical space-time: x,y,z,u

Moreover, cosmological theories make it probable that space in the real astronomical universe corresponds only approximately to the relationships of Euclidean geometry and that the approximation can be improved by replacing Euclidean space with a space of constant positive curvature. Such a space can be mathematically defined as a three-dimensional hyperspherical "surface"

$$R^2 = x^2 + y^2 + z^2 + u^2 \tag{2}$$

in a hypothetical Euclidean space of four dimensions with mutually perpendicular x,y,z, and u coordinate axes.

The assertion that the foregoing statement has no operationally comprehensible content—*i.e.*, no content provable by performable measurements—is designated conventionalism, a view that is based on a remark by a French mathematician, Henri Poincaré, who was also a philosopher of science, that a fixed non-Euclidean space can be mapped point by point on a Euclidean space so that both are suitable for the description of the astronomical reality. The range of this remark is limited, however, in that this mapping, though it can indeed carry over points

into points, can in no way carry over straight lines into straight lines. Hence, many philosophers of science have held that, as long as astronomical light rays are held to be straight lines, the question of a possible curvature of space (*i.e.*, a deviation from Euclidean conditions) will by no means be solved by some arbitrary convention; that it signifies, instead, a problem to be solved empirically. If the universe in fact has a positive constant curvature, then every straight line has a length that is only finite, and its points no longer correspond, as in the Euclidean case, to the set of all real numbers.

In a very definite manner, cosmological facts have further indicated that time is by no means unlimited both forward and backward. Rather, it seems that time as such had a beginning about 10^{10} to 2×10^{10} years ago; thus, with an explosive beginning, the cosmic development began as an expansion.

The foregoing discussion has considered only the replacement of Euclidean spatial concepts by an elementary non-Euclidean geometry corresponding to a space with a constant curvature. According to Einstein, however, the fundamental idea of a still more generalized Riemannian geometry, so-called after Bernhard Riemann, a geometer and function theorist, must be brought into play in order to produce a local-action theory of gravitation.

Riemannian space: g_{kl}

Riemannian geometry is a further development of the theory of surfaces created by the 18th- and 19th-century German mathematician and astronomer Carl Friedrich Gauss, often called the founder of modern mathematics, a theory that aimed to investigate the curved surfaces of three-dimensional (Euclidean) spaces with exclusive regard to their own inner dimensions and no consideration of their being imbedded in a three-dimensional space.

Gauss thought that the points on such a surface could be specified by reference to two arbitrary coordinates u and v defined with the help of two single-parameter families of curves, u = constant and v = constant. The square of the infinitesimal distance between two adjacent points of the surface, ds^2, is then a quadratic form of the differentials du and dv, belonging to the pair of points, namely,

$$ds^2 = g_{11}du^2 + 2g_{12}dudv + g_{22}dv^2, \qquad (3)$$

in which the coefficients g_{kl} are functions of position. One can then calculate the curvature corresponding to the location of the pair of points according to a prescription given by Gauss, a curvature that measures the deviation from Euclidean plane behaviour that exists at this point. The curvature is a definite function of the g_{kl} and their first derivatives.

Riemann extended Gauss's considerations to the case of a three-dimensional space that can have different curvature properties from place to place (expressed by several functions of position that are collectively called the curvature tensor); and Einstein generalized these ideas still further, applying them to the four-dimensional space–time continuum, and thereby attained a reduction of the Newtonian action-at-a-distance theory of gravitation to a local-action theory.

Contents of the natural order. Among the most basic constituents of the physical world are symmetries, fields, matter, and action.

Symmetries

Symmetry is one of the chief concepts of modern mathematics, which combines the different symmetries belonging to an object or a concept into groups of relevant symmetries. The a priori investigation of the totality of possible groups, defined with respect to some operation (such as multiplication), comprises a division of modern mathematics called group theory.

Three-dimensional Euclidean space displays several important symmetry properties. It is homogeneous; *i.e.*, arbitrary shifts in the origin or zero point of the coordinate system produce no change in the analytic expression of the geometrical laws. It is also isotropic; that is, rotations of the coordinate system leave all geometrical laws in effect. Further, it is symmetric with regard to mirror reflections. It is tempting to suppose that these symmetry properties of space are also valid for the physical processes that occur in space, and this is indeed true over a wide range of cases, but not in all cases (for exceptions, see below *Charge/parity/time symmetry*).

Vector fields: directional quantities

That Newtonian mechanics and Maxwellian electrodynamics display in fact all of the symmetries of Euclidean space is revealed by the fact that they can be formulated in the language of vector analysis. Passing over the more familiar Newtonian mechanics, a few points about Maxwell's theory may be mentioned. This theory can be made to satisfy the requirements of operational thinking by ascribing to the electric and magnetic field strengths the significance of measurable physical realities, which makes it unnecessary to interpret them as states of a mysterious, hypothetical substance or ether, for which, in any case, the special theory of relativity (with the equivalence of all inertial systems) has no place.

Mathematically interpreted, a vector a represents a quantity with both magnitude and direction, which preserves its length or value and its direction when displaced. The vector field—*i.e.*, the association of a vector with every point in space (*e.g.*, electric field strength, or electric current density)—and the line integral (or summation) of a vector field V along a curve K leading from a point P to a point P' are basic concepts in vector analysis. To obtain the line integral, the curve K is divided into infinitesimal elements ds, the scalar (numerical or nonvector) product of ds with the value of V at that point is taken, and the results are summed with an integration.

A small surface area envisioned with a given sense of rotation around its boundary curve can also be described by a vector. In this instance, the vector, dF is perpendicular to the surface and forms, with the sense of rotation about the boundary, a right-handed system. Its magnitude is the area of the surface. The flux of the vector field V through the surface dF is called the scalar product $V \cdot dF$.

If V has the property that the line integral along every closed curve K is equal to zero, then V is said to be irrotational. This property is equivalent to the requirement that the vector field be a so-called gradient field; *i.e.*, that there exist a scalar field quantity W with the property that the difference in the value of W at two points P and P' is equal to the line integral of the vector field V from P to P' (along any arbitrary curve K). If V is, for example, an electrostatic (charged) field, the significance of being irrotational is that one can gain no mechanical work in leading a small test charge around any closed curve; the work involved is equal to zero. For an unclosed curve K, however, the movement of the test charge yields an amount of mechanical work that is proportional to the potential difference between the endpoints of the curve. The components of the gradient of W, expressed in partial derivatives ∂, are

$$\text{grad } W = \frac{\partial W}{\partial x}, \frac{\partial W}{\partial y}, \text{ and } \frac{\partial W}{\partial z}. \qquad (4)$$

Curl; divergence; charge; and current density

If the vector field is not irrotational, there can then be constructed from it an adjunct rotational field, called curl V, by considering a small (infinitesimal) surface area dF located at a point P and forming the line integral of V along the boundary curve of dF. Then, when this line integral is divided by the magnitude of the surface area, the component of the curl V parallel to the vector dF is obtained.

On the other hand, the flux of a vector field V out of a closed surface can be formed by integration. If this flux is always zero (for every choice of a closed surface), V is called source-free. Otherwise, there is a so-called divergence of V at a point P, which is defined as follows: one divides the net flux of V out of a small surface that surrounds P by the volume enclosed by the surface. The limit of this quotient for infinitesimally small surfaces is called the divergence of V at P or the source field div V.

The formulation of the basic laws of electrodynamics given by Maxwell is called the Maxwell equations. These equations contain, for example, the statement that, in a vacuum, the source field of the electric field strength is proportional to the spatial electric charge density, symbolized by the Greek letter rho, ρ, and that the magnetic field strength is source-free (divergence equal to zero). Thus, magnetic monopoles having no correlate of opposite sign

do not exist. Remembering that every source-free vector field may be expressed mathematically as a rotation field (and vice versa), it is possible to derive the magnetic field strength H as a rotation field from a vector field A, which is usually called the vector potential of H

$$H = c \cdot \operatorname{curl} A. \tag{5}$$

The fundamental law of the conservation of charge results from Maxwell's equations in the form of the continuity equation

$$\dot{\rho} + \operatorname{div} i = 0, \tag{6}$$

in which ρ is the time derivative of the charge density, and the vector field i is the electric current density.

In the case of a vacuum, the Maxwell equation that expresses Faraday's law of induction takes the form of a proportionality between the rotation field of the electric field strength and the time derivative of the magnetic field strength:

$$H = c \cdot \operatorname{curl} E. \tag{7}$$

It is a significant fact that Maxwell's theory leads to a localization of energy, which in electromagnetic fields is propagated somewhat in the manner of a substance, with a density that, for the vacuum case, is

$$W = \frac{E^2 + H^2}{8\pi}. \tag{8}$$

There remains also the unsolved problem of clarifying the relation of gravitation to quantum theory, which is much aggravated by the fact that gravitational energy allows of no similar localization.

In both mechanics and electrodynamics, the fundamental equations have such a form that they can be understood as the conditions for a variational or an extremal principle: that, through the fulfillment of these conditions, a certain integral receives an extreme value. This integral, which has the dimensions of action—*i.e.,* of energy times time—is one of the most fundamental quantities of nature. Although the concept of action is less obvious to man's physical intuition than that of energy, it is of even greater significance, as it appears also in connection with the quantum laws. For the basic constant of all of quantum physics, which always occurs in the laws of this domain, is likewise of this dimension: namely, Planck's quantum of action

Matter and action

$$h = 6.626 \times 10^{-27} \text{ erg-seconds.} \tag{9}$$

Modalities of the natural order. In a purely phenomenalist theory of matter—*i.e.,* a theory that does not go into the details of atomic physics but considers matter only in a first approximation as a spatially extended continuum—numerous material properties are ascribed to every type of matter, properties such as density, electrical conductivity, magnetizability, dielectric constant, thermal conductivity, and specific heat. To be complete, a theory must provide a means of deriving all of these material properties theoretically from the laws of atomic physics.

The hiatus-free causality envisioned throughout the science of physics before the rise of quantum theory cannot be separated conceptually from the far-reaching assumption that all physical processes are continuous. It had been supposed that continuous changes in antecedent causal processes would issue in continuous changes in the sequence of processes that are causally dependent upon them. Quantum physics, however, has expressly breached the old philosophical axiom that *natura non facit saltūs* ("nature does not make leaps") and has introduced a granularity not only in the matter filling space but also in the finest processes of nature. It is therefore only logical that, with respect to causality, the quantum theory would arrive at new and modified ideas as well. Renouncing unbroken causality, it speaks only of a probability that is statistical and a predetermination for the discrete saltatory events of which physical processes consist—a view that must now, in spite of Einstein, be regarded as irrevocable.

Causality and statistical approach

The special theory of relativity demands that the fundamental validity of the local-action principle be acknowledged: all actions have only finite velocities of propagation, which cannot exceed the velocity of light. Thus, in relativistic cosmology it is quite possible that two partial regions of the total spatial manifold may exist between which no causal interaction can occur: causal influences could then assert themselves only inside the so-called interdependent regions in the space–time manifold. These remarks also apply to the quantum theory, in which, however, instead of a causal dependence of physical processes upon each other, there is only an induction of statistical probabilities for possible quantum transitions.

Levels of the natural order. Moving in quite different directions, the theory of relativity on the one hand and the quantum theory on the other have diverged from the earlier ideas of classical physics, which were considered unalterable. There are some physical problems, however, that can be thought through only by appealing to both the relativistic and the quantum-theoretical modifications. A so-called joint relativistic and quantum-mechanical theory suitable for such problems is quantum electrodynamics, the development of which, however, is not yet complete. Its development was greatly hindered at first by certain mathematical difficulties (so-called divergences), which it later became possible to mitigate by renormalization—*i.e.,* by a technique of correcting the calculated results. The more generally conceived quantum theory of wave fields finds a broad area of possible application in the physics of the different kinds of elementary, though short-lived, particles produced by the huge high-energy accelerators. In its final form, the theory of elementary particles should not only formulate, in general, the laws valid for all known elementary particles but should also allow a deductive derivation for all possible kinds of elementary particles—analogous to the derivations of elements in the periodic table. Heisenberg endeavoured to set up this far-reaching problem, which has been called the world formula, for a solution. Imposing mathematical difficulties, however, have arisen in the attempt to clarify its consequences for a quantitative comparison with experience, and considerable further work may still be required.

Special problems in the philosophy of physics. *Problems at the formal level.* Euclidean space, in contrast to imaginable spatial structures that deviate from it, is distinguished by the simplicity of the topological properties (those preserved through rubberlike stretching and compressing, but without any tearing) that arise from its unusually simple continuity relationships. One may ask, then, whether the empirical knowledge of modern physics gives any cause to consider deviations from the topological relations of Euclidean space. The American physicist John A. Wheeler, author of a new theory of physics called geometrodynamics, has speculated about this question. In particular, he has pointed to the possibility of so-called worm holes in space, analogous to the way in which the cylindrical surface of a smooth tree trunk is changed topologically if a worm bores a hole into the trunk and emerges from it again elsewhere: the surface of the trunk has thus obtained a "handle." Similarly, one can envision certain handles being added to three-dimensional Euclidean space. Whether this hypothesis can be fruitful for the theory of elementary particles is yet to be determined. From the methodological and epistemological standpoints, it is obvious that a geometrical structure is here being assumed, the measurement of which is fundamentally hindered by the lack of rulers with calibrations smaller than the structure itself. Presumably, the practical possibility of appealing to such topological modifications of the ordinary notion of space is to be found in astrophysics rather than in elementary particle physics. Viktor A. Ambartsumian, an Armenian-born astrophysicist, is convinced that the processes involved in the origins of galaxies are connected with explosions in which the matter of new stellar systems arises from prestellar material; it has been found tempting to suppose that this prestellar material exists in regions with unusual topological properties.

Special topological properties

The basic idea of the special theory of relativity can also be understood as a statement about the symmetry properties of the four-dimensional space–time manifold. The special principle of relativity states, in fact, that the same physical laws are valid in all of the various inertial coordinate systems—in particular the law that the velocity of light in a vacuum always has the value c. This equivalence of the space–time coordinates x,y,z,t with other coordinates x', y', z', t' that are linear, homogeneous functions of the unprimed coordinates can be expressed by the equation

$$x^2 + y^2 + z^2 - c^2t^2 = x'^2 + y'^2 + z'^2 - c^2t'^2. \qquad (10)$$

In this formulation, the isotropy of space—its sameness in all directions—appears as a special case of a more comprehensive symmetry property of the space–time manifold. When $t = t'$, the special case of a purely spatial rotation of coordinates is obtained; and in the general case, in which the primed coordinates are moving with velocity u with respect to the unprimed, the famous Lorentz transformations are obtained, which, to adjust to the finiteness of c, add a factor, symbolized by the Greek letter gamma, $\gamma = \sqrt{1 - (u^2/c^2)}$, to the ordinary Galilean transformation, $x = x' + ut'$, thus yielding

Lorentz transformations

$$x = (x' + ut')\gamma. \qquad (11)$$

The group of symmetries of the four-dimensional space–time manifold thus produced is called Poincaré group.

Problems at the quantum level. Problems of particle theory, complementarity, and symmetry have arisen in studies at the quantum level.

Whereas the atomic nuclei beyond hydrogen-1 (the proton) are compounded structures, consisting of neutrons and protons, modern physics also deals with numerous elementary particles—neutrinos; π (pi), μ (mu), and κ mesons; hyperons; etc.—that are thought of as uncompounded. The elementary particles of each particular kind show no individual differences. Each elementary particle has a corresponding antiparticle, which, for charged particles, always carries a charge of opposite sign. (The γ [photon], π^0, and Z^0 particles are understood to be their own antiparticles.) Whether Heisenberg's world formula can provide a complete framework for all possible kinds of elementary particles is undecided.

Elementary particle theory

Every type of elementary particle has a definite value for its spin, either integral (*e.g.,* photons) or half-integral (*e.g.,* electrons, protons, neutrinos). Particles with half-integral spin obey Fermi–Dirac statistics; those with integral spin obey Bose–Einstein statistics, which differ in form as $u/(1+u)$ differs from $u/(1-u)$—u being any function. The conformity to law that underlies the Fermi–Dirac statistics for electrons was first recognized by Wolfgang Pauli and formulated as the Pauli exclusion principle, which played a decisive role in settling upon the shell structure in the periodic system of the elements.

The basic duality of waves and corpuscles is of universal significance for all kinds of elementary particles, even for composite particles in those experiments that cannot lead to a breakup of the particles into their component parts.

An electron (and analogously any other elementary particle or even, for example, an alpha particle) can appear just as well in the form of a wave as in that of a localized corpuscle. In an idealized thought experiment, one can imagine that the position of an electron can be ascertained with a gamma-ray microscope. If the electron is described in terms of wave processes in the sense of Schrödinger's wave mechanics, a very sharply concentrated wave packet appears at the stated position. In an investigation of this packet by Fourier analysis—a technique that analyzes a function into its sinusoidal components—wave components of quite different wavelengths occur; thus an electron in this condition has no definite value for its de Broglie wavelength and consequently none also for its translational momentum. Then, as stated in the so-called de Broglie relation,

$$\lambda = \frac{h}{mv}, \qquad (12)$$

there is for an electron moving free from impinging forces a corresponding wavelength, symbolized by the Greek letter lambda, λ, that is inversely proportional to its momentum mv. And conversely, an electron moving inertially with a definite momentum (which in the limiting case of small velocities is equal to the product of the mass and the velocity vector) has no definite position. If an electron that is moving inertially (especially an electron at rest) is constrained by the use of a gamma-ray microscope to "make up its mind," as it were, on a location, then the probability of its appearance at a point in space is the same for all locations. More precisely stated, the probability of the appearance of the electron in a definite volume is proportional to the magnitude of this volume.

In particular, it will be helpful to consider an electron moving in the x-direction and to suppose that it has a wave amplitude that depends only upon x, an electron in which the most representative wavelengths are confined to a narrow interval while the amplitudes that are discernibly different from zero are likewise confined to a certain interval Δx. If, on the other hand, Δp is the range of discernible momentum values—computed from the discernible wavelengths that represent them according to the de Broglie relation (12)—then the product of the uncertainties Δx and Δp cannot be smaller than Planck's fundamental quantum of action h. This statement comprises the famous Heisenberg uncertainty relation, which expresses the "complementarity" of position and momentum—as Niels Bohr characterized it.

Heisenberg's uncertainty principle

If one assumes, as above, that all physically possible states of an electron can be represented by Schrödinger's wave mechanics, then the complementarity of the position coordinate x and its corresponding conjugate momentum p_x is a simple mathematical fact. When one thinks primarily of physical-measurement experiments, it should be emphasized that stringent limitations are imposed on the simultaneous measurement of the position and momentum of a particle which, according to the uncertainty principle, make it impossible to measure simultaneously both of these complementary quantities with unlimited precision. In an experiment that measures its position, the electron is forced into a sharper localization; and its particle nature is evident. By contrast, in an experiment that measures its momentum, an interference experiment is involved; the electron must be able to display a certain wavelength, which requires an adequately extended region in space for its reacting. These two complementary and opposing demands can be brought into harmony only in the sense of a compromise; and the Heisenberg uncertainty relation formulates the best possible compromise.

Thus, it becomes at the same time clear that the state of the electron given in the wave-mechanical description before carrying out a new measurement experiment can establish only a statistical prediction for the result. The probability density for the appearance of an electron at a point in space is given by the square of the absolute value of the (complex) Schrödinger wave amplitude; for a definite result in measuring a wavelength or a momentum, the square of the absolute value of the (complex) Fourier coefficient belonging to it provides the standard. The general statistical transformation theory of quantum mechanics (as developed by Dirac and Jordan) gives a complete review of the measurable physical quantities for a microscopic mass point (or a system of such points). According to this theory, two different measurable quantities A and B can be simultaneously determined with unlimited precision only if the operators or matrices that describe A and B commute—*i.e.,* if $AB = BA$.

A transformation in which the nucleus emits an electron and a neutrino is called beta decay, an example of nuclear radioactivity. The forces that thus come to light are those of the so-called weak interactions. It has been experimentally determined that for these forces the symmetry associated with reflections in a mirror does not hold. At least in certain circumstances, however, a remnant of this symmetry continues to hold, in which the so-called CPT (for the initials in charge/parity/time) theorem applies. This theorem states that basic physical laws remain of unaltered validity when a reflection of the space coordinates

Charge/parity/time symmetry

as in a mirror is combined with an interchange of positive and negative charge (which is largely synonymous with the interchange of particles and antiparticles) and with a reversal of the direction of time. Whether or not this symmetry law is valid without exception is by no means fully clarified at present.

Problems at the macrophysical level. Proceeding from the properties of atoms and molecules that are described in terms of quantum theory, a theory of macrophysical substance aggregates has been built using statistical mechanics. The theories of heat, of gases, and of solid-state aggregates (crystal lattices) have been extensively clarified. Only the liquid state still poses certain unsolved problems for the statistical theory of heat.

Mach's principle of cosmic influence

In any case, Newtonian mechanics may be derived as a macroscopic consequence of the laws of the mechanics of atoms, and its validity for the motions of astronomical bodies presents no problem. It is not so simple to prove, however, that the statements of Newtonian mechanics for rotating bodies (*i.e.,* the mechanical laws of centrifugal force and the Coriolis force) may be established from Einstein's general theory of relativity. Ernst Mach, a physicist and philosopher of science whose train of thought has substantially fertilized the modern development of physics from the point of view of the theory of knowledge, raised objections against Newton's idea that centrifugal force is a consequence of the absolute rotation of a body; he asserted instead that the rotation of a body relative to the very distant giant mass of the universe was the true cause of centrifugal force. This idea, often referred to as Mach's principle, has been corroborated, though in a different form, within the conceptual framework of Einstein's general theory. An irrotational coordinate system—specifically, a system not rotating with respect to the fixed stars (or, better, to the spiral nebulae)—is distinguished from a rotating system by the difference in the metric field for the two cases (*i.e.,* in the properties of their respective space-times as expressed by the equation—(3) above—for the interval between two events). It is true that the local metric field (in the vicinity of the solar system) is influenced by the distant masses of the universe, but of course only in the sense of a local-action principle and therefore in no way such that the metric in the solar system is directly given as a function of these distant masses and of their motions.

The question of the precise circumstances in which Mach's principle can still be defended on the basis of Einstein's theory is somewhat complicated and thus remains obscure. In any case, it is certain that a deduction of this principle from Einstein's theory can only be given in conjunction with a complete solution of the cosmological problem; *i.e.,* of the problem of what are the overall geometric and dynamic properties of the universe considered in its totality. The remaining problems involved in justifying the application of classical Newtonian mechanics in astronomy by means of Einstein's theory contain, however, no additional fundamental difficulties.

Entropy and the direction of time

There is one more influence of cosmological relationships upon macroscopic physics, which arises in connection with thermodynamics. The existence of irreversible processes in thermodynamics indicates a distinction between the positive and negative directions in time. As Clausius recognized in the 19th century, this irreversibility reflects a quantity, first defined by him, called entropy, which measures the degree of randomness evolving from all physical processes by which their energies tend to degrade into heat. Entropy can only increase in the positive direction of time. In fact, the increase in entropy during a process is a measure of the irreversibility of that process. In contrast, it is true of the quantum theory of the atom that the positive and negative directions in time are equally justifiable (in the sense of the principle of CPT symmetry). Consequently, it is difficult to understand how statistical mechanics can make possible a thermodynamics in which the entropy grows with time.

It is true that there are fluctuating thermodynamic phenomena, even in a system in overall thermodynamic equilibrium—and here theory and experiment agree. Thus, the states that arise within any small partial volume of the system may be not only those that are thermodynamically most probable but also transitory deviations from the most probable state. The mention of these fluctuations, however, does not help to remove the above paradox.

Most physicists now hold that, until recently, this problem was treated erroneously in the usual textbook presentations. In the statistical theory of heat, entropy was regarded as proportional to the logarithm of the thermodynamic probability, and students came to regard it as a necessity of thought that nature progresses from states of lower probability to states of higher probability. In truth, however, the increase of entropy is a real physical property of the positive direction in time. Nonetheless, it was supposed that shaking a vessel containing red and white balls (or even grains of sand) in an originally ordered condition with the two colours neatly separated must result in a condition of extreme intermixing of balls. This result does not correspond, however, to some necessity of thought but to an empirical property of the real universe in which men live and experiment.

Boltzmann's theorem

This interpretation, which was held for a long time and has only quite recently been recognized as erroneous, was allegedly supported by a famous mathematical theorem of Boltzmann, which seemed to show that, in an ideal gas for which the entropy—measured by its number of particles and its total energy—was not yet at its maximum value, the entropy must increase. If collisions of gas molecules are characterized by velocity vectors that are mechanically allowable (both before and after the collision), and if these vectors must satisfy both energy and momentum conservation, then what Boltzmann actually proved is that the entropy increase follows only when a correct count of the collision rate is made, according to which every kind of collision of gas molecules has a frequency of occurrence proportional to the product of the number of collision pairs that were present and the velocities that existed before the impact.

As one can subsequently see, the positive direction in time is already marked out by the collision rate count in a manner that no longer corresponds to the CPT principle. Although it is in fact possible to reason out the continuous increase in entropy on this basis, the paradox is not overcome. The question then remains of how it is physically justifiable—*i.e.,* how it can correspond to reality—to regard this principle of collision rates as valid even though it fundamentally contradicts the CPT principle.

An answer to this paradox can now be given, thanks to the insight of modern theoreticians—among them Hermann Bondi, a mathematician and cosmologist—who have shown that the entropy principle must be understood in the sense that in the universe as a whole, one definite time direction is singled out, namely, the one for which the universe expands. The thermodynamic distinction of a positive direction in time—with increasing entropy on the macroscopic level and with collision rate counts on the microscopic level—results from an expansion of the universe. Surprisingly, the Hubble expansion of the system of all the galaxies—so named after Edwin Hubble, an extragalactic astronomer—thus displays physical effects right down to the level of everyday physics; specifically, when two bodies at different temperatures are brought into thermal contact, the temperature equalization that results is an irreversible process corresponding to an asymmetry of the positive and negative directions in time that depends upon the expansion of the universe.

Problems at the cosmological level. A mathematical discovery by Alexander Friedmann has become of great significance for the mathematical derivation of cosmological models from Einstein's general theory of relativity. According to Friedmann, if the average mass density is constant throughout space, the gravitational field equations can be satisfied by a metric that embraces a three-dimensional space of constant curvature together with a time coordinate t such that the radius of curvature $R(t)$ is a definite function of time; and these cosmologies turn out differently depending upon whether the curvature of space is positive, negative, or zero. Among the models of the universe that are mathematically allowable are models in which the time coordinate may run through all values

from zero to infinity, models in which the time is limited to a finite interval, and models in which it may run from minus infinity to plus infinity.

The steady-state theory

For a time, many specialists working in the field of cosmology found the so-called steady-state theory, first projected by an astronomer, Sir Fred Hoyle, especially convincing. In a modified version, this theory was adapted to the Friedmann model by Bondi. By adopting the so-called perfect cosmological principle, which holds that the broadest features of the universe are the same at all times as well as at all places, the theory then satisfied the unusually high symmetry or homogeneity requirements not only of a three-dimensional space with constant time but also of the entire space-time manifold. This high-degree homogeneity was so convincing to many authors that, in deference to it, a fundamental deviation from Einstein's field equations was tolerated: Bondi and Hoyle supposed that a small but constant creation of hydrogen occurs in the intergalactic vacuum. This hypothesis was introduced in order to achieve, in spite of the Hubble expansion of space, a mass density that remained constant in the universe.

This theory, which in spite of its deviation from Einstein's field equations certainly advocates an allowable hypothesis worthy of consideration, no longer seems tenable, however, because of the discovery of background radiation with a present temperature of 3° Kelvin, which is interpreted as a remnant of an original "big-bang" beginning of the universe. It thus appears that it is no longer possible to uphold the steady-state theory or the perfect cosmological principle upon which it is based. Instead, one must favour either a Friedmann model, which has a beginning, from which it expands monotonically and without limit; or a Lemaître model, in which a quantity lambda, λ, called the cosmical constant, arises that is, mathematically, a constant of integration, and physically, a force of cosmic repulsion that partially neutralizes that of gravitational attraction, and which lends a curvature to space even in its empty regions. For both of these models the time coordinate increases without limit from some initial value, which would naturally be called zero. For the beginning of time, one thinks, moreover, of a singularity $R(0) = 0$ and thus of a space that at the null point of time is still a mass point. Cyclical models that alternately expand and contract in an endless sequence have also been discussed.

The empirical cosmological data, some of which, indeed, are more estimated than ascertained, seem to suggest that, in the present-day universe, the positive energy corresponding to the total rest mass of all the material existing in the universe may be exactly equal to the negative gravitational energy existing in the universe; thus, the total energy would then be equal to zero. This interesting singularity, however, needs further support. At one time, Dirac advocated the speculation that the total mass of the universe is not constant in time but is increasing—at a rate somewhat slower, however, than that in the steady-state theory. Ambartsumian's notion concerning prestellar material, which was mentioned above (see *Problems at the formal level*), could perhaps be considered support for this idea. Many further discussions have followed another conjecture by Dirac, according to which the gravitational constant G should be liable to change in the course of cosmic development. This constant would thus have to be considered a scalar field quantity, which in a Friedmann universe is approximately independent of the three space variables but dependent on the time variable. In spite of extensive theoretical deliberations on this theme, no decision has yet been reached.

Fundamental ratios

The way has been opened for some fundamental conjectures on certain emerging themes by the fact that the product of the mean mass density in the universe and the gravitational constant has the same order of magnitude as the square of the reciprocal of the radius of curvature of the universe. The aforementioned relation between the mass and gravitational energy in the universe presents a different expression for this ratio. The total mass of the universe divided by the proton mass probably has approximately the order of magnitude 10^{80}, according to present cosmological notions. The order of magnitude of the radius of curvature of the universe is approximately 10^{40}, when expressed as a multiple of an elementary length of which the order of magnitude is approximately that of the nuclear radius. Whether it is justifiable to presume that there is here a functional dependence—*i.e.*, a proportionality of M to R squared—is a question for the present still undecidable. The speculative attempt of Dirac to find an answer, however, is still—at least provisionally—judged with skepticism by the majority of physicists. (P.W.J.)

PHILOSOPHY OF BIOLOGY

The sharp increase in man's understanding of biological processes that has occurred in recent years has stimulated philosophical interest in biology to an extent unprecedented since the development of evolutionary theory in the 19th century. Biologists and philosophers alike have devoted much attention to a variety of issues regarding the subject matter and the methodology of biology, resulting in a sizable output of written material, formulating philosophical questions that are still arising and framing answers to acknowledged difficulties. Most of the problems of the philosophy of biology are old questions now being investigated afresh in the light of biological advances and new standards of philosophical rigour. In this account contemporary questions will be stressed.

The range of topics. An investigation of recent writings in biophilosophy reveals a continued preoccupation with unanswered—some say unanswerable—questions about evolutionary theory and a growing concern for a critical reappraisal of the question of whether biology is an autonomous discipline unamenable to reduction to mere physical and chemical underpinnings. Until the mid-20th century the biological sciences suffered from a lack of attention by philosophers; the principles that were generated were far less rigorously examined than were those of the physical sciences. There is now renewed hope, however, for a fresh approach to the age-old puzzles regarding life and its raison d'être. This hope rests on the recrudescence of interest in all biological matters as a direct result of an increased understanding of biological processes, of the changing quality of life, of the growing awareness of man's stewardship of the Earth, and of the exploration of space. Biology has just begun to make the sort of impact that the physical sciences have already made. It has generated a life technology with genetic engineering, organ transplants, and artificial organs. Each innovation, each technical masterstoke, each conceptual knot united emphasizes the need for a definitive philosophy of biology, and developments toward this goal are now under way. Good biological work has been accomplished by investigators with varied philosophical outlooks ranging from Neo-Thomism to skeptical naturalism. No inevitable metaphysics evolves from the study of biology or any other natural science; nevertheless, some of the general conclusions of biology have a philosophical interest, defining the limits of reasonable belief about the nature of the living world.

Natural distinctions

Categorical discontinuities that are recognized for the purpose of scientific methodology often seem impossible to justify as "natural" distinctions. Many biologists have noted, for example, that it is easier to study life than it is to define it. Properties such as metabolism and reproduction undeniably characterize organisms and might be said to define them, yet such a definition is arbitrary to the extent that such properties are logically independent. What is true of all life forms today may not have been true of the very earliest ones and, what is more, might not be true of extra-terrestrial ones that might be encountered in the future. There is not as yet a set of nonarbitrary characteristics that mark the distinction between living and nonliving systems. Moreover, in the course of analysis, it becomes necessary to arrange all of the phenomena of nature in a more or less linear, continuous sequence of classes and then to describe events occurring in the class of more complex phenomena in terms of events in the classes of less complex phenomena (principle of hierarchical continuity). Within each class, however, there are numerous interrelations observed between events of the same order of complexity. It is thus possible to recognize a number of

Living and nonliving systems

more or less autonomous disciplines, each permitting generalization, but ordered so that the more complex events treated by one discipline can also be analyzed in terms of less complex events treated by another discipline. It is possible, for example, to establish a body of generalizations about human society independent of the behaviour of individual persons; a number of generalizations about individual behaviour without consideration of the physiology of the sensory, conductor, and effector mechanisms involved; and a large body of generalizations about muscle or nerve physiology without considering the molecular mechanisms involved. A particularly striking feature of the hierarchy is that an increase in complexity is coupled with the emergence of new characteristics. The origin and development of life from small systems that synthesized biochemicals to organisms that perform highly complicated functions suggests that the hierarchical arrangement of nature and the sciences is correlated with the temporal order of evolution. The maintenance of a steady state by metabolism, reproduction, responsiveness, modification of response by experience, tradition, and social phenomena are just some of the more dramatic examples of emergent phenomena. The emergence of new qualities as evolution proceeds might generally characterize the universe.

Moreover, photosynthesis, on the one hand, and reproduction followed by natural selection, on the other, provide a mechanism by which physically less probable systems can emerge locally from physically more probable ones. Though it frequently has been supposed that physical evolution is at an end, there is no reason to suppose that this is true of social development, for which Sir Julian Huxley, a biologist, philosopher, and educator, provided an evolutionary context. In his Romanes Lectures, published in 1943, Huxley wrote:

> It is only through social evolution that the world-stuff can now realize radically new possibilities. Mechanical interaction and natural selection still operate, but have become of secondary importance. For good or evil, the mechanism of evolution has in the main been transferred [in man] onto the social or conscious level. . . . The slow methods of variation and heredity are outstripped by the speedier processes of acquiring and transmitting experience. . . .
>
> And in so far as the mechanism of evolution ceases to be blind and automatic and becomes conscious, ethics can be injected into the evolutionary process. Before man that process was merely amoral. After his emergence onto life's stage it became possible to introduce faith, courage, love of truth, goodness—in a word moral purpose—into evolution. It became possible, but the possiblity has been and is often unrealized.

It may well be that social evolution is only in its early stages. These stages, moreover, have for the most part taken place in a period during which systematic knowledge was undeveloped. A Russian mineralogist, Vladimir Vernadsky, the founder of biogeochemistry, regarded the envelope of the Earth as passing from a stage determined primarily by biological processes to one determined by conscious human effort. He called this layer of consciousness the noösphere. The concept was later extended, notably by Pierre Teilhard de Chardin, a French priest and paleontologist, who began the building of a new philosphic bridge between biology and religion.

The nature of biological systems. Questions about the character of biological systems and of the biological world merge with those about the concepts and methods required for their understanding. Although the two cannot be wholly separated, in this account those matters most clearly related to the substantive philosophical aspects of biology will be stressed. Methodological issues will be touched on in respect to the concepts being elucidated.

Origin and definition of life. Space exploration has directly influenced the development of life-detecting devices. This technological need spurred intensive study regarding the kinds of evidence living things display reflecting their aliveness. In his *Chance and Necessity* (1972) Jacques Monod, a biologist, deals with the invariance of genetic endowment, morphological autonomy, reproductive invariance, and teleonomy (the tendency to have a purpose or project written into their molecules) as the major properties of living systems; he considers that they involve the chance and necessity that determine the course and character of the entire biological world.

Philosophic views

Philosophers have long deliberated over the definitive features of living systems. The distinction between living and nonliving, which was widely discussed at the turn of the 20th century, has lost much of its interest for current biology. A growing conviction, intuitively felt by many biologists, is that no clear line can be drawn between the living and the nonliving. The bridge between what is and is not obviously alive consists of a range of problematic agents, including viruses and genes, which appear to be living at times and nonliving at other times.

Viewpoints on the nature of life. Basically and traditionally, there are three distinct philosophical stands regarding the biological nature of life: vitalism, mechanism, and organicism.

Vitalism

Essentially, vitalism holds that there exists in all living things an intrinsic factor—elusive, inestimable, and unmeasurable—that activates life. In its classic form, as espoused by many biologists at the turn of the 20th century—in particular, by Hans Driesch, a German biologist and philosopher—it has suffered severe criticism. Ernest Nagel, a philosopher of science, rang its death knell in 1951, when he wrote in *Philosophy and Phenomenological Research* (11:327 ff.):

> Vitalism of the substantival type . . . is now a dead issue . . . less, perhaps, because of the methodological and philosophical criticism that has been leveled against the doctrine than because of the infertility of vitalism as a guide in biological research and because of the superior heuristic value of alternative approaches.

And whereas most biologists concur in renouncing this so-called naïve vitalism, some continue to espouse a so-called critical vitalism, perhaps indistinguishable from organicism (see below).

Mechanism

Simply stated, the view of the mechanists is that organisms are no different from subtle machines: the whole is the sum of its parts, which are arranged in such a way that an internal energy source can move them in accordance with a built-in program of purposeful action. In the mechanist's view, advances in molecular biology corroborate this claim and demonstrate that in principle organisms are no more than complicated physical systems. This is, in essence, the reductionist position, which states that biological principles can be reduced to physical and chemical laws. Antireductionists, of course, contend that molecular biology cannot explain all aspects of living forms.

It has often been said that, whereas biologists may think as vitalists—and hold the conviction that organisms are more than just complex machines—they perforce become practicing mechanists in the laboratory, required by the demands of scientific inquiry to view their experiments in terms of the measurable parameters of physics and chemistry. K.F. Schaffner, an American philosopher, suggested in 1967 that, even though reductionism may be correct, a better strategy may be to strive toward an independent biology.

Organicism

The basic claim of organicism is that organisms must be interpreted as functioning wholes and cannot be understood by means of physics and chemistry alone. Few scientists today call themselves organismic biologists or endorse the doctrines put forward by such organismic theorists as Ludwig von Bertalanffy and Edward Stuart Russell. Nevertheless, most antireductionists subscribe at least to part of the organismic doctrine, in particular to its wholistic claim. Russell, a foremost proponent of organicism, stated in his work *The Interpretation of Development and Heredity* (1930):

> Any action of the whole organism would appear then to be susceptible of analysis to an indefinite degree—and this is in general the aim of the physiologist, to analyze, to decompose into their elementary processes the broad activities and functions of the organism. But . . . by such a procedure something is lost, for the action of the whole has a certain unifiedness and completeness which is left out of account in the process of analysis. . . . In our conception of the organism we must . . . take account of the unifiedness and wholeness of its activities . . . since . . . the activities of the organism all have reference to one or other of three great ends, and that both the past and the future enter into their determination. . . . Bio-chemistry studies essentially the *conditions* of action of

cells and organisms, while organismal biology attempts to study the actual modes of action of whole organisms, regarded as conditioned by, but irreducible to, the modes of action of lower unities.

In some special sense, then, an organism is regarded as being more than a simple sum of its parts; an additional "something" has accrued to it as a result of the unique arrangement of its components. As Morton O. Beckner, a philosopher of biology, asserted in an article in *The Encyclopedia of Philosophy* (5:549):

> In the history of biology it is difficult to disentangle vitalistic and organismic strands, since both schools are concerned with the same sorts of problems and speak the same sort of language. The distinction between them was drawn clearly only in the twentieth century. Organismic biology may be described as an attempt to achieve the aims of the murky organismic-vitalistic tradition, without appeal to vital entities.

Further (p. 551):

> Organismic biology is to be interpreted as a series of methodological proposals, based on certain very general features of the organism—namely, the existence in the organism of levels of organization with the biological ends of maintenance and reproduction. These features are sufficient to justify "a free, autonomous biology, with concepts and laws of its own," whether or not the higher levels are ultimately reducible to the lower ones.

Organisms as systems. The concept of an organism as a cybernetic, or automatic-control, system is currently influential in biology.

Holism and emergence

The holistic concept of an organism—*i.e.,* the theory that the determining factors in biology are its irreducible wholes—owes its success primarily to the existence of control and regulation mechanisms operating at the molecular level that determine development and behaviour. The character of such systems at all levels of analysis—molecular through total organism—is nothing other than a sophisticated kind of cybernetics. Holism and reductionism are similar in this respect. Closely allied to organicism is the old problem of emergent properties dealt with earlier: at each successive level of organization, qualities emerge that cannot be anticipated by the components and that confer an added dimension to each hierarchical level in the biological world.

General systems theory

A theoretical and methodological program called general systems theory—presented in its fullest and most persuasive form by Bertalanffy—is an extension of the tenets of organismic biology. It is an attempt to provide a common methodological approach for all of the sciences, based upon the idea that systems of any kind—physical, biological, psychological, and social—operate in accordance with the same fundamental principles. Ideally, it should be possible to deduce the principles applying to a particular sort of system from the more general ones. This approach is one still very much in need of development.

Teleology

Attributions of purpose (teleology) appear frequently in biological writing. Not only do biologists say that parts of organisms have a purpose with respect to the whole, but some hold that life itself is inherently purposive. But the term purpose is both vague and ambiguous. That every biological system—from self-replicating molecules (DNA) to biotic communities—involves specific and identifiable functions is undeniable. But whether, or in what way, functional ends like the reproduction of a cell resemble human intentions or purposes is a matter of some controversy. Even if this matter were settled, a larger question would still remain, viz., whether a biological system as a whole can have a goal that is in some way similar to a human goal—*i.e.,* whether it is programmed with an ultimate purpose. Although resolution of this matter has long been and will continue to be a critical point in the philosophy of biology, much has been done to clarify the issues involved.

Philosophy in evolution. *Teleology and determinism.* Vitalists and those who subscribe to a Lamarckian view of evolution involving the inheritance of acquired characteristics claim that evolution involves a deterministic finalism, or directedness toward an end. Most evolutionists—among them George Gaylord Simpson—reject that claim and hold that natural selection is the non-random element in evolution, that which gives evolution direction. Other evolutionists—among them Theodosius Dobzhansky—argue that the chance factors in mutation and selection, in addition to the unpredictability of environmental change, make it impossible to formulate deterministic laws even in experimental populations, let alone in natural populations. Similar considerations by others have led to the claim that evolutionary biology is a paradigm of an after-the-fact exploratory science and that the course of evolution can never be predicted.

The species problem. Whether biological species can be said to have a real existence in the world is a question that has been receiving much consideration. The issue may be posed in the words of Benjamin Burma, a paleontologist, who, writing in *Evolution* (3:369), asked:

> What, then, is a species? It would seem thus far to be the whole of any one series of breeding populations. . . . [But the] definition as it stands unfortunately puts all living and fossil animals in one species, since there is a continuity of germplasm back from John [an individual animal] to the original primordial cell, and from it forward to every living animal (not to mention plant). Thus, if we ignore time, we end up with only one species. . . .

The temporal difficulty, however, is not the only stumbling block to the question of species reality; for, if the species is redefined as the whole of any one series of breeding populations as it exists at any one time, then there is an infinity of species, since time itself is infinitely divisible. On the basis of these and other objections, some biologists have concluded that species have only a subjective existence merely as convenient labels for arbitrary assemblages and have only a minimum of biological significance. On the other hand, there are proponents of the idea that species have an objective reality. Ernst Mayr, a U.S. evolutionist representing the latter group, has written—also in *Evolution* (3:372):

> In all multidimensional situations an inference has to be made (Simpson, 1943) on the basis of the objective species of the non-dimensional system. The subjectivity of this expanded species concept by no means invalidates the species concept per se. The species of the local naturalist or of the paleontologist within a given horizon is clearly delimited against other species and can thus be considered as having objective reality.

Although the controversy is confused by semantic difficulties, one of the chief contributions of the philosophy of biology has, in fact, been to separate mere linguistic puzzles from matters of substance. Many taxonomists are guilty of ambiguity of reference; they often fail to distinguish their entities clearly, with the result that there is widespread befuddlement over just what stand is held by whom. The problems are now clearer than they have ever been, and with few exceptions biologists and philosophers tend to agree about the nature of biological species and the definition of the species category.

Evolutionary theory. Although most of the issues connected with evolution as a theory are methodological ones, two issues go beyond the limits of logic. Some philosophers have tried to demonstrate, for example, that evolutionary theory is circular and offers no real understanding of the process of evolution. Others have argued that the notions of types of organisms must be used to understand evolution and that evolutionary change takes place when a new type emerges.

The views of Simpson and Schindewolf

Two clear viewpoints regarding evolutionary theory have come to the fore since 1950. One is expressed in detail by George Gaylord Simpson, in his work *The Major Features of Evolution* (1953), and the other is put forward by a paleontologist, Otto Schindewolf, in his *Grundfragen der Paläontologie* (1950). In 1959, Marjorie Grene, a philosopher of biology, writing in the *British Journal of the Philosophy of Science* (9:11 ff.), summarized their positions as follows:

> Professor Simpson is the principal American spokesman of neo-Darwinism. . . . He sees evolution as a continuous series of minute changes in innumerable directions, in which all alterations of any significance, larger as well as smaller, quicker as well as slower, are determined by the great cooperating "pressures" of mutation, geographical isolation, and selection, with adaptation as the universal effect, and criterion, of systematic change. The basic concept, ultimately is variation in

the occurrence of genes; out of such variations all the systematic relations of living things have been gradually evolved.

Schindewolf's principles are simpler. He sees typical shapes, and sees again and again what appear to be new shapes. Therefore he assumes that living things are able to originate novel types. Mutation, he agrees, must have been the mechanism by which they originated; but the adaptive control of mutation occurs only within, not between types. The basic pattern is of change from type to type, and always, as we have seen, with the more general appearing before its specialised subdivisions.

The controversy between these two opposing viewpoints is a complex one filled with both philosophical and scientific issues. In the opinion of most biologists, Schindewolf's view is persuasive only with respect to the paleontological evidence and is not supported by the experimental study of evolution in current organisms. Most of them thus tend to accept the synthetic theory in more or less the form expressed by Simpson. It remains possible, however, that the process that Schindewolf is talking about is fundamentally different from that explored in population genetics and that typostrophic mutations are so rare, on the time scale of man, as to be beyond hope of detection in the laboratory.

Evolution as a world view. Very few attempts have been made in recent years to employ the concept of evolution as a scheme for viewing all knowledge and experience. Sir Julian Huxley, who remains one of the best representatives of such an effort, continues to claim that the entire universe is in a process of evolution, which, however, has different aspects, viz., physical, biological, and social. Life and nonlife alike must be understood as part of the process of cosmic evolution, and from this follows a host of metaphysical and ethical implications. The other chief representative of this viewpoint is the evolutionist priest Teilhard de Chardin, who has woven into the fabric of cosmic evolution the panoply of a Christocentric religion that sees the perfection of all things in an "Omega" point toward which evolution is moving.

Metaphysics of the more piecemeal kind—exploring the implications that biological knowledge has on beliefs and attitudes—is fostered by Simpson, a consistent antagonist of Huxley and Chardin. Simpson suggests, for instance, that knowledge of man's origins and of the process that has brought him to his current state in no way threatens belief in his own uniqueness. Man is an animal, but a very special sort of animal. Other matters of a similar kind—purpose in nature and man's evolutionary future—are considerations that constitute the implications of biology in general and of evolutionary biology in particular.

Biology and ethics. *The question of innate aggressiveness.* One of the best known issues threatening accepted beliefs about moral responsibility is probably that raised by the proponents of the theory of innate aggression, in particular by such spokesmen as Konrad Lorenz, an Austrian student of animal behaviour, and Robert Ardrey, a U.S. writer. If there is an instinct for aggressiveness, then the notion that it is acceptable to blame individuals and society for outbreaks of violence or war loses its validity. The thrust must then be elsewhere: not in faultfinding but in shoring up against what is felt to be pedestrian and inevitable. As Ardrey puts the theory in his *African Genesis* (1961):

Views of the ethologists

> But we were born of risen apes, not fallen angels, and the apes were armed killers besides. And so what shall we wonder at? Our murders and massacres and missiles, and our irreconcilable regiments? Or our treaties whatever they may be worth; our symphonies however seldom they may be played; our peaceful acres, however frequently they may be converted into battlefields; our dreams however rarely they may be accomplished. The miracle of man is not how far he has sunk but how magnificently he has risen. We are known among the stars by our poems, not our corpses.

Raymond Dart, a South African anatomist and anthropologist, in an article entitled "The Predatory Transition from Ape to Man," published in the *International Anthropological and Linguistic Review* (1:201–208), expressed the thesis of innate depravity on which Ardrey's more popular presentation is based.

Another aspect of the innate aggression inherited from man's primate forebears is militant enthusiasm, which Lorenz described in *Das sogenanannte Böse: zur Naturgeschichte der Aggression* (1963; Eng. trans., *On Aggression,* 1966):

> In reality, militant enthusiasm is a specialized form of communal aggression, clearly distinct from and yet functionally related to the more primitive forms of petty individual aggression. Every man of normally strong emotions knows, from his own experience, the subjective phenomena that go hand in hand with the response of militant enthusiasm. A shiver runs down the back and, as more exact observation shows, along the outside of both arms. One soars elated, above all the ties of everyday life, one is ready to abandon all for the call of what, in the moment of this specific emotion, seems to be a sacred duty. All obstacles in its path become unimportant; the instinctive inhibitions against hurting or killing one's fellows lose, unfortunately, much of their power. Rational considerations, criticism, and all reasonable arguments against the behavior dictated by militant enthusiasm are silenced by an amazing reversal of all values, making them appear not only untenable but base and dishonorable. Men may enjoy the feeling of absolute righteousness even while they commit atrocities. Conceptual thought and moral responsibility are at their lowest ebb. As a Ukrainian proverb says: "When the banner is unfurled, all reason is in the trumpet."

Montagu's view

Equally notable opponents of the theory of innate aggression see it much as M.F. Ashley Montagu, a British–U.S. anthropologist, does, as "original sin revisited," and deplore the tendency to neglect authoritative studies in favour of simplistic popularization. In *Man and Aggression* (1968), he writes:

> While the findings of these disciplines [anthropology and the behavioral sciences] are wholly opposed to the deeply entrenched view that man is an innately aggressive creature, most people tend to dismiss these findings out of hand or ridicule them as a rather eccentric idealistic heterodoxy, which do not deserve to become generally known. In preference to examining the scientific findings they choose to cast their lot with such "authorities" as William Golding who, in his novel *Lord of the Flies,* offers a colorful account of the allegedly innate nastiness of human nature, and Robert Ardrey who, in *African Genesis* and more recently in *The Territorial Imperative,* similarly seeks to show that man is an innately aggressive creature. . . .
>
> . . . when through the distorting glass of his prejudgments he looks at a tool it becomes not simply a scraper but a weapon, a knife becomes a dagger, and even a large canine tooth becomes "the natural dagger that is the hallmark of all hunting animals," while in "the armed hunting primate" it becomes "a redundant instrument." "With the advent of the lethal weapon natural selection turned from the armament of the jaw to the armament of the hand." But the teeth are no more an armament than is the hand, and it is entirely to beg the question to call them so. Virtually all the members of the order of primates, other than man, have large canine teeth, and these animals, with the exception of the baboons, are predominantly vegetarians, . . . that such teeth may, on occasion, serve a protective purpose is entirely secondary to their main function, which is to rip and shred the hard outer coverings of plant foods.

Further responses to Ardrey's and Lorenz' thesis are the interpretations of field studies of primate groups, such as those on the gorilla, chimpanzee, and orangutan. These researches suggest that the majority of such groups are singularly free of belligerence. According to Montagu,

> The myth of the ferocity of "wild animals" constitutes one of Western man's supreme rationalizations, for it not only has served to "explain" to him the origins of his own aggressiveness, but also to relieve him of the responsibility for it—for since it is "innate," derived from his early apelike ancestors, he can hardly, so he rationalizes, be blamed for it! And some have gone so far as to add that nothing can be done about it, and that therefore wars and juvenile delinquents, as Mr. Ardrey among others tells us, will always be with us! From one not-so-minor error to another Mr. Ardrey sweeps on to the grand fallacy.

The matter remains moot; but there appears to be a growing consensus that, given a certain genetic constitution—and within the bounds of that endowment—whatever man is, he learns to be, especially in respect to values, morality, and customs. Baser appetitive needs, however, may have a genetic component that is greater than an environmental one.

New understanding of environmental factors and the consequences of man's actions with respect to them has made

it clear that man has acquired responsibilities that he did not recognize before. It has become increasingly accepted that standards and values with respect to the environment must be established; this is perhaps the most dramatic case in which recent biological knowledge has generated a crisis of a moral kind. The classic work *Science and Survival* (1966) by a biologist, Barry Commoner, is particularly noteworthy in connecting theoretical and philosophical issues about reductionism and holism to practical matters of environmental understanding and problem solving.

The metaphysical issue of man's place in nature is now being construed as one that requires that man make value decisions, assign responsibilities, and plan for the future of his planet. Environmental problems have become intertwined with problems of social planning, racial tension, transportation and housing crises, genetic engineering, and a host of other current concerns.

Evolutionary ethics. The question of whether nature provides guides to the actions of humankind has held a fascination for many biologists. Those who call themselves evolutionary ethicists say that it does. The defenders of evolutionary ethics contend that external moral standards exist in the facts and process of evolution.

Social Darwinism

Toward the end of the 19th century, Herbert Spencer, in England, and others advanced a series of principles that came to be called Social Darwinism. It espoused such ideas as the inevitability of progress, survival of the fittest, and the struggle for existence, expressions that have become bywords although they have since been discredited in their original sense, as applied to social phenomena. Social Darwinism, as C.H. Waddington, a biologist, explains in his book *The Ethical Animal* (1960), has been superseded by

> . . . the more recent phase of evolutionary ethical thought beginning in the early 1940s, [which] comprises a number of rather different methods of approach. At one extreme we have discussions framed in terms of extremely wide scope, which treat of evolution not only in the animal world but throughout the cosmos, and attempt to relate such broad concepts to man's religious and spiritual life. The pre-eminent example of this tendency in recent years is Teilhard de Chardin, but a rather similar approach can be found in the works of several biologists, such as Conklin, Holmes, and Huxley. The opposite tendency, which of course is also found expressed to various extents in these authors, particularly in Julian Huxley, is the attempt to demonstrate, in a logically coherent argument, a real connection between evolutionary processes and man's ethical feelings.

Evolution and morality

Some biologists continue to insist, therefore, that biological facts can provide a yardstick by which to measure the morality of a given course of action. Julian Huxley, for one, has long claimed that moral principles can be found in nature and in the evolutionary process in particular:

> When we look at evolution as a whole, we find, among the many directions which it has taken, one which is characterized by introducing the evolving world-stuff to progressively higher levels of organization and so to new possibilities of being, action, and experience. This direction has culminated in the attainment of a state where the world-stuff (now moulded into human shape) finds that it experiences some of the new possibilities as having value in or for themselves; and further that among these it assigns higher and lower degrees of value, the higher values being those which are more intrinsically or more permanently satisfying, or involve a greater degree of perfection.

Huxley further asserts that, although the Golden Rule, the policy of action based on sympathy—doing as one would be done to by others—may be an immediate good, it ultimately leads to the suppression of those qualities most needed for survival and the continuation of a species. Rather, he argues:

> The facts of nature, as demonstrated in evolution, give us assurance that knowledge, love, beauty, selfless morality, and firm purpose are ethically good. . . . In the broadest possible terms evolutionary ethics must be based on a combination of a few main principles: that it is right to realize ever new possibilities in evolution, notably those which are valued for their own sake; that it is right both to respect human individuality and to encourage its fullest development; that it is right to construct a mechanism for further social evolution which shall satisfy these prior conditions as fully, efficiently, and as rapidly as possible.

Simpson, however, contends, in the article "Biological Sciences," in *The Great Ideas Today* (1965):

> The facts and the processes of evolution are neither ethical nor unethical. The questions of good or bad are simply irrelevant to this field, with the important reservation that evolution has produced a species, *Homo sapiens,* concerned with ethics. Denial of man's naturalistic origin and animal nature is flatly false, and any ethic based on such denial is invalid. Evolution controverts primitive creation myths, but it is consistent with higher values in the Judeo-Christian tradition and those in most now-current religions and philosophical systems. One need only think of the brotherhood of mankind—a biological fact, not only an ethical idea.
>
> Beyond such considerations as those, efforts to combine science and religion may be noble in intention but usually end up distorting or stultifying both. One of the most striking examples at present is the cult, as it may fairly be called, of Pierre Teilhard de Chardin. He preaches—necessarily posthumously, for the Roman Catholic Church suppressed his views during his life—a mystical Christianity ostensibly derived from evolutionary principles. But since the mysticism is primary, the evolutionary principles are distorted and downright falsified for seeming coherence with the nonscientific, nonnaturalistic premises. In turn, the mystical views advanced as having that false basis are thereby vitiated. The result (in my opinion) has been a disservice to true relgion and to true science.
>
> At the same time, no one can deny the purity of Father Teilhard's intentions or the correctness of his view that evolution and religious feeling should be considered congruent aspects of the nature of man. It is almost as irrational to deny evolution as to deny gravity. The management of life and the goals of aspiration, to be sane, must take account of all such truths of nature. They need not thereby become brutal or earthbound.
>
> (J.R.Mn./R.C.Y.)

Philosophy of history

The term history may be employed in two quite different senses: it may mean (1) the events and actions that together make up the human past, or (2) the accounts given of that past and the modes of investigation whereby they are arrived at or constructed. When used in the first sense, the word refers to what as a matter of fact happened, while when used in the second sense it refers to the study and description of those happenings. The notion of philosophical reflection upon history and its nature is consequently open to more than one interpretation, and contemporary writers have found it convenient to regard it as covering two main types of undertaking. On the one hand, they have distinguished philosophy of history in the traditional or classical sense; this is conceived to be a first-order enquiry, its subject matter being the historical process as a whole and its aim being, broadly speaking, one of providing an overall elucidation or explanation of the course and direction taken by that process. On the other hand, they have distinguished philosophy of history considered as a second-order enquiry; here attention is not focussed upon the actual sequence of events themselves but, instead, upon the procedures and categories used by practicing historians in approaching and comprehending their material. The former, often alluded to as speculative philosophy of history, has had a long and varied career; the latter, which is generally known as critical or analytical philosophy of history, has only risen to prominence during the 20th century.

SPECULATIVE THEORIES

The necessity for a meaningful view of history

The idea of an order or design in history. The belief that it is possible to discern in the course of human history some general scheme or design, some all-encompassing purpose or pattern, is very old and has found expression in various forms at different times and places. The reasons for its persistence and vitality are numerous, but two very general considerations may be identified as having exercised a fairly continuous influence. First, it has often been supposed that, if the belief in an overall pattern is abandoned, one is obliged to acquiesce in the view that the historical process consists of no more than an arbitrary succession of occurrences, a mere agglomeration or patchwork of random incidents and episodes. But such a view (it has been contended) cannot be seriously entertained, if only because it conflicts with the basic demand for system

and order that underlies and governs all rational enquiry, all meaningful thought about the world. Second, it has frequently been felt that to refuse to allow that history is finally intelligible in the required manner implies a skepticism concerning the value of human life and existence that constitutes an affront to the dignity of human nature. The 18th-century German philosopher Immanuel Kant, for example, spoke of the "repugnance" that is inevitably experienced if the past is viewed

> as if the whole web of human history were woven out of folly and childish vanity and the frenzy of destruction, so that one hardly knows in the end what idea to form of our race, for all that it is so proud of its prerogatives.

In more recent times, a comparable attitude is discernible beneath Arnold Toynbee's uncompromising repudiation of the idea that history is "a chaotic, disorderly, fortuitous flux, in which there is no pattern or rhythm of any kind to be discerned." Thus, it has been the object of a long line of theorists, representative of widely divergent outlooks, to demonstrate that such pessimism is unjustified and that the historical process can, when appropriately viewed, be seen to be both rationally and morally acceptable.

Theological origins. Western speculation concerning the meaning of history derived in the first instance chiefly from theological sources. The belief that history conforms to a linear development in which the influence of providential wisdom can be discerned, rather than to a recurrent cyclical movement of the kind implicit in much Greco-Roman thought, was already becoming prevalent early in the Christian Era. Traces of this approach are to be found in the conception of the past set forth by St. Augustine in his *City of God* and elsewhere; it is, for example, compared on one occasion to "the great melody of some ineffable composer," its parts being "the dispensations suitable to each different period." Yet the cautious subtlety of Augustine's suggestions and the crucial distinction he drew between sacred and secular history make it important not to confuse his carefully qualified doctrines with the cruder positions advanced by some of his self-proclaimed successors. This applies, *par excellence*, to the work of the most renowned and thorough of these, Jacques-Bénigne Bossuet. Written 1,250 years after Augustine's death, Bossuet's *Discours sur l'histoire universelle* (1681; "Discourse on Universal History") is imbued throughout with a naïve confidence that the entire course of history owes its pervasive character to the contrivance of a "higher wisdom." In the eyes of Bossuet, to grasp and understand the great procession of empires and religions was "to comprehend in one's mind all that is great in human affairs and have the key to the history of the universe." For the rise and fall of states and creeds depended in the end upon the secret orders of Providence, the latter being the source of that manifest historical justice and retribution to which, on nearly every page, the annals of the past bore clear and unmistakable witness. Bossuet's vast survey was, in fact, the last major contribution to its genre. Though it made a considerable impression when it was first published, it appeared just before the discoveries of Sir Isaac Newton effected a massive transformation of the European outlook, and the book's impact was short-lived. Thus, the development of historical speculation in the 18th century was generally marked by a tendency to reject theological and providential interpretations in favour of an approach more closely aligned, in method and aim, to that adopted by natural scientists in their investigations of the physical world.

Augustine and Bossuet

Secular approaches: the Enlightenment and beyond. For many Enlightenment and post-Enlightenment thinkers, the project of establishing a science of history and society, comprising hypotheses and laws of an explanatory power analogous to that attained by theories in the physical sciences, acquired an almost obsessive importance. The age of religious and metaphysical conjectures concerning the destiny of human affairs had, in their opinion, come to a close. The task that now presented itself was one of constructing, upon the basis of hard observable facts, interpretations that would not only rescue the human studies from ignorance, uncertainty, and primitive superstition but also put into men's hands an instrument for predicting and controlling their fate. Thus, the idea of creating a universally valid social science, capable of accounting for the phenomena of history in terms of causal principles comparable to those employed in the natural sphere, came to be linked with the promotion of reformist and revolutionary ideals. Men such as Condillac and Condorcet in the 18th century and Henri de Saint-Simon, Auguste Comte, John Stuart Mill, and Henry Thomas Buckle in the 19th century all believed that it was feasible to apply scientific procedures to the study of human development. But equally—though in widely different ways—they were men deeply concerned with practical objects and committed to changing existing institutions and ways of life. To these men, theory was complementary to practice; knowledge was power.

19th-century skepticism

Yet even in the 19th century, when speculation of this type was at its height, there were informed skeptics—Joseph de Maistre and Arthur Schopenhauer, for example, and later the great Swiss historian Jacob Burckhardt—who challenged the optimistic and rationalistic presuppositions on which it was founded. It was pointed out that notions such as that of the perfectibility of man or of the existence of some foreseeable goal toward which the course of events was inexorably leading were not empirically established truths but mere articles of faith; in subscribing to them, historical theorists often appeared to be tacitly importing into their allegedly scientific interpretations teleological conceptions of a kind that it had been their declared intention to banish forever from social enquiry. These objections have been repeated and amplified by 20th-century critics such as Karl Popper, who have also maintained that the theorists in question were, in any case, working with an unacceptably crude notion of scientific reasoning and that their high-sounding generalities conspicuously failed to measure up to the requisite standards of conceptual precision and observational testability.

Although such strictures have considerable force, they should not obscure the significant contribution that had been made toward extending human knowledge and understanding. The tendency, for example, to insist upon the relevance of scientific modes of procedure to the areas of historical and social investigation at least achieved the salutary effect of throwing into relief the inadequacy of previous work in these domains; moreover, it indirectly brought to the fore the entire question of the status of history as a legitimate form of thought. For, if history should prove resistant to attempts to assimilate it to other accredited branches of enquiry, it would be necessary to show why this was so and to exhibit those features of historical thinking that lent it its distinctive and irreducible character.

The new science: Vico and Herder. Among the 18th-century theorists, two writers can indeed be picked out who—while remaining firmly within the speculative tradition—at the same time possessed sufficient genius and prescience to realize that the solution to the problem of establishing history as a reputable discipline might be found by pursuing a course different from one modelled upon the methodology of the natural sciences. Partly because of the obscure and scholastic manner in which it was written, Giambattista Vico's *Scienza nuova* (3rd ed. 1744; "New Science") was a work whose importance remained for a long time wholly unrecognized, and it is only fairly recently that its significance and originality have been fully appreciated.

Vico's notion of unique historical knowledge

Central to the book is the contention that the kind of knowledge that men can achieve of their own actions, creations, and institutions is of a radically different type from the knowledge that is acquired by the observation and investigation of the nonhuman or natural world; knowledge of the former variety is, moreover, held to be in principle superior to that of the latter. For, in Vico's opinion, in order truly to know something it is necessary in some sense to have made it: it followed that, whereas the reality studied by the physical scientist is the creation of God and therefore only properly known by God, the "world of nations" that forms the subject matter of history is the creation of men and is therefore something that men can "hope to know." Thus, Vico was led to

stress the differences rather than the analogies between historical and other forms of enquiry; in particular, he emphasized the need for the historian to enter imaginatively into the spirit of past ages, re-creating the outlooks and attitudes that informed them as opposed to seeking to impose upon them inappropriate or falsifying interpretations—"pseudomyths"—that derived from the cultural ethos of his own time. Vico propounded a cyclical theory of human history, according to which "nations" or societies pass through determinate stages, and he combined this with the idea that a providential principle is in some manner immanent within the various forms of life that men construct. He employed such conceptions, however, in a fashion that underlined man's nature as a historical being, whose powers and capacities do not conform to a fixed or static pattern but are necessarily subject to change and development in the course of time.

In a similar vein, the German writer Johann Gottfried von Herder, in his influential *Ideen zur Philosophie der Geschichte der Menschheit* (4 vol., 1784–91; Eng. trans., *Outlines of a Philosophy of the History of Man*, 1800), implied that it was vital to view human actions and achievements from a standpoint that took proper account of "time, place and national character"—in other words, cultural milieu and the inevitable limits imposed by historical situation and circumstance. In its general direction, Herder's historical thought reflected the Enlightenment preconceptions of man as a progressive being. Herder's chief importance lies, however, in his insistence upon the misconceptions involved in treating the products of past thought and action as if they were the manifestations of an unchanging human consciousness and as if they could be explained by reference to abstract laws eternally valid for men everywhere. According to Herder, such an approach failed to recognize the complex influences that act upon human beings as members of particular historical societies; each of these societies possessed its unique life-style, subtly but inescapably determining the mentalities of those born within its confines in a manner that rendered futile all attempts to reduce human propensities and needs to the terms of some simple set of abstract formulas.

Herder's rejection of static laws

Many of Vico's and Herder's ideas appear familiar today, but it is easy to forget that the emergence of what has come to be known as the "historical sense" is a comparatively recent phenomenon, one that represents a genuine revolution in European thought. It is largely because of this revolution that social and political theories of the kind elaborated by men such as Thomas Hobbes and Benedict de Spinoza in the 17th century seem oddly artificial to 20th-century eyes, so remote are the categories in which they sought to explain human life and behaviour from those that have subsequently found acceptance.

History as a process of dialectical change: Hegel and Marx. The suggestion that there is something essentially mistaken in the endeavour to comprehend the course of history "naturalistically" and within an explanatory framework deriving from scientific paradigms was powerfully reinforced by conceptions stemming from the development of German Idealism in the 19th century. Hegel's "philosophy of the spirit" made its appearance upon the intellectual scene contemporaneously with Saint-Simonian and Comtean Positivism, rivalling the latter in scope and influence and bringing with it its own highly distinctive theory of historical evolution and change. Hegel's stress upon the "organic" nature of social wholes and the incommensurability of different historical epochs owed evident debts to Herderian ideas, but he set these within an overall view that pictured the movement of history in dynamic terms. Regularities and recurrences of the sort that typically manifest themselves in the realm of nature are foreign, Hegel maintained, to the sphere of mind or spirit, which was characterized instead as involving a continual drive toward self-transcendence and the removal of limitations upon thought and action. Man was not to be conceived according to the mechanistic models of 18th-century Materialism; essentially he was free, but the freedom that constituted his nature could only achieve fulfilment through a process of struggle and of overcoming obstacles that were themselves the expression of his own activity; it was in this sense that Hegel claimed that spirit was "at war with itself"—"it has to overcome itself as its most formidable obstacle" (*Lectures on the Philosophy of History*). In concrete terms, this meant that historical advance did not proceed through a series of smooth transitions. Once the potentialities of a particular society had been realized in the creation of a certain mode of life, its historical role was over; its members became aware of its inadequacies, and the laws and institutions they had previously accepted unquestioningly were now experienced as fetters, inhibiting further development and no longer reflecting their deepest aspirations. Thus, each phase of the historical process could be said to contain the seeds of its own destruction and to "negate" itself; the consequence was the emergence of a fresh society, representing another stage in a progression whose final outcome was the formation of a rationally ordered community with which each citizen could consciously identify himself and in which there would therefore no longer exist any sense of alienation or constraint. Somewhat curiously, the type of community Hegel envisaged as exemplifying this satisfactory state of affairs bore a striking resemblance to the Prussian monarchy of his own time.

History as a dynamic process

The notion that history conforms to a "dialectical" pattern, according to which contradictions generated at one level are overcome or transcended at the next, was incorporated—though in a radically new form—in the theory of social change propounded by Karl Marx. Like Hegel, Marx adopted a "directional" view of history; but, whereas Hegel had tended to exhibit it as representing the unfolding in time of an inner spiritual principle, Marx looked elsewhere for the ultimate determinants of its course and character. Man, according to Marx, was a creative being, situated in a material world that stood before him as an objective reality and provided the field for his activities; this primitive truth, which had been obscured by Hegel's mystifying abstractions, afforded the key to a proper understanding of history as a process finally governed by the changing methods whereby men sought to derive from the natural environment the means of their subsistence and the satisfaction of their evolving wants and needs. The productive relations in which men stood to one another, resulting in such phenomena as the division of labour and the appearance of economically determined classes, were the factors fundamental to historical movement. What he termed the superstructure of society—which covered such things as political institutions and systems of law, ethics, and religion—was in the last analysis dependent upon the shape taken by the "material production" and the "material intercourse" of human beings in their struggle to master nature: "it is not the consciousness of men that determines their being, but, on the contrary, their social being that determines their consciousness." Hence, the inner dynamic of history was held to lie in conflicts arising from changes in the means of production and occurring when modes of social organization and control, adapted to the development of the productive forces at one stage, became impediments to it at another; they were to be resolved, furthermore, not by abstract thought but by concrete action. Thus, the Hegelian conception of spirit as involved in a relentless struggle with itself and with what it had created underwent a revolutionary transformation, explosive in its implications.

Economic determinism

Marx's interpretation of the historical process, with its stress upon necessity and the operation of ineluctable laws, has often been portrayed by its proponents as being scientific in character. It has, however, more than one aspect, and it would be an error to identify its underlying methodology with that associated with Comtean Positivism. Generally speaking, the basic categories within which it was framed derived from a theory of human nature that had more in common with the postulates of German romantic thought than with those of British and French Empiricism: to this extent, the logical structure Marx sought to impose upon the data of history belonged to a tradition that stressed the differences rather than the resemblances between the human and the natural world.

Twentieth-century systems. The tendency to detect in history the presence of large-scale patterns and compre-

hensive uniformities continued into the 20th century in the work of a number of writers, most notably Oswald Spengler and Arnold Toynbee. Spengler's *Decline of the West* (originally published in German, 1918–22), wherein the history of mankind was presented in terms of biologically conceived cultures whose careers conformed to a predetermined course of growth and decay, was widely acclaimed during the years of disillusionment that followed World War I; and a somewhat similar reception was given to Toynbee's massive *A Study of History* (1934–61) immediately after World War II. Toynbee, like Spengler, undertook a comparative study of civilizations, thereby repudiating attempts to treat the past as if it exhibited a single linear progression: at the same time, he diverged from Spengler in suggesting that current Western society might not after all be necessarily doomed to extinction and in tempering a predominantly deterministic mode of thought with reservations that allowed a place for human free will and the possibility of divine intervention. Yet, as some of his critics were quick to point out, such qualifications were not easy to reconcile with his original insistence upon the need to adopt "a scientific approach to human affairs"; nor was it clear that his own use of inductive methods to establish the laws governing the development of civilizations was above logical suspicion or reproach. Toynbee's experiment might have been impressive as an individual achievement; nevertheless, with the multiplication of objections and in a theoretical climate that had become skeptical of speculative system-building of any kind, the very feasibility of engaging upon a project of the type he had undertaken came to be seriously questioned. It was felt increasingly that philosophy of history in the traditional sense—resting largely upon uncriticized assumptions concerning the nature of historical enquiry and its relations with other disciplines—had reached something of an impasse; if history was still to be treated as a proper subject for philosophical examination, it must be along lines quite different from those previously pursued.

Toynbee's system

ANALYTICAL PROBLEMS

The concept of history. The task of trying to delineate the specific character of historical knowledge and understanding, rather than of seeking to construct vast speculative schemes in the earlier manner, first began to attract the attention of philosophers toward the end of the 19th century. To such thinkers as Wilhelm Dilthey and Benedetto Croce, the claim that, in the absence of some all-embracing system of a teleological or quasi-scientific kind, the course of history could be regarded as constituting nothing better than a meaningless chaos appeared to be totally unacceptable. History is intelligible, they believed, in the sense that historians make it so; moreover, this was the only type of intelligibility it was either necessary or legitimate to demand. What could reasonably be looked for was a clearer and deeper insight into the conditions that render historical knowledge possible, an elucidation of the presuppositions upon which historical enquiry is founded and of the principles according to which it proceeds. It was with such an investigation in mind that R.G. Collingwood, a British philosopher who owed much to Crocean ideas, wrote in his *Autobiography* that "the chief business of twentieth-century philosophy is to reckon with twentieth-century history." By contending that the philosopher should eschew the grandiose ambition of providing a synoptic vision of the entire historical process and concern himself, instead, with the articulation and justification of existing historical procedures, Collingwood and his continental precursors made, in effect, a crucial contribution toward setting philosophy of history on a new path. Their proposals were, moreover, given additional impetus by the widespread acceptance of analytical approaches in other branches of philosophy. In consequence, contemporary thinkers have tended to focus attention upon the explication of concepts and terms that perform a key role in historical thought and description as these are actually carried on: among other things, they have been led into discussing the ways in which historians typically divide up and classify the past, the manner in which they argue for and substantiate their interpretations, and the logical structure of the explanations they are accustomed to offer.

The work of R.G. Collingwood

Explanation and understanding. Both Croce and Collingwood, in their criticisms of earlier theorists, were especially anxious to expose what they believed to be recurrent and fundamental misconceptions regarding the method and subject matter of history: central to these was the assumption that historical occurrences could be subsumed under, and explained in terms of, universal laws of the sort that played an essential part in scientific interpretations of inanimate nature. This assumption was, in their opinion, a gross error. As Collingwood put it, the moment had arrived for history to be released from "its state of pupilage to natural science." With this in mind, he went on to develop an account of historical understanding according to which the historian explains events by exhibiting them as the expressions of past thinking on the part of self-conscious purposive agents—thinking that the historian must imaginatively reconstruct or re-enact in his own mind—rather than by showing the events to be instances of general uniformities or regularities that are established by induction. In propounding this view—which Croce, though he formulated it less clearly and precisely, basically shared—Collingwood set in motion a controversy concerning knowledge and explanation in history that has been central to much subsequent discussion. As Collingwood himself was fully aware, a position similar to his own had been originally advanced (though in a very different context) by Vico; and it is indeed noteworthy that the general division, evident at the speculative level, between those who wished to comprehend historical phenomena in ways suggested by the physical sciences and those who, by contrast, argued for an altogether distinct pattern of interpretation has tended, in recent times, to re-emerge at the level of methodological and conceptual analysis.

The unity of science approach

Thus, on one side of the dispute, there have ranged philosophers who have taken their stand upon what has been called "the unity of science" and who have insisted that the categories and procedures appropriate to the human studies do not enjoy a unique or privileged status that somehow sets them apart from those characteristic of systematic empirical enquiry in other domains. In a classical 18th-century discussion, David Hume had argued that, if two events were said to be causally related, this could only be in the sense that they instantiated certain regularities of succession that had been repeatedly observed to hold between such events in the past: to presume otherwise was to fall back upon an unacceptable belief in "intuitable" connections that had no warrant either in reason or experience. This doctrine may be said to have been given more rigorous expression among Positivist philosophers of the present century in the shape of what is variously known as the "deductive-nomological" or "covering law" theory of explanation; as originally applied to history by Carl Hempel, it amounted to the claim that explaining a given historical occurrence in terms of some other event or set of events necessarily involves an appeal, which need not be more than tacit, to laws or general propositions correlating events of the type to be explained with those of the kind cited as its causes or conditions. Although the proposed analysis has received a variety of different formulations, each designed to meet specific objections that have been raised against it, its adherents have not wavered from the conviction that some such account must be in principle correct if explanations in history are to be open to rational assessment of the sort properly demanded within any legitimate branch of empirical investigation. It is for this reason, together with others, that they have been strongly opposed to *Verstehen*, or "empathy," theories of historical knowledge, regarding the contention that historical understanding presupposes an allegedly direct identification with the mental processes of past human agents as representing at best a heuristic recommendation of doubtful utility, at worst an obscurantist doctrine that transparently fails to provide an objective criterion whereby divergent historical interpretations can be evaluated.

Resistance to the Positivist approach has come from more than one direction. To a number of practicing historians, for instance, the account offered has appeared implausible

inasmuch as it overlooks the "irreducible particularity" of historical occurrences and because it postulates an unjustifiably high degree of reliance upon the presence of discernible uniformities in the sphere of human affairs. So far as philosophers are concerned, dissatisfaction has been voiced both by those to whom the Croce-Collingwood notion of historical thinking as the "re-enactment of past experience" has seemed to contain an important element of truth and also by those followers of Ludwig Wittgenstein who have been impressed by the skepticism concerning the adequacy of scientific models apparent in his later discussions of mental concepts. A leading representative of the former group, W.H. Dray, not only constructed a series of arguments to demonstrate the deficiencies of the covering-law theory but further proposed an alternative conception of "rational explanation," which—it was suggested—fitted many of the familiar ways whereby historians seek to render the past intelligible. Thus, Dray has maintained that the function of much historical explanation consists of showing the actions of historical persons to have been "appropriate" when viewed within the perspective of their specific beliefs, aims, and principles: it was this consideration, he claims, that was uppermost in the minds of theorists who were concerned to stress the part played by imaginative or empathetic understanding in historical reconstruction, their point being primarily a "logical" one and not necessarily carrying any of the dubious epistemological implications attacked by Positivist critics. From a different standpoint, Anglo-U.S. writers influenced by Wittgenstein have challenged the entire assumption that explanations involving the notions of human intention and purpose are susceptible to a Humean pattern of causal analysis; they have also (for example, in the work of Peter Winch) stressed the extent to which historical descriptions of past behaviour require to be framed in terms the agents themselves would have recognized as giving meaning to their activities, terms embodying references to ideas and conventions that defined the social reality in which they participated.

The criterion of appropriateness

Objectivity and evaluation. Fundamental issues concerning the status of historical enquiry of the kind just mentioned have arisen in another crucial area of discussion, centring upon the question of whether—and, if so, in what sense—history can be said to be an objective discipline. Some modern philosophers have inclined to the view that the entirely general problem of whether history is objective cannot sensibly be raised; legitimate questions regarding objectivity are only in place where some particular piece of historical work is under consideration, and in that case there are accepted standards available, involving such matters as documentation and accuracy, by which they can be settled. To others, however, things have not seemed so clear, and they have drawn attention to the doubts that may be felt when history is compared with different branches of investigation, such as chemistry or biology: by contrast with such enquiries, the historian's procedure, including the manner in which he conceptualizes his data and the principles of argument he employs, may appear to be governed by subjective or culturally determined predilections that are essentially contestable and, therefore, out of place in a supposedly reputable form of knowledge. One topic that has been recurrently examined in this connection has been the role of evaluation (specifically, of moral evaluation) in historical writing—a subject, incidentally, about which historians themselves are apt to exhibit a certain uneasiness. Nevertheless, recommendations to the effect that value judgment can and should be totally excluded from history and, indeed, from the social studies as a whole have met with a mixed philosophical reception. Among Positivists and Logical Empiricists, traditionally skeptical of the rationality of value judgments and anxious in any case to reduce the differences between the human and the natural sciences, they have found some measure of support. But that has been by no means a general response. Thus, objectors have pointed out that the language the historian customarily uses, adapted as it is to the assessment and appraisal of human motives and characteristics, makes some degree of evaluation unavoidable; they argue that, even if the possibility of a drastically revised historical vocabulary allows the ideal of a *wertfrei*, or objective history, to be theoretically conceivable, such an ideal can scarcely be seriously entertained as a realizable practical goal. These considerations have been reinforced by the further point that every historian, insofar as he has to select from the mass of material confronting him, is necessarily committed to forming judgments ascribing relative importance and significance; such attributions cannot, however, be simply read off from the facts and must, rather, be said to depend upon the prior acceptance of certain critical standards. To this extent, then, one is required to acknowledge the presence in historical writing of an ineliminable evaluative component, which is liable to obtrude itself into even so "objective" a field as that of causal analysis: it is notorious that disputes between historians as to the "true" causes of occurrences such as wars or revolutions often appear to resist resolution at a purely empirical level, and it has been persuasively maintained by some philosophers that the basic grounds for such disputes may often be traced back to one historian's adherence to a moral or political standpoint not shared by his opponent.

The question of value judgments

Conclusions. Although the topics discussed above have occupied a central position in 20th-century critical discussion, they represent only a sample of the issues with which analytical philosophers of history have been concerned: other problems that have attracted attention have related to the freedom and responsibility of historical agents, the nature and description of historical events, and the role of narrative in history. Here, as elsewhere, the approach adopted has often produced results of considerable interest, throwing a revealing light on features of historical enquiry that are easily missed or ignored by theorists in the grip of some powerful dogma or ideology. Even so, it has perhaps been accompanied by a too ready acquiescence in the view that history is "in order as it is," the philosopher's function being confined to offering a purely descriptive elucidation of typical modes of historical thought and argument. In accepting this conception of their role, analytical philosophers of history have no doubt been partly, and understandably, influenced by a desire to avoid emulating the heady ambitions of their speculative predecessors. Yet, normative questions regarding the validity or adequacy of established procedures within any domain can always be legitimately raised; in the case of history, there seems to be no compelling reason to assume that such problems necessarily lie beyond the scope of philosophical criticism and appraisal. (P.L.G.)

Philosophy of religion

In addition to treating what is commonly called the philosophy of religion, this section considers a wide spectrum of situations, experiences, and issues recognized as "religious" and endeavours to appraise the characteristic approaches and attitudes not only of the adherents of particular religions but also of those who stand outside any particular religion, whether as sympathizers or caustic critics. Outside the scope of this section, however, are questions relating to the study of religions and its methodology or questions relating to the types of argument by which one interpretation of a religious claim is preferred to another (see also RELIGIONS, THE STUDY AND CLASSIFICATION OF).

RELIGION AS A FACT IN HUMAN EXPERIENCE, CULTURE, AND HISTORY

Evidences of religious attitudes and loyalties exist in every sector of human life—in human experience in general; in "culture," the complex interweaving of attitudes, concerns, and views; and in history, the record of social and personal behaviour.

The findings of psychology. Religion incorporates certain characteristic feelings and emotions such as wonder, awe, and reverence. The religious person tends to show a concern for values, moral and aesthetic, and to seek appropriate action to embody these values. He is likely to characterize behaviour not only as good or evil but also as holy or unholy and people as not only virtuous or unvirtuous but also as godly or ungodly.

Characteristic feelings and emotions of religion

As a feature of human existence, religious life can be studied, for example, in terms of psychology, sociology, and history. Among the first books in the psychology of religion were two by Jonathan Edwards, an 18th-century American theologian: *A Faithful Narrative of the Surprising Work of God* (1737) and *A Treatise Concerning Religious Affections* (1746). About a century later, during a period of religious "revivals," interest developed concerning the age at which conversions most often took place—the period of adolescence. Reflections on such facts, and in this sense the psychology of religion, only came, however, with the works of two American psychologists: Edwin Diller Starbuck's *Psychology of Religion* (1889) and the classical treatment by William James's *Varieties of Religious Experience* (1902). Generally, the psychology of religion has shown that though religion for some is a crisis experience, for others it is a natural growth.

As psychology became more analytical it became more interested in the abnormal, in neuroses and dreams, in the techniques of hypnosis, and in the kinds of experience induced by drugs. When Freud spoke of religion as an illusion, he maintained that it is a fantasy structure from which a man must be set free if he is to grow to maturity; and in his treatment of the unconscious he moved toward atheism. The study of the unconscious by the Swiss psychiatrist Jung, however, suggested that dominant archetypes (implying innate tendencies to form symbolic images) are supplied by a racial unconscious, thus providing a psychological approach to belief in God.

Classifications of individuals

In classifying individuals into different types, psychology has distinguished between religious people who are: "extrovert" or "introvert" (Jung), "healthy minded" or "sick" (William James), and "objective" or "subjective" worshippers (J.B. Pratt). There is always the danger, however, that psychological distinctions may beg too many philosophical questions.

One of the most widely accepted studies of religious experience in regard to feelings was written by the modern German Protestant theologian Rudolf Otto. In his *Idea of the Holy,* Otto analyzed what is distinctively religious in terms of the unique concept of the "numinous"; *i.e.,* something both awesome and appealing, both fearful and attractive.

Psychology, however, is concerned not only with individuals but also with what is known about group behaviour, which can also be of importance in any study of the Christian Church or other religious institutions regarded as communities of religious people. The authority of a religious leader, like that of all leaders, is derived from his symbolic character and the extent to which the leader and his followers share a common ideal.

The findings of sociology. The ideas and images of a religion are much influenced by the social culture in which it emerges. Some of the oldest social institutions and practices, such as those concerning birth and death, marriage and the family, and art and music, have developed in a religious context. Religion has often been a driving force in the reform of social abuses, but also it has been associated with reaction and oppression. More recently, the sociology of religion—influenced by contemporary sociology—has been concerned with making use of sociological criteria and of demographical and statistical studies in planning the church's mission and appraising its significance.

The findings of the history of religions. Conclusions in the history of religions have been largely determined by the particular ideas of man or history with which the study was approached. Some scholars have supposed that at the dawn of human existence there was a belief in a single god and that only later there occurred a development into a belief in many gods as well as animism (a belief in souls or spirits in man and other aspects of nature). Other scholars have supposed an evolutionary development of religion, which only reached monotheism—considered to be the highest form of religious belief—after a long period of purification. The two approaches sponsor, respectively, two contrasting myths about primitive man. According to the one, there was once a golden age of innocence and harmony; according to the other, the life of the earliest man was nasty, brutish, and short.

Granted the ubiquity of religion and its diversity, historians have found no universal essence expressible in terms of common beliefs. What is probably common to all religions is nothing more than the claim that reality is not restricted solely to what is yielded by sense experience itself.

The ambiguity of religion's influence on culture

The role of religion in culture. Religion has had a strong but ambiguous cultural influence. The thought that a man depended for his life and existence on a power not his own has encouraged some persons to be lazy, as it has inspired others to greater effort. A conviction about another world has led some religious people to disvalue human life; it has led others to view human life as having the significance of a state of probation. It has been plausibly argued by some (*e.g.,* the German sociologist Max Weber) that Protestantism provided a seedbed for modern capitalism; Catholicism, according to others, easily accommodates a Socialist point of view.

Because a religious view is generally associated with a conviction about the inadequacy of "things seen and temporal," religion as a cultural influence usually shows itself dissatisfied with things as they are. Often, however, when confronted with novelty, religion has tended to be conservative. Thus, religion has alternately opposed or fostered social and cultural development.

VIEWS WITH TRANSCENDENT REFERENCES

A situation is regarded as religious when through its spatiotemporal features what can be termed depth or another dimension can be disclosed objectively. In this sense, there cannot be such a thing as a religion that is nontranscendent. On the subjective side, there will be a matching self-disclosure, a "coming to one's self" that occurs as a response to a vision of the eternal in and through the temporal.

Relation to an ultimate power or being, to values, or to ideals. Different religious approaches can be distinguished by the different interpretations they give of what is objectively disclosed, of what in this sense is the transcendent. In primitive religion, for example, the transcendent is always interpreted in terms of an ultimate power or activity expressing itself, whether singly (monism) or with multiplicity (pluralism), through the objects and events of the world.

Animism, totemism, and interpretations of cosmic power

Animism views the world as having life, power, and feeling as do men. A monistic view of the universe is conceptually akin to the view according to which people or objects exert the peculiar influence they do and have the strange significance they possess because of mana—a power or force somewhat similar to the scientific concept of energy—that they embody. Animism becomes more diversified and pluralistic when it becomes spiritism, which locates the cosmic life, power, and feeling in particular objects. Totemism involves a highly complex system of beliefs and practices whereby an animal or plant becomes a totem, or a focal symbol for the life and well-being of a tribe. Just as tribal communities are sustained by a power that the totem symbolizes and expresses, so the patterns of tribal behaviour are maintained by taboos. Persons, things, and behaviour are taboo, or are prohibited to members of a society, when they are judged to be so highly charged with sacred power that ordinary "profane" persons must keep their distance.

These primitive viewpoints have a certain conceptual kinship with what the more sophisticated religious viewpoints have labelled with such terms as theism, polytheism, pluralism, and Idealism. Theism interprets the one cosmic, life-giving power in personal terms—different versions varying in their views of the adequacy of those personal terms. Polytheism posits a multiplicity of cosmic personal powers on whose activity (whether in cooperation or conflict) the universe depends. Pluralism views cosmic power as mediated and expressed through a multiplicity of ultimates (*e.g.,* finite persons) or otherwise views the universe as best understood in terms of ultimate atomic units, with no claim made for the absolute supremacy of any one of them. In this way, pluralism—even of a personal kind—differs from theism, which holds that God is a Supreme Person.

Absolute Idealism maintains that activity is an ultimate category but makes no claim, as does theism, for this activity to be personal. Instead, it takes a biological organism as its dominant model. Theism, like deism, has sometimes posited an ultimate personal power or being beyond, above, and certainly separated from the changing scenes of life, whereas absolute Idealism posits an ultimate power or being that is considered to be the whole, of which the changing scenes of life are but a part.

Seeking salvation in a life beyond. Religion is not merely a matter of being aware of a transcendent dimension nor is it merely a claim for a broader and more comprehensive view of the reality. Fundamental to religion is the conviction that through a right relation with a cosmic power or powers, man will find his salvation. Various views of such salvation have been held. Salvation has been regarded as something attainable only after this life. Other views, however, tend to posit a salvation for man through escape rather than fulfillment. Alternatively, salvation may be viewed as something anticipated in the present but fulfilled perfectly after this life. Salvation also has been interpreted in terms of fellowship with God or as a state of bliss needing no God (as by the early 20th-century British philosopher J.M.E. McTaggart), as a state of ultimate peace that arises when man sees his peculiar and rightful place in the whole universe (as by Spinoza), and as a state of bliss in which man cannot properly speak of himself as a self-conscious individual as in the Buddhist state of Nirvāna.

Interpretations of salvation

VIEWS WITH ANTHROPIC REFERENCES

Inner attitudes and dispositions. A religious view of the universe contends that a new dimension and depth can be disclosed within the person who responds. Though religious faith has its characteristic inner attitudes and dispositions, they must be of a transcendently self-involving kind, and there must be a depth to any attitude or disposition before it can be called religious. Thus, the attitude of awe is related to the feeling of fear. For fear to become awe, however, it must be characterized by a particular depth and self-involvement that come from responding to the presence and activity of God, or of the sacred or holy that call it forth.

Religion relates to the whole of a man's personality and because of this totality of human response, people speak of "conversion" in relation to religious attitudes. Generally, a person who becomes religious or ceases to be religious undergoes a profound transformation. Persons who have become converted to religion speak of the world as having taken on a fuller and richer dimension; those for whom the religious vision has disappeared speak of a world as having become flat, dead, and bleak.

The concept of conversion

Behavioral discipline with prescribed practices. Many religions bind their adherents to specific practices and particular moral codes. Thus, conversion has often shown itself in radical changes of behaviour; *e.g.*, an alcoholic becoming a total abstainer. Such behaviour as murder, lying, breaking promises, stealing, and commiting adultery have been condemned by the world religions. So strong is the ethical element in Confucianism that some regard it more as an ethical system than a religion. Yet, ethical (and ceremonial) codes can be transformed imperceptibly into no more than current social conventions and mere customs. Whether such codes have changed or not, their range and detail vary widely. Pork is eaten by Christians but is considered to be unclean by Jews and Muslims. Muslims and Buddhists abstain totally from alcohol; Christians and Jews need not. A Sikh will not shave his beard; but Hindus, Christians, and Muslims are free to do so if they wish. In contrast with Christians, Buddhists will not kill animals, and Muslims may practice polygamy.

Participation in a social institution. Whatever the diversities, religious faith is not only self-involving, but it has a social dimension as well. Hermits apart, religion brings people together as children of one family having a common father. For Christians, the significance of the universal religious community, the church, has been variously interpreted. Some, with a Protestant emphasis, have viewed the church as a voluntary institution created ad hoc for the convenience of its members to enable them to gather together to worship, to sing hymns, and to share common interests and beliefs. The Catholic view is that the church is a social institution that is derived from God and whose structure expresses the givenness of God himself.

To be of religious significance, however, social practices and moral codes, like inner experiences must have depth, a transcendent dimension, or they become superficial and dangerous parodies of religion, all the more dangerous for being in their outward features so similar.

THE VIEW FROM WITHIN AS PRIVILEGED

The assertion that the view of religion from within is privileged needs careful analysis.

First, religious faith is logically privileged insofar as it is characterized by a self-involvement, commitment to which partial commitments can only point. A temporary loyalty, however intensive, is at best a distant pointer to a conversion. Further, because religious faith is grounded in a disclosure, there is something logically privileged about it in the same sense that some are "privileged" to understand a joke when others do not. Yet, even though religion has a disclosure basis, it is still true that just as there are techniques for jokes so also are there techniques for meditation, whether in Christianity or in other religions. By virtue of such techniques men can have a reasonable expectation of a view of religion from within. In another sense, the view that religion from within is privileged may merely mean that if a man believes something and is committed, he is more involved than a man who does not believe.

One aspect of the logically privileged position of religion might be called its semantic privilege; *i.e.*, the fact that a religious vision cannot adequately be expressed. One fundamental problem for religious language, according to linguistic analysts, is to discover more reliable rather than less reliable ways of talking. One need not presuppose, however, so fundamental a distinction between the sacred and the secular that men become committed to total silence on religious matters. When St. Paul, for example, wrote of being "caught up" (in II Cor.) he "heard things that cannot be told, which man may not utter." If this statement of Paul's were generally true of religion, however, religious people would be so privileged that they would be living in a segregated silence.

Some scholars have argued that the privileged character of religion makes it unsuitable as a proper study for the philosopher, who must in principle be detached, not committed, and have an openness to all truth. The lack of finality in philosophical thought is contrasted with religious commitment and the final claims sometimes made for religious doctrine. Nevertheless, insofar as anyone has a coherent world view, there will be some degree of commitment.

Religion is not, however, altogether beyond argument, and those who are outside a religion can still have some inkling of what is being discussed within a religion and the manner in which it is being discussed, especially when the social, cultural, historical, and psychological embodiments of the religion are described. For this reason Western Christians and Jews, for example, are able to know something about the primitive religion of an African or Indonesian tribe.

Thus, much about religion can be known by those outside it, however, views about the nature of religion and definitions of religion have a systematic inadequacy about them. Like everything of the spirit, religion cannot be described so as to make clear to the detached observer the characteristic quality and depth of religious awareness and commitment.

THE DIMENSION OF RELIGION FOR INSIDERS

The essence or core of religion. For the insider, the essence of religion is given in a moment of vision and disclosure. Friedrich Schleiermacher, a German philosopher of the 18th and 19th centuries, described the basic religious experience in terms of a kiss or an embrace. Attempts to understand such a unity can only be made in terms of

the particulars into which the unity subsequently breaks, and such particulars then fall broadly into subjective and objective compartments.

Faith, prayer, and ritual

The subjective and objective aspects of religion. Faith describes a subjective state that accepts what a disclosure discloses and is akin to personal trustfulness, to a conation or striving that, according to Spinoza, all living things display. Prayer is the utterance of words (rite) with or without some dramatic context (ceremonial) designed either to carry one into the presence of what is worshipped or to express appropriate sentiments in the presence of what is worshipped. Most prayers incorporate words that function in both ways. Ritual is especially concerned with events in human life that have disclosure possibilities and in which mystery is at its highest.

Mystery in the context of religion refers to situations, such as birth, reproduction, death, and suffering, in which there are numerous possibilities for new insights and yet further insights. Public worship must constantly renew and realize in the liturgy the possibilities of the past disclosures. If the outward expressions and forms come to dominate, ritual can become an empty shell, and religious practices can become devoid of religious significance.

Effects of religious beliefs and practices. One of the effects of religious beliefs and practices is sacralization, a process in which certain persons, days, or objects become regarded as sacred. If such objects are granted more than the status of symbols, they may become objects of idolatry or superstition.

Belief in salvation, which often accompanies religious commitment, can have various practical results. If salvation is viewed as something that inspires progress and may be accomplished in the realm of time, such doctrines of salvation encourage social reforms and projects that envision an abundant life for humanity. If, on the other hand, salvation is viewed as something that is beyond the realm of time and set entirely apart from this world—something for which at best this world is a probation and at worst a sink of misery and iniquity from which the sooner man is released the better—such doctrines of salvation can be excessively individualistic and may even encourage oppression and tyranny.

Conclusions about nature and history

Religious belief has sometimes led men to detailed conclusions about nature and history. Good harvests have been interpreted as due rewards for appropriate worship or good behaviour, or both; calamities have been viewed as the results of sin, either ceremonial or moral. If God is believed to be in control of history, a nation that does what is right and follows his guidance, as expressed through its prophets and other religious personages, is expected to experience national prosperity and success. In previous periods, when this did not occur, the ensuing calamities were attributed to the backslidings of earlier generations.

Many observers of religion claim that in the modern world few would suppose that God intervenes in this direct and predictable way. According to this view, God's activity in the world, apart from being expressed in its constant creativity and conservation, is effective through man's own intellectual and physical activity. Insofar as man's own creativity is exercised, however, within the framework of the order that the world displays, and in no way violates it, one cannot exclude a similar creativity on the part of God. Admittedly, the fact that man expresses his activity through an intermediate organism (his body) indicates that there is no exact parallel between God and man; nevertheless, because God's activity terminates with the universe, the analogy with God's activity might very well be expressed in the number of ways in which man can effect creative development in his own body.

INTERNAL CRITICISMS OF RELIGION

Internal criticisms of religion have their basis in the imbalance that occurs when one aspect or one understanding of religion is allowed to dominate the rest. Heresies have arisen when one way of understanding has been developed without balancing it with another. In the development of doctrines concerning the nature and person of Christ within Christianity, for example, heresies arose when a particular model (*e.g.*, that of fatherhood and sonship) was believed to be capable of infinite development. The model of the Father–Son relationship was pressed too far, and the Son was subordinated to the Father in a way inconsistent with Christian orthodoxy, thus leading to what became heresy. Sectarianism develops when religious insights are associated exclusively with one particular doctrinal or theological phrase, such as justification by faith, or with one particular theological view regarding religious practices; *e.g.*, baptism. Because religion is at once infinite and mysterious, it is important that religious belief does full justice to a wide variety of approaches.

Another criticism of religion has been that it has tended to be overintellectual; and when this trait has been combined with moral laxity and factional rivalries, it has led to protests about the arrogance of intellectualized religion, often leading to the opposite error of supposing that belief does not matter as long as common sentiments are shared. Religious believers have not always recognized that for the most part their belief explicates metaphors, images, and symbols. Though ways of religious reasoning are appropriately informal and variegated, having their origins in a multitude of images and symbols, it nevertheless is considered a religious duty to produce the most reliable overall discourse based on the various images and models.

Balancing intelligibility and mystery

The fundamental difficulty of all religious understanding, however, is to balance intelligibility and mystery. If the intelligibility is neglected, religious belief can become dishonest and religious men can lose integrity; if mystery is neglected, there may be splendid controversy and exercises in logical appraisal, but the heart of religion will have disappeared.

The basic difficulty of all religions and of historical religions in particular is to effect a constant rebirth of symbols in changing cultures. In the course of time some of the most powerful images and symbols lose their fertility in promoting ideas that inform a religious community. This might be said of the image of sacrifice in the Christian religion. Religious practices and institutions, though they may have social merits, can all become stereotyped routine, as happens when they fail to preserve a sense of reverence and fail to disclose the givenness of the sacred or holy. Because religious belief is so important and influences all aspects of a society, there is a tendency for religious institutions to become authoritarian and oppressive. If a religious institution becomes interwoven with political views it can become tyrannical. Religion's only compulsion, according to some scholars, must be the compelling power of a vision, as the modern English–American philosopher Alfred North Whitehead expressed it: "The power of God is the worship He inspires." The authority of any religion is the authority of a vision, the authority of that which, in being disclosed, inspires men and leads them to fulfillment in their lives. For a Christian, the final authority is the love of God in Christ, and love is not love if its power is anything but inspiration. For other religions there is the compelling inspiration of that to which—Nirvāṇa or the Qur'ān, the Buddha or Muḥammad—point.

Internal criticisms of religion usually focus on such themes as narrowness, sectarianism, traditionalism, conventionalism, materialism, and immorality. Some criticism is also reserved for religiosity, which, though granting a dimension of faith, treats faith in an altogether superficial and often unbalanced way. Religiosity represents an excessive preoccupation with religion that is depicted in an incoherent and oversimplified relating of religious faith to intellectual views and social and personal practices.

THE REJECTION OF RELIGION OR RELIGIOUSNESS

Because religious commitment is so all-embracing and tends to influence thought, feelings, and behaviour, it is not surprising that there are many reasons why religious claims have been rejected.

Rejections on the grounds of alleged incoherence. Religious claims have been rejected because of their alleged logical or moral incoherence.

Alleged logical incoherence. Logical incoherence may arise internally or externally and in relation to different issues. In regard to internal coherence, critics have maintained that man should be able to expect that God would

see to it that there could be no possibility of ignoring his existence or of making mistakes about religious beliefs and behaviour, if religious convictions are so important. They have also claimed that it is altogether too naïve, though inevitable, to think of God as made in the image of man. Some have rejected theistic belief because of the incoherence of the idea of God, which must—they claim—combine so many incompatible predicates; *e.g.,* God is eternal, yet acts in time, or he is loving and yet incapable of suffering or feeling.

External incoherence, in conflict with other views of world

Religious beliefs have been alleged to be externally, as well as internally, incoherent because of their conflict with other views about the universe, especially scientific views. The doctrines of heaven and hell, in particular, which have given great personal and social significance to religious belief, have been rejected by many critics when these doctrines were viewed literally. Yet it has been the supposed actuality of heaven and hell that has given religious persons their hope and their terror respectively. Absolute Idealism, it has sometimes been alleged, is incoherent insofar as it states that time is not "real" and that evil does not really exist. This is not to say, however, that there is no temporal succession or nothing evil, claims that would be obviously incoherent. What is being claimed is that within a particular interpretation of the universe, time and evil are not left as ultimate categories but are in some sense derivative from other categories.

It has been argued that by far the greatest problem of external incoherence that belief in God has to face is that of the evil and suffering that characterize the world. Critics have stated that if God cannot rid the world of evil and suffering, he is not all-powerful; if he could, but he won't, then he isn't all-good; if he is powerful and good but not all-wise, then, even though he is trying his best, there are bound to be disasters. The most serious classical expression of this problem was given by David Hume, in his *Dialogues Concerning Natural Religion* (1779). With such considerations in mind, some philosophers, such as John Stuart Mill, have been willing to argue for a limited God—*i.e.,* the great fellow-sufferer who understands and has compassionate sympathy.

Alleged moral incoherence. Though religious conviction shows itself in moral behaviour, it has been argued that religious people have not shown outstanding moral qualities. An 18th-century English philosopher and churchman, Bishop George Berkeley, when presented with this objection, remarked that nothing evil can be attributed as such to the Christian religion and that the only legitimate comparison is that between a person who is a Christian and what the same person would have been otherwise. The distinctiveness of the Christian faith, however, has sometimes been supported by arguing a stark contrast with morality. The 19th-century Danish philosophical theologian Søren Kierkegaard, for example, by a too literal misreading of the biblical story of Abraham and Isaac (Gen. 22), supposed that religious obedience must be in radical opposition to moral duty. However that may be, religious men often may be only too well aware of their moral lapses, their sins, and for this very reason they seek the grace and power of God. The good that they would do they do not do, and the evil that they despise they continually do, as St. Paul noted in his letter to the Romans. In this moral predicament, those with a Christian commitment believe that the grace and power of God comes to inspire and release them from the dominion of sin. This does not mean that the Christian never sins, but it does mean that he is assured of ultimate victory over sin. The Christian Church is viewed not as a society of saints but a school for sinners.

The problem of sin and grace

The exclusiveness of religious sects is regarded by those outside the sects as hardly to the sectarians' credit. For Christians, sectarian exclusiveness is viewed as a scandal to the gospel that they preach. On the other hand, the criticisms of Puritanism that hold it as inevitably negative and oppressive sometimes fail to see that it may be neither negative nor oppressive if it is grounded in a spiritual and religious vision.

The doctrine of grace (the view that God grants man abilities that man does not merit by his own efforts) has sometimes appeared to make God himself—interpreted as the spirit dwelling in a man—the actual agent of good behaviour. In this way, some interpretations of the doctrines of grace have compromised man's freedom and come close to denying man responsibility for his actions.

Outside Christianity, critics have pointed to the gap between religious profession and moral action, though within Christianity, with its strong emphasis on moral transformation, the gap has been very wide and the criticism most challenging. In Hinduism for instance, Gandhian reformational and nonviolence ideals have not mixed well with social corruption or with the type of neutralism that allowed China to persecute Tibetan Buddhists. Again, the Buddhist who goes to a temple is not necessarily compassionate as his religion dictates, and the Muslim who attends services in a mosque may be less filled with an inner sense of justice and patience than with thoughts of a holy war. In Sri Lanka (Ceylon) and Vietnam nationalist loyalties have given rise to a violence untypical of Buddhism. In the last resort, however, each religion will appeal to its doctrine of salvation when presented with a gap between its moral ideals and the actual actions and behaviour of its adherents.

Other grounds for the rejection of religion. *Rejection of historical beliefs, practices, and institutions as spurious or irrelevant.* When a religion appeals to historical events, other grounds for its rejection arise. The Old Testament view of history appears to have been exceedingly selective in order to emphasize a particular point about God and his activity. The miracles of Jesus—both those relating to his own person (his birth and Resurrection) and those that he himself performed (especially nature miracles)—conflict with what men experience in the normal course of their natural lives and experience. Prayers requesting favourable weather, plentiful crops, or safety in a journey are characterized by many as spurious and irrelevant. Ideas of God intervening in the universe, according to such critics, satisfy neither science nor religion. From a scientific point of view, "laws" of nature are no longer viewed as divine prescriptions; and the word law becomes, in fact, misleading. Furthermore, in order to allow for miraculous intervention of this kind, God's providential care is viewed as a compromise. He thus becomes the absentee landlord who absented himself from the world, which must take care of itself except for some spectacular visitation. According to this view, the only coherent way to speak of an intervention of God is to interpret it in the context of personal intervention.

Religious institutions have been criticized on the grounds that they conflict with the ideas of the founder and are supported by claims that cannot be historically verified. These claims, according to critics, depend on taking certain historical events on which the religion is founded, and reinterpreting them by theological speculation or a very full imagination, to produce, for example, a doctrine of papal supremacy according to which Christ is believed to have given explicitly to the successors of St. Peter final jurisdiction over the church.

Rejection of religious sentiments or dispositions as valueless. According to some views, anyone who prizes "another world" must despise this world and be uncertain in his attitude toward the world around him. In this way, it is said, religion dries up the sources of its activity and attacks such happiness as this world can provide—though promising happiness hereafter, which has been called "pie in the sky" or "opiate of the people" by critics of religion. A humanist concern to liberalize and relax laws (*e.g.,* on abortion and divorce), to abolish capital punishment, and to encourage birth control has always been opposed, according to humanists, by Christian orthodoxy, which they interpret as having a negative and conservative attitude that has proceeded from a nervous fear of a decline in moral standards. At the same time, humanists would continue, moral standards have hardly been upheld by sectarian strife and persecutions. Further, they point out, too often the church, in its desire to indicate what abides, has confused what is abiding with current social and political institutions and traditions inherited from the

Humanist criticisms of religion

past, generally resulting in an illiberal obscurantism and a reactionary outlook.

Some critics of religion have contended that almost all scientific progress has been hindered by religious beliefs and attitudes. Biology, physics, and geology, they have claimed, only made the rapid progress that they did when they were freed from a context of religious belief by the 17th-century philosopher René Descartes, who devised a metaphysical myth of the separation of mind and body.

Naturalistic or skeptical views of the origin and development of religion. In the matter of the origins and development of religion, many (*e.g.*, the psychologist James Henry Leuba in his *Psychology of Religious Mysticism* [1925]) have argued that there is a close connection between mysticism and hallucination, between hysteria and ecstatic institutionalized inspiration as, for example, in Pentecostal churches. Religious people, according to such views, often have personality weaknesses and are psychologically disturbed. Freud, the founder of psychoanalysis, maintained that inner conflicts—often the result of repression, particularly in relation to sex—become expressed in peculiarities of behaviour and mood, especially in the vivid imagery of dreams that erupt from the unconscious area of one's personality. By comparing the symbolism of dreams and mythology, Freud held that belief in God—in particular, the father image—merely perpetuates in fantasy what the individual must in actual fact overcome as part of his growth to maturity, thus giving religious belief a treatment that not only made belief in God unnecessary but positively unhelpful.

Carl Jung, a former disciple of Freud, gave a different account of the psychology of the unconscious. Each person displays a libido, a fundamental striving that is creative and purposive and of which there is evidence in the symbolic language of dreams. Behind all such symbolic language are archetypes (innate tendencies to form symbolic images), which all humanity shares and which inspire a person to move toward a balanced integration to which the energy of the libido would creatively move, if given proper freedom and encouragement. Thus, Jung posits a racial or impersonal unconscious in which, at the deepest level, all individual human beings share. Jung's archetypes raise the metaphysical question of whether they are symbols of an existent God or gods—a question that psychology leaves open. For many psychologists it is a question of little interest, because for them the archetypes themselves suffice in practice.

In addition to such naturalistic or skeptical views about the origin and development of religion are other claims that religion is merely an infantile reaction to fear, a more or less harmful sublimation of sex, a projection of wishful thinking, or a social device for use in the class struggle. On the other side, however, it is likely to be pointed out that one must be careful not to indulge in the genetic fallacy: no account of the origin and development of anything, of religion in particular, is necessarily a reliable analysis of what that particular phenomenon is now; a single explanation of the origin and development of a phenomenon as complex and variegated as religion is difficult to describe and maintain. It is also necessary to beware of the "really only," or reductionist, fallacy. To say "*x* is really only *y*" is, in effect, denying the significance of *y* language despite the fact that *y*-talk as well as *x*-talk already occurs; *e.g.*, persons are really only "machines," or worship is really only a social occasion. Over-simplification streamlines discourse at the cost of adequacy and truth.

Problem of the genetic and reductionist fallacies

Some have thought of religion as no more than a body of stories designed to encourage a noble attitude toward life and humanity. If, however, one asks why or how these attitudes encourage and why a particular attitude is valued, what begins as a simple account of religion becomes, in the end, as complicated as any. Another criticism of religion, arguing for its redundancy, claims that the progress of man in society can and should be determined by scientific considerations. This contention, however, goes beyond the particular conclusions of the individual sciences; it is to make a philosophy out of science. On the one hand, such a scientific view of man and society would be open to philosophical criticism, not the least if it were suggested that man's subjectivity—that which makes him the unique person he is—has to be analyzed in terms of the objects of science. On the other hand, if science becomes a philosophy, it might be said to have assumed a religious dimension itself.

In the realm of religion in the latter part of the 20th century, in what might still be called the Christian societies of the West, the attitude of very many people lies in an intermediate zone between religious belief and atheism, but the content appears rather to be given to agnosticism. Such persons believe in God but dislike any kind of formal worship, pray only on exceptional occasions, and find it difficult to have a sense of sin but admire saintliness. They are critical of the need for a Christian ministry except insofar as a priest or pastor can show sympathy and act as a vehicle of social concern. They are distrustful of dogma and critical of Christian sectarianism. They may be uncertain of Christ's divinity, but the words and example of Jesus are viewed as a guide to the good life. This outlook has many affinities with the "natural religion" of the 18th century in which the ethical example and teachings of Jesus were emphasized. Though, as in the 18th century, there may be an intent to reject revelation, persons holding such an outlook may rather be rejecting certain stylings of Christian revelation.

Ambiguities in attitudes of persons in Christian societies

Examples of occurrence of such a "natural piety" can also be found in religions other than Christianity, though significantly not in Islām—unless the Bahá'í movement be taken as an approximation of this outlook. This attitude, for example, has provided the basic cohesion for the State of Israel in the latter half of the 20th century. Further, the spread of technology has gradually been alienating many Hindus and Buddhists from their traditional beliefs, but the Hindu has continued to treasure his spiritual ideology, which may well give to technological development its needed direction and wider setting. Buddhism in Japan, and perhaps elsewhere in the East, is still valued in the 20th century insofar as it supplies a local religious dimension to a society whose public and industrial life has been increasingly Westernized. Thus, an attitude has arisen that is sympathetic to the broad claims of religion, but has been critical, if not disdainful, of theological dogma and rivalries.

THE ACKNOWLEDGMENT OF RELIGION OR RELIGIOUSNESS AS VALID

Traditional justifications. *Religion as pointing to an ultimate power, being, or value.* More generally, persons who are outside the particular religions and who have nevertheless acknowledged religion as significant often seem to base their views on a fundamental feeling of absolute dependence. The grandeur of the universe, the character of the moral struggle, reflections on human nature, and an awareness of moral values inspiring men to reform society have all joined together to point men to an ultimate power or being—a "power, not ourselves, which makes for righteousness," according to the 19th-century English poet Matthew Arnold.

The fundamental difference in the latter part of the 20th century between the secularist and the religious person most likely has been between someone who takes a narrower and someone who takes a wider view of humanity. That there is an acknowledged need in modern times to give a moral direction to technology seems to many to bring with it the need for a religious view of the universe, even though they may not themselves be adherents of a particular religion.

Religion as producing wholesome spiritual or moral effects. Others point to examples of the wholesome moral and spiritual effects that religion has had. They mention that society has ceased to practice child exposure and there has been a notable development in the status of women in society. Religion, where it is not parodied, misrepresented, or misunderstood, broadens rather than narrows vision. Insofar as human nature is inadequately understood, if no place is granted to the spirit of man, human nature, it is argued, will never find satisfaction except through the self-realization and self-fulfillment that come from responding to the inspiring ideal.

The pragmatic view

Alternatives to traditional beliefs, practices, and institutions. *The quest for authentic existence.* In the 20th century various alternatives to traditional religious beliefs, practices, and institutions have become apparent. Chief among these is the quest for authentic existence. This has been encouraged and portrayed by various Existentialists (those who view man in terms of his existing thoughts and actions rather than in terms of his "essence"), who have been concerned in one way or another with emphasizing the significance of certain situations. In this way, they have given their own versions of salvation—that situation in which a person finds his true significance. For some, such as the German philosopher Martin Heidegger, a sense of authentic existence is given to each person when he realizes his true subjectivity, which his life in the world and his social transactions so often conceal. Authentic existence is often contrasted with cosmic anxiety—*i.e.,* anxiety of a deep and far-reaching kind to which the antidote is to find oneself and one's freedom in a total commitment to what is called the ground of Being.

Existentialists of an atheistic persuasion, such as the philosopher and Nobel laureate Jean-Paul Sartre, regard human existence as absurd and other people as hell, because, though one needs other people, they can never be other than "other people"—their subjecthood, their freedom is inaccessible. Love is, thus, doomed to permanent frustration. The need to know others like oneself is matched by its impossibility. According to Sartre, this condition only reflects the absurdity of man's own existence, which is always attempting to overcome a radical estrangement between man as the object of scientific study and man himself (*en soi*) and the subjectivity man knows in consciousness (*pour soi*). Suicide is the final absurdity, for in getting rid of *en soi,* what man is, *pour soi* disappears at the same time.

Religious Existentialism as a God-given fellowship

This pessimistic estimate of human life and its apparent absurdity, however, has been converted into a religious view by other Existentialists, such as Gabriel Marcel, another French philosopher, who point to a participation—a mysterious self-involvement that persons can have intersubjectively with each other—in a kind of fellowship that is viewed as God-given. According to this view, man needs to open himself to the presence and grace of God for a dynamic transformation in which the mysterious transcends the purely problematic. Common to all Existentialists, however, is the view that the authentic man is not merely satisfied with playing a role, with being a cog in industrial society. One way or another, the quest for authentic existence is to discover the means by which man can recapture and enjoy occasions of self-disclosure. So significant are these occasions that they have been viewed by some theologians to be the paradigm for the kind of situation that the Christian gospels recount.

Secular religion. Another feature of 20th-century development has been society's rediscovery of the significance of the secular. This change has led to an outlook and attitude that has been characterized as "religionless Christianity," a Christianity influenced by its residual social and political ideal, but bereft of its specifically religious practices, doctrines, or institutions. Such practices as traditional intercessory prayer are dismissed as empty approximations to magic; doctrine is condemned as outdated and expressed in terms of past cultures; institutions are criticized as oppressive and conservative.

Behind all this suspicion of structures and doctrinal schemes and practices, however, is a desire to get back to basic principles and origins, to learn again what is distinctive about the religious point of view. According to some proponents, such a goal might be attained by beginning with the secular, with activities in the secular world, not least with compassionate service, by seeing where the need arises for religious conviction and by ascertaining what contribution faith will make to secular endeavour. Though secular religion broadens out into a more sympathetic and a more positive attitude than agnosticism, it is never as explicit or particularized as orthodoxy.

Marxism. Marxism, which provides remarkable evidence of the power of dominant key ideas to inspire and direct man, is undoubtedly one of the greatest challenges to traditional religious belief. Based on the socio-economic philosophical thought of the 19th-century thinker Karl Marx, Marxism can be said to be a quasi-religion on two counts. First, Marxism had connections with the metaphysics of G.W.F. Hegel, an 18th–19th-century German philosopher who interpreted reality in terms of a spiritual Absolute. Furthermore, the thinking of Marx had religious overtones, whether from his own Jewish background or from a Christian atmosphere, not least in Britain where he lived from 1849 to 1883. Second, Marxism can be called a quasi-religion insofar as it calls from its followers a devotion and a commitment that in their empirical character greatly resemble the commitment and devotion that characterize religious people. Marxism has undoubtedly fired the spirit of man and given to revolutions, whether in Russia or China, a powerful direction that has maintained stability and avoided anarchy. Furthermore, like a religion, it has provided themes of fulfillment and hope—a revolution interpreted as the initiation of a Communist world society that would be a final consummation. There are many logical similarities between the doctrine of the Marxist millennium and the Christian doctrine of Christ's Second Coming. Marxism has also stressed the significance of cooperating with the immanent spirit of the times—something comparable to the providence of God—in economic and military struggles that are viewed as the travail by which society would be reborn. The main difference between Marxism and Christianity in the 19th and early 20th centuries, according to some scholars, was that for many the Christian vision encouraged men to endure tyranny, while the Marxist view inspired men to rebel. Yet, once it can be established that religion is not the servant of oppression, is not necessarily linked with an illiberal regime, and does not use concepts of "other worldliness" to make men content with tyranny and injustice, then religion may yet have a place in the Communist state. Such a religion would not have to concern itself with the kind of supernaturalism that Marxism now rejects; it would not have to appeal to an invisible world entirely other than the present world. It is not without significance that Marxism has its own form of public ceremonial and its own language of glorification. If it has to be granted that many religions have a ceremonial, a symbolism, and a moral code that has lost the vision they once had, Marxism is a social program, a doctrine, and a ceremonial searching for a vision that haunts it and that may at some time bring it to fruition. In this regard, Chinese Marxism is particularly significant insofar as Marxism in China cannot escape some interweaving with Chinese Buddhism. Chinese Buddhism brings with it a natural framework of absolute Idealism, which may yet supply Marxism with the spiritual dimension that for many critics appears to be Marxism's main inadequacy, something it lost when it shed its Hegelian metaphysics and became the anti-God Materialistic world-view of the U.S.S.R.

Marxism as a quasi-religion

HISTORY OF THE PHILOSOPHY OF RELIGION

Most philosophies have incorporated religious views in the wide sense of being concerned with a reality beyond appearance, and in this sense they have provided a philosophy of religion.

Developments in the West. *Ancient and medieval concepts.* For the Greek philosophers Plato and Aristotle, wonder was the beginning of philosophy. From such wonder, according to Plato, emerged religious knowledge that was also mediated through Ideas, eternal entities or concepts in which the things of time participate. In performing every good act, man realizes his link with eternity and the Idea of the Good. For the moment, however, man, as in a cave, is chained by his earthly existence so that he cannot see the light outside; he can only see shadows on the wall, which are signs and tokens of the eternal light behind him. This was Plato's way of styling the relationship between time and eternity, between appearance and reality, and it is a styling that found a particular welcome in the Christian tradition and not least by Christian Platonists, whether of the 2nd or 17th centuries. Plato's philosophy also led to belief in God, and his *Timaeus* is a philosophical creation story.

Religious ideas of Plato and Aristotle

Aristotle, impressed with organic life in man and animals, took as his fundamental category growth and development. The nature of anything was thought of as a form by which its movement and development as an organism was to be understood. It was as if the form supplied the driving force. In this context, God was thought of as pure form, as final cause, and as prime mover. Aristotle provided for St. Thomas Aquinas, the great medieval philosopher of Western Christendom, the foundation on which he developed Scholasticism, which has been a distinctive feature of Christian philosophy of religion since the 13th century. Other medieval philosophers, such as Erigena, with his pantheism (God in all); Abelard, with his critical questions; Eckehart, with his mysticism; and Duns Scotus and Bonaventure, with a wider view of reason than could be contained in the Scholastic philosophy, all illustrate the variety and independence of Christian thinkers.

Modern concepts. Descartes, the "father of modern philosophy," is significant in terms of his reacting against external authority in matters of belief, seeking a fresh basis for certainty, and finding it in the existence of his own mind. He must think in order to doubt his existence, hence his famous statement, *Cogito ergo sum* ("I think, therefore, I am"). Henceforward, much significance was given to the individual mind, and the resulting myth of the body–mind separation enabled both physics and biology to develop without the risk of ecclesiastical interference. Only in recent years has the inadequacy of the Cartesian body–mind myth come under general criticism not only because of the metaphysical problems it poses but also because it fails to do justice to the unity of personality that recent developments in medicine, such as those pertaining to psychosomatic disorders, presuppose.

The myth of the body–mind separation

Many of Descartes's 17th- and 18th-century successors can be best understood by reference to him. Nicolas Malebranche, a French Cartesian philosopher, and the occasionalist philosophers, were more radical than Descartes; they dispensed with any unity whatever in man himself and linked together man's mind and body by means of the constant correlation effected by God himself, claiming that mental events were merely "occasions" for God effecting material change. For Spinoza, the whole universe had not only Descartes's two attributes of mentality and materiality but an infinite number of attributes, and it could be alternatively named God or Nature. Each existent in the world could be pictured as a particular whirlpool in an infinitely deep sea made up of endless layers of particular fluids of which man knows only two—mentality and materiality. Gottfried Leibniz viewed Descartes's minds as the only ultimate existents, so that even material things were colonies of souls. God was viewed as the supreme monad (the ultimate substance) that establishes coherence and harmony among all other monads. What appears to men as the external world is, so to speak, the result of blurred vision on the part of those groups of monads that are human beings.

After Descartes there appeared the British Empiricists: John Locke, George Berkeley, Joseph Butler, and David Hume. Locke, though rejecting some of Descartes's characteristic doctrines, nevertheless took over Descartes's view of the human mind and then concerned himself with the philosophical psychology of how the mind comes to have the ideas it possesses. By the time of David Hume (died 1776), the mind was viewed as nothing more than a collection or bundle of ideas thought of as very similar to images, which means, as Hume frankly admitted, that it becomes impossible to do justice to the subjectivity that makes each person distinctively the person he is. The significance of Berkeley (died 1753) in this sequence is that he saw the need for an extended Empiricism that took the notion of personality seriously and that regarded activity as a key concept. Indeed, for Berkeley the fundamental unit for thought was "activity-directed-towards-and-terminating-in ideas," and it was the activity of God directed to those ideas, which make up the external world, that gave to this world its continuous independent existence. His contemporary Butler also argued for a broader Empiricism, which for him centred on the significance of man as a moral agent and on a reasonableness that need not always conform to a mathematical paradigm. In a matter of great consequence, a man's action can be reasonable even though there may be little supporting evidence for his decision and though, indeed, the evidence may be very much against it. It may, thus, often be a moral duty to act in such problematical circumstances. This led to Butler's famous doctrine of probability—"probability is the very guide of life"—a view that influenced the treatment of belief in *The Grammar of Assent* (1870), by the English theologian John Henry Newman.

Immanuel Kant has been called the second founder of modern philosophy. With Kant, late 18th-century philosophy began to take an interest in human knowledge, its varieties, scope, and limits. In Kant's critical philosophy, which emerged in his old age, he showed how scientific knowledge left room for morality. Though he was inclined to interpret all religious assertions in terms of morality, belief in God was justified as the holding of a regulative idea that brings coherence into all of man's thinking. The foundation of this idea is to be found, in fact, in those experiences of unity to which moral ideals, beauty, and the notion of a purposive universe all point. This idea of unity, largely implicit in Kant, was developed by Hegel, who came to regard the universe and its cultural, social, and political progress as but manifestations in time of an unchanging absolute spirit. In this way, Hegelianism provided a spiritual interpretation of the universe, but it regarded particular religions as no more than visual aids toward understanding Hegelian truths. A century later, the British philosopher F.H. Bradley was able to use a Hegelian approach in a much more empirical and far less intellectual context. Whatever form Hegelianism took and though its spiritual insights seemed on first view to make it a friend to religion, it has proved to be a position in opposition to Christianity, whether by its minimizing the historical element or by the way in which it compromises belief in a personal God.

The Kantian and Hegelian idea of unity

Since the absolute Idealists, there has perhaps been only one philosopher in the mainstream of tradition—Alfred North Whitehead—who, in taking becoming rather than being as the fundamental category, made "process philosophy" possible. This philosophical view maintains a metaphysics that not only provides an interpretative scheme linking God, man, and the world but one that incorporates scientific and historical thinking, though in taking growth and process as fundamental, Whitehead seems, to some, to have an evolutionary God.

There were two main reactions against Hegelianism. The first, initiated by Kierkegaard, viewed Hegelianism as altogether too detached and objective and its ways of reasoning entirely unsuited to the deepest experiences of human life, the tragic situations in which human beings find themselves. From Kierkegaard, the Existentialist movement began. Also, in reaction against Hegel, were the modern Empiricists, such as Bertrand Russell and G.E. Moore from England, whose watchword was clarification in their attempts to create a straightforward, unambiguous language. This movement passed easily into Logical Positivism (a philosophical position that accepts only scientific knowledge as factual and rejects metaphysics), which challenged not only the truth but the meaning of theological assertions.

Developments in the East. *Buddhist concepts.* Among the religious philosophies of the East, the conservative Theravāda (Way of the Elders and another term for Hīnayāna) Buddhism regarded all existence as a succession of transitory states: what alone was permanent was Nirvāṇa, a deathless realm the existence of which was revealed to the Buddha himself in the Enlightenment that came to him while he meditated beneath the bo tree (late 6th century BC). About Nirvāṇa, the wise will say little more except to affirm its existence and to express their conviction that the plurality of individual souls that man knows in this world cannot in the same way exist in that deathless realm where there is no rebirth. Such ideas find a natural home in the philosophical standpoint of absolute Idealism, and Nirvāṇa can be regarded as an alternative word for the Absolute. Broadly speaking, Buddhism is agnostic both about a personal creator and personal im-

The concept of Nirvāṇa

mortality, though Theravāda Buddhism explicitly rejects belief in a creator. Undoubtedly, the dominant theme of Buddhism is the quest for release from the changes and chances of this world, which will lead to the serenity and peace of Nirvāṇa. A Buddhist saint is someone who has indeed become the Absolute, which thus incorporates and transcends all human imperfections and struggles and all the imperfect ideas, ideals, and deities of popular religion and popular ways of thinking. The difference between the *arhat* of Theravāda, and the *bodhisattva* of the Mahāyāna is one between two different routes of realizing Nirvāṇa—the one through self-concentration; the other through self-sacrifice for the welfare of others. The difference is one between two "saintly" routes to the one saintliness—being possessed by and dwelling in the Absolute.

Thus, Buddhism, by embracing what is, in effect, a metaphysical concept of the absolute, not only could but did hold together a complex mythology within a unifying philosophical insight and was able, as in Japan and China, to incorporate a complex popular pantheon of the cult of ancestors. Furthermore, it could combine a popular devotion to a personal lord with a mystical contemplation that had encouraged the development of Buddhist monasteries. In sponsoring such a broad synthetic (all-embracing) view, the philosophical significance of Mahāyāna (Greater Vehicle) Buddhism emerged. Such developments began about 100 BC and lasted for several centuries; it was Mahāyāna Buddhism that spread to China and East Asia to influence and modify the religions native to those areas. In Mahāyāna, the humanitarian saviour notion of the *bodhisattva* has some echoes in Kenotic Christianity (*i.e.,* emptying oneself to become a suffering servant), and attitudes to the Buddhist scriptures have parallels with those of Christians toward the Bible. Common to both is the view that revelation can express itself in developing forms and that it is a mistake to concentrate on the texts themselves, sacred though they are, rather than on that which transcends them and of which they are symbols and to which they point. In this respect, one may contrast the open and exploratory attitudes of many Buddhists and Christians toward their sacred books with the closed and rigid attitudes of most Muslims.

Confucian, Taoist, and Japanese concepts. Prior to the introduction of Buddhism into China in the 1st century AD, the two main strands of religious thought in that country were represented by Confucianism and Taoism. Confucianism displays a reverential propriety that is expressed and developed in social relationships and fulfilled in Heaven. Taoism claims that the wise man will constantly seek harmony and rapport with Tao (the Way), which, at one and the same time, is the way for men to follow if they would reach blessedness and the principle that underlies and sustains the world. As a concept that is both moral and cosmological, Tao has a logical status similar to that of the Logos (or Word, the active principle of God in creation and revelation) in Christian philosophy. The Taoist thinks little of the ways of the world, including the decorum of the Confucianist; his outlook rather encourages a laissez-faire policy toward the world and even withdrawal from its affairs. The immediate mystical experience of Taoism or the inspired behaviour of Confucianism can easily blend with Buddhism, which sets both within a metaphysics of the Absolute.

In Japanese religion are found the same two themes that are found in most religions, though in their extreme forms they are mutually exclusive. On the one hand is mysticism—specifically, nature mysticism, of which the mysticism of the Zen practitioner belonging to an intuitive meditative form of Buddhism is a specific example. In Zen Buddhism, religion is scarcely distinguished from an aesthetic experience in which shrines, gardens, mountains, woods, and streams reveal a mysterious beauty and in which the exercise of the intellect is at a minimum. In contrast to mysticism, there is devotion to a supreme personal lord, at one time symbolized in the emperor as a descendant of Amaterasu, the Sun-Goddess. At other times Shintō ("Way of the gods") devotion focussed on particular shrines and particular deities, just as Zen Buddhism could concentrate on a particular image or on particular events. In both types of devotions, however, it could be argued that such particularity was fulfilled and transcended in the unity revealed to a mature mystical insight. These different philosophical positions have an interesting reflection in the Christian position in which the Christian claims to find evidence of God's presence and activity in particular places and situations (especially in the incarnation of Christ), though at the same time allowing for God to be omnipresent.

Hindu concepts. This mixture of a mystical contemplation, which sees the divine everywhere, and a personal devotion to a particularized divinity recur in Hinduism. The most characteristic feature of Hinduism, however, is the doctrine of an eternal soul and its rebirth. The universe is pictured as the arena in which the immortal soul engages in a succession of incarnations from which man seeks release, a release that true contemplation can give him, especially when approached through Yoga (a mental, physical, and spiritual meditation technique). At the same time, a sensitivity to the numinous (spiritual) has left open the possibility of and certainly encouraged personal devotion. The most famous of Indian scriptures, the *Bhagavadgītā* ("Song of God") has for its recurrent theme the majesty, glory, and terror of God and the devotion due to him, though as in Christianity these attributes are compatible with a loving God. In the matter of revelation and incarnation, it is an open question as to how far the Hindu conception of revelation is the same or similar to Christian or Muslim conceptions. The Hindu view of *avatāra* ("incarnation"), however, implies many incarnations and in a Christian context would demand many Christs; thus, the concept of *avatāra,* a salient feature of Vaiṣṇavism (centring on the veneration of Vishnu, the preserver), cannot be easily reconciled with the uniqueness attributed to Jesus.

Doctrine of an eternal soul and rebirth

Depending on the particular questions that determine a particular content of discussion, Hinduism can talk of a plurality of souls, when it would concentrate on the theme of reincarnation, but, especially when influenced by Buddhist (and also pre-Buddhist) ideas, it can also sponsor an absolutism, or a monism; yet, again, it can come very close to a traditional Western theism. On the whole, however, it might be said that Hinduism holds together in a creative tension both theism and monism, though often it appears that in conceptual foundations and philosophical discussion the theistic strand predominates. Even in its classical period (600 BC to 450 BC) Hinduism was characterized by an astonishing variety of doctrines and cultures. Indeed, it well illustrates a characteristic of Indian thought that is becoming more acceptable to Western ways of thinking—the notion that there are many different approaches to the truth, which matches the concept of a multiple theology. It was regarded, however, as a retrograde step when these varieties of culture, ritual, and mythology became hardened into social strata and castes.

In the medieval period, Śaṅkara (*c.* 788–820), the leading exponent of Advaita Vedānta, or nondualism, is the most significant Hindu figure in the philosophy of religion. Arguing in a way very reminiscent of absolute Idealism, he claimed that the only existent was an absolute and that all else was an illusion. In this context he equated *ātman* (the individual soul) with Brahman (the universal or absolute soul), Both were viewed as one in a cosmic consciousness. For Śaṅkara, only ignorance or lack of insight into the nature of being prevents a man from realizing his identity with Brahman and thus becoming here and now aware of the freedom that is his. Śaṅkara also allows as permissible, without being accepted as the truth, talk of God as personal and as creator and of men as separate souls related to one another and to him. This, however, is only considered a way of talking—salvation in the Absolute transcends all such imperfect discourse. The same logical problems recur here in the concept of Nirvāṇa in Buddhism. In Hinduism, the *Upaniṣads,* Hindu philosophical treatises, and the *Bhagavadgītā* use the imperfect language of finite man, who has not yet found release, and in this way they can only point beyond themselves to that which they cannot adequately express. Here again are ideas reminiscent of some of those in Western philosophy of religion in the

Concepts of *ātman* and Brahman

modern world: the importance of theological reticence, the limitations of theological language, and, in another context, the significance of "existential situations."

Twentieth-century Hinduism has been chiefly characterized by attempts to purify and reform the doctrines of its medieval period, to deepen its spirituality, to reassert its moral dimension, and to inspire social reform. Mahatma Gandhi and Sri Aurobindo, the founder of a spiritual community and a Communist, were significant in such ventures. Aurobindo has been compared with the French Jesuit paleontologist and theologian Teilhard de Chardin insofar as both have a repeated experience of cosmic consciousness and a profound belief in evolution, both of which point to a divinization of man.

Islāmic concepts. At the heart of Islām is an experience of awe before the one, all-powerful, mysterious creator Allāh. Thus, its dominant theme has been surrender, though it must not be forgotten that it has nurtured mystics to whom the mysterious and awesome God has revealed himself through created things. Allāh controls man's destiny, whether to salvation or damnation, which points to the ultimacy of God, to his majesty and power. The concept of heaven inspired warriors to fight to the death; the concept of hell encouraged loyalty by showing what terrible punishments awaited the disloyal. The Qur'ān (the Islāmic sacred scriptures) is regarded as an infallible book—a transcript of a tablet that is eternal in the heavens. Islām shows pre-eminently the strength and limitations of a total surrender based on clear-cut beliefs, themselves arising from a basis in infallible texts, the whole being translated into vigorous political and social practices associated with a rigorous ritual and ceremonial discipline. Its mixture of both rigour in theology and vigour in politics in India and the Middle East from the Middle Ages to the 20th century can perhaps be compared with the same mixture as has been seen in the Protestant and Catholic communities in Ireland since the 17th century. However much the concept and practice of holy warfare is repugnant to many minds today, in the context of Islām it implies a sensitivity to evil and a conviction that evil has to be resisted and overcome in a total dedication. In this way the faith of Islām has shaped human history by obedience to a resolute and powerful God. Islām also illustrates the point that predestination need not bring with it a submissive fatalism. Furthermore, it has to be granted that Islām has allowed, within itself, for some allegorical interpretations of the scriptures—explicitly by the Ṣūfīs (mystics)—and it has also allowed for differences of piety and beliefs and even intellectual exploration on the part of particular disciples. Nevertheless, to other religions Islām has shown itself to be very conservative and with a distrust of compromise and a passionate desire to proselytize.

The strength and limitations of a total surrender

East and West: common ground. In reviewing the different philosophical understandings of religions in both East and West, two points clearly emerge. First, that however great the variety, there is almost universal agreement that "what there is" is not restricted to the facts and features of the world as they are given to or received by man's senses. Secondly, what has been for the philosophy of religion in the East almost a permanent problem is coming to be a crucial problem for the West, viz., how to preserve both the concept of absolute spirit and the significance of personal individuality or, alternatively, how far one can speak reliably of God as a person. The West is becoming aware of the problematic character of religious discourse. If, in such ways, Western philosophy of religion can benefit from some of the insights of the East, so also can the East—as a growing interest in the Empiricists of the West demonstrates—gain from the West. Not least, scientific developments have created Eastern interest in the English Empiricists, particularly John Locke; Eastern philosophers also have been impressed by the political liberalism of some modern Western Empiricists, such as Bertrand Russell. The Empirical philosophy of religion, as it has been recently developed in the West, may provide basic approaches and techniques for a closer mutual study of religions in East and West.

BASIC THEMES AND PROBLEMS IN THE PHILOSOPHY OF RELIGION

The problem of God, the Absolute, or the supreme value. *The existence of God.* The so-called proofs of God's existence are of two kinds: independent logical exercises or particular conclusions set within an overall metaphysics. Either way, the discourse of these independent proofs or metaphysical schemes is best viewed as speech designed to evoke a disclosure. A particular argument recommends, as a way of speaking about what the disclosure discloses, a particular brand of discourse offering an interpretation of the world and man and one that develops from a specific key idea grounded in the disclosure. The existence of an Absolute or a supreme value has never been concluded as a result of an isolated logical exercise but has always arisen in the context of a total metaphysics. Thus, a quasi-mathematical structure, for Spinoza; a dialectic method, for Hegel; and evolutionary considerations, for the modern French philosopher Henri Bergson, determined the discourse that these three philosophers used in order to evoke that situation to which God or Nature, the Absolute Spirit, or the life force became for them respectively key concepts of interpretation. Bradley similarly reached a belief in an Absolute Spirit by reflecting on the logical problems of relatedness.

The kinds of proofs of God's existence

The following are some traditional arguments for the existence of God restyled along the lines suggested above:

The ontological argument of Anselm of Canterbury (*c.* 1033–1109) takes a phrase "that than which nothing greater can be conceived" and uses it as a technique for disclosure, directing one without limit to an ever-increasing perspective, in the hope that at some point the light will dawn, whereupon the phrase "necessary being" will be used to develop talk of the God.

The cosmological argument uses as a technique for disclosure such questions as "Why is this thus?" or "Why is there anything at all?" In receiving replies to these questions in causal terms, the cosmological argument builds up an ever-increasing causal spread until a disclosure occurs, whereupon the phrase "first cause" specifies what is disclosed and advocates certain ways of talking.

The argument from design takes a story with acknowledged disclosure possibilities—*e.g.*, the interrelated parts of a watch—and uses this as a catalyst to evoke a disclosure around some ever-broadening purpose patterns of the universe, in relation to which one can speak of God in terms, for example, of eternal purpose.

What is, in different ways, implied by these arguments is that the word God is unique in its logic, that it works in discourse as no other word exactly works. Thus, one cannot say "God exists" but rather "God necessarily exists." This is sometimes expressed by remarking that the existence of God is not the existence of a physical object or even the existence of a person, though what can be said about persons is less misleading in speaking about God than in speaking about the logic of things. This point is sometimes made, albeit misleadingly by saying that God does not exist, but this is only a picturesque way of saying that he does not exist in the way that a table exists.

The nature of God. These reflections are of wider applicability in relation to the nature and attributes of God. Such attributes are spoken of in terms of personal models, such as wisdom, goodness, power, love, mercy, righteousness, and so on. These models, however, will always need qualification by words such as infinite, perfect, and all. What is quite clear is that grammar itself is no clue to the logic of phrases such as "infinitely wise." Although that phrase is similar in grammar to one such as "exceedingly wise"—a phrase that is entirely descriptive in its logic—it is logically quite different, because "infinitely wise" has both descriptive and what has been called performative force. In other words, it not only describes some matters of fact—some specimens of wisdom—because of the word wise, which works descriptively as a model, but it also generates something—the word infinite acting as an operator, continually directing persons to expand their understanding until a moment of vision emerges. Alternatively, the point that God is not a being has sometimes been made by saying that God is the ground of Being—

Qualifications for personal models

"the ground of" functioning as a qualifier, operating as the model of beings, or things. The emphasis of such qualifiers is twofold. First, they remind one of the inadequacy of all language used to speak of God—language authorized by particular models that, arising in a moment of vision or disclosure, naturally originate speech about what the disclosure discloses. Secondly, qualifiers constantly point one back through developed discourse to that moment of vision in which the discourse originated and in which alone one knows what the discourse is speaking about. The logic of models and qualifiers is a way of combining the intelligibility and mystery that any philosophy of religion must preserve.

Language about God thus develops as a multiple discourse, having various strands of which each is authorized by a particular model and of which each must, somewhere along the line, be modified by the presence of the others. Thus, theological understanding is a complex interweaving of different strands, and not least is the task of the philosopher of religion to produce the most comprehensive, coherent, consistent, and simple discourse he can. When problems arise that seem to be problems about the nature of God—for example, the conflict between different attributes—these are most profitably translated into problems of language. They then become problems of how to create discourse of the kind that in the end produces the best understanding of a cosmic disclosure with a single individuation, in which all the pertinent discourse originates and about which all the different strands endeavour to speak.

Natural and revealed theology

The knowledge of God. Natural theology is the name given to the kind of discourse about God and the world that originates in natural moments of vision without reference to God's revelation of himself in an incarnation, and in this sense "natural theology" is distinguished from "revealed theology." Among some philosophers—*e.g.,* Locke—the distinction is one between general and special revelation. In natural theology are generally included the "proofs" of the existence of God, discussions about the immortality of the soul, and discussions about God's providential control of the world, which provides for man a state of moral probation.

Some have viewed religious experience as affording direct evidence for the existence of God. In any discussion of religious experience, however, it is important at the outset to distinguish religious experience in general—a sense of awe or reverence, or a sense of the numinous—from mystical experience. The language of mystics is notoriously confusing to those not accustomed to the mystical idiom, and a leading question is how far mystical experience can establish the kind of objective reference it claims. Words such as immediate, direct, and intuitive refer rather to the way in which the experience occurs as a disclosure rather than justifying one in taking as guaranteed the interpretation that this disclosure appears to bring with it. If one already has an interpretative scheme, then mystical experience may provide an instance of such a scheme, but this has been rightly described as supporting belief in God "on the way back" rather than "on the way out." The concept of revelation is used by Christians to describe the way in which God's activity is uniquely disclosed in Christ, and faith relates to the human attitude and response that matches revelation subjectively. Revelation is sometimes contrasted with discovery, the former being said to relate to a passive subject, the latter to an active subject, but the distinction is largely one of emphasis. Philosophers of religion are now inclined to view revelation in terms of activity that waits to be interpreted rather than as a revelation of propositions. Revelation thus relates to events rather than to doctrine. According to this view, doctrine could never have the ultimacy and finality that necessarily belongs to the givenness of God in his incarnation or incarnations.

Special problems. *Freedom.* Among the classical problems in the philosophy of religion are those of free will, self-identity, immortality, evil, and suffering. The freedom of the will is a claim for the uniqueness of the subject, known in occasions of activity in which the subject "comes alive" and realizes his subjectivity as that which cannot be reduced to the behaviour patterns and facts—*i.e.,* the objects—of the natural and social sciences. Such freedom is realized in responding to a situation that has equally come alive objectively to inspire a person and call forth such response. Some claim the predictable character of human behaviour rules out man's freedom; others state that the extent to which human behaviour is unpredictable argues for freedom. This controversy, however, does not in any way solve the problem of freedom; it only makes evident what kind of problem the problem of freedom is, viz., how far human nature is capable of being analyzed into behavioral terms without any residue.

Self-identity and immortality. When there has been a self-disclosure of transcendence, of what cannot be characterized in space and time, one cannot say that any self so disclosed entirely comes to an end. In this sense, there is an argument for personal immortality, though one can only talk sensibly about it by expressing immortality in terms of continuing personal life. In Christianity this becomes speech about the resurrection of the body, and in Hinduism it becomes speech about reincarnation in this world or in the universe at large. All detailed talk about a future life, whether in Christianity or other religions, is only a way of spelling out and pointing back to that experience of man's transcendence here and now, in terms of language that expresses the claim that such a transcendent element is not annihilated by death. To be articulate about immortality, emphasis is placed on features of life that, at first view, have high significance and point here and now to experiences in which man's self-disclosure is most often found—*e.g.,* inspiring music or the intimate and deep fellowship of a particularly significant meal. General claims for immortality in relation to an objective disclosure (whether it be spoken of in terms of God or moral ideals) have to be distinguished from, though they have evident similarities to, the Christian claim for eternal life. Eternal life is, according to the Christian view, a subjective self-disclosure alongside the objective disclosure of God's activity in Jesus Christ, and it is as unique as the uniqueness of God in Christ, a uniqueness that is, however, an inclusive, and not exclusive, uniqueness.

Arguments for personal immortality

Evil and suffering. The problem of evil arises (1) from the loss of a sense of God's presence in the face of evil or suffering and (2) from an apparent conflict between the language used to describe God (*e.g.,* all powerful, all good, and all wise) and that to describe the world as being characterized by evil and suffering. The solution proffered by the Book of Job in the Old Testament is that of evoking such a sense of awe around the created universe that, discovering in this way a renewed sense of God's presence, one accepts both evil and good and contents himself verbally by acknowledging a final incomprehensibility.

Other solutions relate good and evil to God and thus seek consistency by relating good and evil to God's primary and secondary will or to God's willing and permitting, respectively. In demanding some overall purpose to complete such a story, however, these solutions point to others that seek to resolve the conflict between good and evil within some reconciling model, which is then used to specify, with suitable qualification, a purpose or attribute of God. Thus, the conflict necessarily involved in the creation of a community of freely responsible persons is used as a model to illuminate a personal conflict exhibited, for example, by war. Also, the conflicts resulting from general rules imposed for the sake of training are used to provide a model to illuminate the disharmony exhibited in, for example, earthquakes or floods. These models are then developed and amplified in order to lead one to a renewed disclosure of God's presence. These solutions—by raising questions about God's character—perhaps point to another solution that attributes to God redeeming love—something that, as directed to evil, can be creative of personal maturity and fulfillment in a way not otherwise attainable. This attribute must then be appropriately qualified so as to lead to a renewed disclosure of God's presence, in this way enabling one both to face evil and to talk of it more coherently in relation to God.

In the matter of absolute Idealism, which is the kind of metaphysics implied in Eastern religions generally, evil

and good are transcended in the Absolute Spirit that is beyond good and evil. Logically, this is akin to the solution of the Book of Job.

THE PRESENT SITUATION IN THE PHILOSOPHY OF RELIGION

Interest in religious language

In the latter part of the 20th century in western Europe and the United States there has been an Empirical philosophy of religion, the interest of which has been in religious language and the kind of Empirical basis there can be for religious discourse. The definitive question has been concerned with what are the patterns of religious reasoning and what is the character of religious language if such discourse points back to and articulates situations of the particular kind that have been discussed above. The approach originated in what has been called Logical Positivism. According to the verification principle, which gave what the Positivists considered to be the touchstone of meaning, an assertion had meaning if and only if it was verifiable at least in principle by sense experience. Logical Positivists were not at all daunted when this seemed to exclude the whole of theology and a good deal of ethics from meaningful discourse.

Since the 1950s, however, there has been a reaction against the Positivist's veto, and the works of the Austrian British philosopher Ludwig Wittgenstein are symptomatic of those who broadened Empiricism so that it has become interested in displaying and elucidating the variegation of language, in setting language in actual contexts, and in relating it to specific situations. Significantly, this mellowing of Empiricism has been accompanied by a growing interest in personality and the self. This newer emphasis of Empiricism unites with Existentialism in suggesting that personal situations may very well provide helpful parallels to religious situations. There has been introduced into the philosophy of religion a renewed sense of the significance of mystery and a new emphasis on theological reticence. With this has come a renewed awareness of the significance of metaphor, myth, and symbol, and there has also emerged a significant use of the concept of the model. The use of terms like myth and mythological, it is important to recognize, does not mean that the assertions so called are false. Myth includes stories that try to articulate what is objectively given in a certain religious situation. Myths also relate to historical events—though the myth may be selective in its choice of such events—when speech about these events is used to articulate a claim of a transcendent kind. In other words, myth, metaphor, symbol, and model are all ways of expressing in ordinary language an extraordinary point.

Questions of reference and criteria

The present stress on metaphors and models in religious language, however, inevitably raises two far-reaching questions: the question of reference and the question of criteria. The former concerns the possibility of the assurance that one is talking about anything at all. The latter concerns what the criteria are for better and worse ways of talking. The question of criteria has been answered in terms of the logical character and the empirical pattern of the multimodel discourse to which the different strands arising from the different metaphors or models give rise. That this discourse talks about something must in the end rest on the claim that, in a disclosure situation, a subject is relatively passive—*i.e.* aware of an activity bearing on his own and thus aware of something other than himself about which he is talking.

In this context the significance of the Existentialist approach is to underline, as does recent Empiricism, the importance of a wider view of human experience than ordinary scientific experience might allow and to point one to highly significant personal situations that cannot be netted in scientific terms. The phenomenological approach, as developed by the Moravian philosopher Edmund Husserl, represents an attempt to be objective and scientific about experience, an endeavour to set out facts uncompromised and unprejudiced by metaphysical frameworks. As an endeavour to reach agreement on what is being talked about and as an attempt to seek the simplest and clearest interpretations, the phenomenological approach has been applauded by many philosophers of religion and theologians. There can be no question, however, about a purely scientific account of a religious situation—that would be a contradiction in terms, and, though there can be a phenomenological approach to religious situations, there can be no phenomenological explanation of them that claims to be adequate. The main contribution of Phenomenology is that of encouraging scholars to describe situations with as much critical analysis as possible.

Logical Empiricism, it might be said, has absorbed something of the Phenomenologists' concern. It has certainly raised questions about and directed interest toward the way situations are talked about and interpreted, the possibility of there being different interpretations of the same situation, and so on. In this way it has provided the tools for and greatly stimulated contemporary interest in hermeneutics (critical interpretations)—a second order appraisal of interpretations together with an interest in their empirical bases.

The future of the philosophy of religion

As to the future of the philosophy of religion, a merging of the Empirical and Existential strands may well be expected. Metaphysical and religious views of the future most likely will combine conviction with tolerance and commitment with openness. The commitment and the conviction will probably come from moments of vision. Claims to finality, fanaticism, and bigotry will disappear, it is hoped, when it is made obvious that no self-guaranteed translations of what disclosure is are given, and tolerance and openness will arise from the acknowledgment that all understanding of these moments of vision is a multiple exploration, an exploration yielding different strands of discourse. Solutions to contemporary problems, social and intellectual, demand a multiple consideration by scholars from many disciplines of all the issues involved in the problem, a consideration set within a framework of faith and morality in which man is interpreted as distinctively human, characteristically a person. From such interprofessional, interdisciplinary groups may emerge a new metaphysics and a new theology linked with, but by no means prescriptive of, assertions in other subjects. In this way there may be created a new culture—scientific, moral, religious, and technological at the same time. To be involved in such groups would seem to be the main task of the philosopher of religion, as of the metaphysician, today. If he is successful and if these interdisciplinary groups are creative, the modern period will then take its place among those that have marked crucial turning points in the history of mankind and its culture. (I.T.R.)

Philosophy of art

DISTINGUISHING CHARACTERISTICS

Difference between the philosophy of art and art criticism

The philosophy of art is distinguished from art criticism, which is concerned with the analysis and evaluation of particular works of art. Critical activity may be primarily historical, as when a lecture is given on the conventions of the Elizabethan theatre in order to explain some of the devices used in Shakespeare's plays. It may be primarily analytical, as when a certain passage of poetry is separated into its elements and its meaning or import explained in relation to other passages and other poems in the tradition. Or it may be primarily evaluative, as when reasons are given for saying that the work of art in question is good or bad, or better or worse than another one. Sometimes it is not a single work of art but an entire class of works in a certain style or genre (such as pastoral poems or Baroque music) that is being elucidated; and sometimes it is the art of an entire period (such as Romantic). But in every case, the aim of art criticism is to achieve an increased understanding or enjoyment of the work (or classes of works) of art, and its statements are designed to achieve this end.

The test of the success of art criticism with a given person is: has this essay or book of art criticism increased his understanding or enhanced his appreciation of the work of art in question? Art criticism is particularly helpful and often necessary for works of art that are more than usually difficult, so that the average person would be unable adequately to understand or enjoy them if left to himself.

The task of the philosopher of art is more fundamental than that of the art critic in that the critic's pronounce-

ments presuppose answers to the questions set by the philosopher of art. The critic says that a given work of music is expressive; but the philosopher of art asks what is meant by saying that a work of art is expressive and how one determines whether it is. In speaking and writing about art, the critic presupposes that he is dealing with clear concepts, the attainment of which is the task of the philosopher of art.

The task of the philosopher of art is not to heighten understanding and appreciation of works of art but to provide conceptual foundations for the critic by (1) examining the basic concepts underlying the critic's activities to enable him to speak and write more intelligibly about the arts, and by (2) arriving at true conclusions about art, aesthetic value, expression, and the other concepts that the critic employs.

The meaning of "art"

Upon what does the philosopher of art direct his attention? "Art," is the ready answer; but what is art and what distinguishes it from all other things? The theorists who have attempted to answer this question are many, and their answers differ greatly. But there is one feature that virtually all of them have in common: a work of art is a man-made thing, an artifact, as distinguished from an object in nature. A sunset may be beautiful, but it is not a work of art. A piece of driftwood may have aesthetic qualities, but it is not a work of art since it was not made by man. On the other hand, a piece of wood that has been carved to look like driftwood is not an object of nature but of art, even though the appearance of the two may be exactly the same. This distinction is being challenged in the 20th century by artists who declare that *objets trouvés* ("found objects") are works of art, since the artist's perception of them as such makes them so, even if the objects were not man-made and were not modified in any way (except by exhibition) from their natural state.

Nevertheless, according to the simplest and widest definition, art is anything that is man-made. Within the scope of this definition, not only paintings and sculptures but also buildings, furniture, automobiles, cities, and garbage dumps are all works of art: every change that human activity has wrought upon the face of nature is art, be it good or bad, beautiful or ugly, beneficial or destructive.

The ordinary usage of the term is clearly less wide. In daily life when works of art are spoken of, the intention is to denote a much narrower range of objects, namely, those responded to aesthetically. Among the things in this narrower range, a distinction, although not a precise one, is made between fine and useful art. Fine art consists of those works designed to produce an aesthetic response or that (regardless of design) function as objects of aesthetic appreciation (such as paintings, sculptures, poems, musical compositions)—those man-made things that are enjoyed for their own sake rather than as means to something else. Useful art has both an aesthetic and a utilitarian dimension: automobiles, glass tumblers, woven baskets, desk lamps, and a host of other handmade or manufactured objects have a primarily useful function and are made for that purpose, but they also have an aesthetic dimension: they can be enjoyed as objects of beauty, so much so that a person often buys one brand of car rather than another for aesthetic reasons even more than for mechanical reasons (of which he may know nothing). A borderline case is architecture: many buildings are useful objects the aesthetic function of which is marginal, and other buildings are primarily objects of beauty the utility of which is incidental or no longer existent (Greek temples were once places of worship, but today their value is entirely aesthetic). The test in practice is not how they were intended by their creators, but how they function in present-day experience. Many great works of painting and sculpture, for example, were created to glorify a deity and not, insofar as can be ascertained, for an aesthetic purpose (to be enjoyed simply in the contemplation of them for their own sake). It should be added, however, that many artists were undoubtedly concerned to satisfy their aesthetic capabilities in the creation of their work, since they were highly perfectionistic as artists; but in their time there was no such discipline as aesthetics in which they could articulate their goals; in any case, they chose to create "for the greater glory of God" by producing works that were also worthwhile to contemplate for their own sake.

This aesthetic sense of the word "art," whether applied to fine art or useful art, is the one most employed by the majority of critics and philosophers of art today. There are two other senses of "art," however, that are still narrower, and, to avoid confusion, their use should be noted: (1) Sometimes the term "art" is restricted to the visual arts alone or to some of the visual arts. Thus, the art department in 20th-century colleges and universities is usually the department of painting and sculpture (and sometimes architecture). But as philosophers of art use the term (and as it is used here), art is not limited to visual art; music and drama and poetry are as much arts as painting, sculpture, and architecture. (2) Sometimes the term "art" is used in a persuasive sense, to include only those works considered good art. "That's not art!" exclaims the viewer at an art gallery as he examines a painting he dislikes. But if the term "art" is to be used without confusion, it must be possible for there to be bad art as well as good art; the viewer, then, is not really denying that the work in question is art (it is a man-made object presented to be contemplated for its own sake) but only that it is worthwhile.

Other senses of the term "art"

The word "art" is also ambiguous in another way: it is sometimes used to designate the activity of creating a work of art, as in the slogan "Art is expression"; but it is more often used to designate the product of that process, the completed artwork or artifact itself, as in the remark "Art is a source of great enjoyment to me." There will be occasion later to remark on this ambiguity.

Countless proffered definitions of "art" are not definitions at all but theories about the nature of art that presuppose that the ability to identify certain things in the world as works of art already exists. Most of them are highly unsatisfactory even as theories. "Art is an exploration of reality through a sensuous presentation"—but in what way is it an exploration? Is it always concerned with reality (how is music concerned with reality, for example)? "Art is a re-creation of reality"—but is all art re-creation, even music? (It would seem likely that music is the creation of something, namely, a new set of tonal relationships, but not that it is the re-creation of anything at all.) "Art is an expression of feeling through a medium"—but is it always an expression (see below *Art as expression*) and is it always feeling that is expressed? And so on. It appears more certain that Shakespeare's *King Lear* is a work of art than that these theories are true. All that seems to be required for identifying something as a work of art in the wide sense is that it is not a natural object but something made or transformed by man; and all that is required for identifying it as art (not as good art but as art) in the narrower sense is that it functions aesthetically in man's experience, either wholly (fine art) or in part (useful art); it is not even necessary, as has been shown, that it be intended by its creator to function in this way.

THE INTERPRETATION OF ART

Works of art present problems of both interpretation and evaluation. Evaluation is not the concern of this article (see AESTHETICS); but one problem about interpretation deserves to be mentioned. Works of art are often difficult, and how to interpret them properly is far from obvious. The question then arises as to what factors should guide efforts at interpretation.

Isolationism and contextualism

At one extreme lies the view known as isolationism, according to which a knowledge of the artist's biography, historical background, and other factors is irrelevant to an appreciation of the work of art and usually is harmful in that it gets in the way, tending to substitute a recital of these facts for the more difficult attempt to come to grips with the work of art itself. If the work of art is not understood on first acquaintance, it should be read (or heard, or viewed) again and yet again. Constant re-exposure to it, so that the recipient is totally absorbed in and permeated by it, is the way to maximum appreciation.

At the other extreme, contextualism holds that the work of art should always be apprehended in its context or setting and that not merely knowledge about it but total appreciation of it is much richer if it is approached with

this knowledge. According to the contextualists, not only literature (ordinarily appreciated contextually) but also the other arts, even nonrepresentational painting and music, should be apprehended in this way.

No critic or art lover need hold to either position in its undiluted form: a person could well be an isolationist about some kinds of art, such as music, a contextualist about others, such as historical dramas and religious paintings. It is essential to be more specific, however, about the factors—other than careful and repeated perusal of the work of art itself—that the contextualist holds are either necessary or extremely helpful in the appreciation of works of art:

1. Other works of art by the same artist. If the artist has created other works, particularly in the same genre, acquaintance with them may enhance appreciation of the work at hand. Quantity of works has no particular merit in itself; but, when, say one of the piano concertos of the 18th-century Austrian composer Wolfgang Amadeus Mozart is heard, the auditor may (often largely unconsciously) compare its mode, thematic material, and method of development and resolution with some of Mozart's 25 other piano concertos. Knowledge of the entire corpus of his work in a certain genre may heighten enjoyment of a particular work.

2. Other works of art in the same genre by other artists, particularly in the same style or tradition. Appreciation of the pastoral poem "Lycidas," by the English poet John Milton, is doubtless enhanced by a study of the pastoral tradition in poetry, with which Milton supposed his readers to be acquainted. To study "Lycidas" in isolation would needlessly deprive the reader of much of the richness of texture of the poem and would even make some of the references in it unintelligible.

3. A study of relevant facts about the artistic medium, such as the instrumental limitations or advantages of pipe organs in the time of the German composer Johann Sebastian Bach (1685–1750) or the modes of presentation of ancient Greek tragedies in the Athenian theatre. An acquaintance with the artistic conventions and idioms in which the artist operated often leads to better understanding of certain aspects of his work and helps to avert misunderstandings of it.

4. A study of the age in which the artist lived—the spirit of the time and its current ideas, the complex influences that molded the artist, even the social, economic, and political conditions of the time and place in which he worked. Sometimes such knowledge is of dubious relevance: it can be argued that no aid to the study of the 82 string quartets and 104 symphonies of the 18th-century Austrian composer Joseph Haydn is provided by reading about the political and economic conditions of his day. It is interesting to study the evolution of the string quartet or symphony from its origin through Haydn to the present; but this would appear to be an evolution traceable entirely within the art form and not dependent on factors outside it. This, however, is not always so: particularly in literature, where a study of such exterior factors seems to be of much more relevance. It would seem important to know, for example, that Milton was aware of the new Copernican astronomy but deliberately chose in *Paradise Lost* to make his cosmos Ptolemaic, the antiquated astronomical system that was already steeped in literature, mythology, and tradition.

External influences on the artist

5. A study of the artist's life. Anthologists of literature constantly assume that this is an important consideration, since they supply detailed biographies prior to their selections by each author. It is true, of course, that knowledge of the artist's life can distract attention from his work, as with those who cannot hear Beethoven's late quartets without constantly thinking, "What a pity it was that he was deaf at the time!" Yet such knowledge may also heighten experience of a work; some would say, at any rate, that it helps to know that Milton was blind when he wrote the sonnet "On His Blindness." It is the relevance of this kind of knowledge to an appreciation of the poem, as a poem, that is in dispute. In every case, however, it should be kept in mind that acquaintance with the artist's biography is a means toward an end, the enhanced appreciation and understanding of the work of art, and that otherwise it is aesthetically irrelevant. The facts about the artist's life are the means and the enhanced appreciation the end, not the other way around, as is often found, for example in psychoanalytic essays attempting to infer facts about the artist's subconscious conflicts from his work; in these cases the work is being taken as the means and the study of his life as the end.

6. A study of the artist's intentions. It is this factor that has prompted the principal controversy in the mid-20th century. When difficulties arise as to what to make of a work of art or when several conflicting interpretations come to mind, how is the difficulty to be resolved? One obvious suggestion is to consult the artist or his records or memoirs or the testimony of people who knew him, to discover what his intentions were with regard to the work or the passage. It is tempting to believe that whichever way he intended it this is the way the work should be interpreted; for in regard to his own work, surely the author's own word should be law.

"Intentional fallacy"

This temptation is hotly decried by other critics as "the intentional fallacy"—the fallacy (if it is a fallacy) of believing that whichever way the artist intended it is by definition the way it really is. A work of art should stand on its own, without help from the artist; if he has not sufficiently realized his intentions within the work, forcing the recipient to go outside for help, this is held to be an artistic defect. Once the artist has completed his work, moreover, and presented it to the world, it belongs to the world and no longer exclusively to him: in the interpretation of it he now becomes just one critic among many, whose word should be respected but not taken as the final authority. Perhaps other critics can think of better interpretations than he did, which give a greater aesthetic reward in subsequent encounters with the work; perhaps there are even acceptable interpretations (such as the Freudian interpretations of Shakespeare's *Hamlet*) that he could not possibly have thought of himself at the time.

In the late 18th and early 19th centuries, Johann Wolfgang von Goethe set forth three criteria for critics to consider in interpreting and evaluating a work of art: (1) What was the artist trying to do? (2) Did he do it? (3) Was it worth doing? The first of the three is intentionalistic; and, says the intentionalist, surely this is plausible: an artist can hardly be blamed for failing to do what he had no intention of doing. It must first be known, then, what he was trying to do.

But the anti-intentionalist points out that the intention makes no difference, only the product does. If the ballerina excuses her fall in the middle of the dance by saying that she intended it, the dance is just as marred aesthetically as if she had fallen accidentally. And if a poet admits that he wrote rubbish and says that this is just what he intended to write, one does not rate the poem any higher because the poet's intention was fulfilled.

The persistent questioner might ask, however, if there are not at least some works of art in which the intentions of the artist have to be known? Suppose that a contemporary critic reads a dull, stodgy, moralistic Victorian novel and says at the end, "What an excellent parody of a Victorian novel!" But it was not a parody; its intentions were deadly serious—and should not this be known in order to interpret and evaluate it properly? Not at all, replies the anti-intentionalist; all the critic has to say is, "As a Victorian novel, this is deadly dull; as a parody of a Victorian novel, it is brilliant; if the author intended it in the former way, so much the worse for him—his work can still be praised for being brilliant parody, even if it wasn't intended as one. He just achieved something better than he knew at the time."

The author's statement of intention

Still, the intentionalist has a point: sometimes the clue to unlock an otherwise intransigent work may come from the author's statement of intention, and a plausible interpretation might be unobtainable without it. Such a suggestion might have come from a reader other than the author; but there is no point in disdaining helpful hints, regardless of their source. If the suggestion does come from the artist himself, that is nothing against it. Perhaps a work is less aesthetically perfect because it requires outside clues to its

interpretation, but few works of art even approach perfection, and they may yet amply repay attention, all the more if some plausible suggestion comes from the outside. The 20th-century Russian composer Sergey Prokofiev intended his *Classical Symphony* to be a parody of the classical symphonies of Mozart and Haydn—and regardless of whether the suggestion that it be construed this way came from Prokofiev (as it did) or from someone else, if it is rewarding to listen to it in this way, then no one gains by refusing to accept the suggestion. A statement of intention is not the only key to unlocking the secrets of works of art, but it is one key among many, and there appears to be no good reason why it should not be used.

THE MEDIUMS OF ART

In the context of every work of art there are three items to consider:

1. The genesis of the work of art.
2. The artifact, or work of art, which is a publicly available object or thing made by the artist and viewed by the audience.
3. The effects of the work of art upon the audience.

The first item comprises all the artist's mental states, both conscious and unconscious, in the creation of the work, including his intention with regard to the work, as well as all the factors that led to these states of mind: for example, the spirit of the age, the socio-economic conditions of the times, his exchange of ideas with other artists, and so forth. Whatever factors helped to form the work of art in the artist's mind fall under this heading. The experiences undergone by the artist in the creation of the work constitute the artistic experience.

The third item includes all the effects of the work of art upon those who experience it, including both aesthetic and nonaesthetic reactions, the influence of the work of art upon the culture, on the state of knowledge, on current morality, and the like. The experience that involves the observer's attention to the work of art for its own sake and not for the sake of some ulterior end is called aesthetic, but of course art has many effects that are not aesthetic. The aesthetic experience belongs to the consumer of art, as opposed to the artistic experience, which belongs to the creator of art.

The second item is what is usually called the work of art itself. According to some writers, such as the Italian philosopher Benedetto Croce (1866–1952), the work of art exists only in the mind of the artist, and the physical artifact then counts as an effect of the work of art. But in ordinary usage, as well as the usage of most philosophers of art,the work of art is identified with the physical artifact, as it exists in the physical medium. What goes on in the creator's mind is already contained in the first item.

Every work of art occurs in a medium; that is, there is some physical object or series of events by which the work is communicated to the recipient (listener, observer, reader) by means of his senses. In painting, the medium is paint; in sculpture, such materials as stone or wood or plastic. It might at first be thought that the medium of music consists of the musical score on which the composer writes the notes; but the written notes are not music, they are a set of visual cues for the production of the tones to be emitted by the various instruments. If every player had a perfect memory, he would have no need for the written score; indeed, music existed long before there were any written scores and was played or sung from memory from one year or generation to the next. It could be said more plausibly that the medium of music consists of the physical sound waves by means of which the sound sensations enter the consciousness of the listener. The medium of literature can truly be said to be words, yet not words as abstract entities conceived in the mind but words as spoken (in oral presentation) or written. The physical medium of literature, then, is either auditory or visual, although what is conveyed through the medium is not.

The medium of music

Classifying arts by their mediums. There are many ways of classifying the arts—by their purpose, by their intentions, by their effects. But the most usual and the most fundamental method of classifying the arts is by their mediums:

Visual art. This includes two-dimensional visual arts such as drawing and painting and also three-dimensional visual arts such as sculpture and architecture. Some of these should doubtless be called visuo-tactual art: buildings are ordinarily touched as well as seen, sculptures could be more fully appreciated if touched as well as seen, and even paintings may sometimes have enough three-dimensionality to repay touch experience. At any rate, all these arts appeal first and foremost, though not exclusively, to the sense of sight; and the artifact is an object in the visual medium.

Auditory art. This includes music in all its forms but not song, opera, and those arts that combine music with literature (see below). Just as the medium of visual art is sight, so the medium of auditory art is sound.

In auditory art there is—unlike visual art—no physical object (other than the score, which as has been seen is not the music). There is only the temporally successive series of sounds: sound waves emanating from the various instruments. While no such tones are being emitted, no sounds exist; only the musical score exists (and the memories of listeners, some of whom might enable the score to be reproduced if it were lost), from which music can be reproduced. Unlike the existence of paintings and sculptures, the existence of musical sounds is intermittent. In what sense, then, does the music exist between performances? It exists only in the sense that it is reproducible from the written score.

Verbal art. The art of literature is clearly different from both visual and auditory art. There are sound values in poetry, particularly when read aloud; but as sound alone, literature would be the most poverty-stricken of arts: what makes the sounds of poetry effective is (at least 99 percent) knowledge of the meanings of the words heard. Listening to the sounds of a poem or play uttered in an unfamiliar language gives some idea of the importance in literature of knowing the meanings of the words. Note that "murmuring," one of the most pleasant sounding words in English, has almost the same sounds as "murdering." It is almost exclusively a knowledge of word meanings that makes it possible to appreciate the art of literature.

The meanings of words

Nor is literature a visual art, although it is customary to read works of literature from a printed page. A critic who said, "I think this poem is a bad one, because it is written in unpleasant small type in double-column pages on yellowed paper," might be giving advice to typesetters and book designers (these two groups are engaged in the practice of visual arts), but he would be saying nothing about the merits of the poem. The printed or written word or for that matter the spoken word is only a vehicle for the meanings. Literature, then, must be placed in a separate class from either auditory or visual arts.

Mixed arts. Other arts variously combine the above three types of arts; this group includes all the arts of performance. Drama combines the art of literature (verbal art) with the visual arts of costuming, stage designing, and so on. Opera combines the art of music (its predominant component) with the art of literature (the libretto) and the visual arts of stage design. Dance combines the visual spectacle of moving bodies (the principal component) with musical accompaniment, sometimes with accompanying words and often with stage design. Song combines words with music. The motion picture combines the visual component (a series of pictures presented in such rapid succession that they appear to be moving) with the verbal component (the script) and usually an intermittent musical background as well.

All the visual arts are also spatial arts, or arts of space; music and literature are both temporal arts, or arts of time. This leads to very great differences in the things each can do. In temporal arts, the parts do not appear together before the audience but appear successively in time, the second moment not beginning until the first one has finished. In spatial arts, the entire work of art is present simultaneously; attention to the parts of it is successive—it is impossible to concentrate on the whole at once, at least on first viewing—but the entire object is nevertheless there, and it is up to the viewer which part he shall examine first. In three-dimensional art, such as

Spatial arts and temporal arts

sculpture and architecture, the entire object is present, but it is impossible even to see (much less to look at) all of it at once: the back of a statue cannot be seen at the same moment as the front and the exterior of a cathedral cannot be viewed by someone inside it.

Temporal arts must be attended to in a certain order: it is impossible to hear the symphony played backward, or the drama, or the movie; even when technically it can be done (as in running a motion picture in reverse), the results usually are an aesthetic catastrophe. The recipient is supposed to attend to the temporal work's various parts in an order predetermined by the artist. For this reason, painting is not capable of telling a story in the way that a novel is; for a story is a series of temporally successive happenings, and a painting can at best take a series of represented persons and objects and show them as they exist at one moment only, one knife-edge of time; whereas a novel can depict the temporally successive happenings in the order of their occurrence (or in a different order, such as flashback).

The German aesthetician and dramatist Gotthold Lessing made this distinction the basis of his study *Laokoon,* contending that the function of visual art is to create beautiful objects and that the artist should select that stance or moment at which the person or object appears most beautiful, to enable the viewer to continue looking at it with pleasure; whereas literature, being temporal, is equipped to tell a story that includes many moments other than pleasing ones (moreover, the scenes in literature are not seen with the eyes but only imagined). Lessing's thesis that each art should restrict itself to what it can do best or is peculiarly equipped by its medium to do is a highly controversial one: it would virtually eliminate program music, for example, and descriptions of nature in novels. The tendency of art today is to attempt to curtail distinctions between time and space rather than to preserve them.

Differences in the arts related to mediums. Very significant differences among the arts occur because of the differences in their mediums:

Literary and nonliterary. The greatest difference among the arts is between the literary and the nonliterary. Literature consists of a system of symbols with assigned meanings. A word is not simply a noise (or a mark on a printed page); a word is a noise or a printed mark with an assigned meaning. In different languages, different noises have been assigned meanings, and the language must be learned in order to understand what is being said. To appreciate the work of the 11th-century novelist Lady Murasaki Shikibu, one must learn Japanese; to fully appreciate Molière, the 17th-century playwright, one must learn French. No other art has this problem: the Englishman can appreciate German music as well as a German—or if he does not, it is not for lack of learning a language.

Assigned meanings of words

Shapes, colours, and tones do not have assigned meanings. That is not to say that these elements when present in art cannot be said to have some sense of the term "meaning." There are many meanings of the word "meaning," and a colour, for example, can have meaning in that it may symbolize something, as red symbolizes courage; or it may have strong emotional or other effects upon the viewer, evoking all manner of strong associations. But a colour or a tone has no assigned meaning: if the question were asked, "What does middle C mean?", the answer would be, "It has no assigned meaning at all; in that sense, it means nothing—it just has certain effects." But if the meaning of a certain word in a poem is not known, the reader is to that extent prevented from appreciating the poem; for the medium of poetry is not noises, not printed marks, but words, and the difference between a noise and a word lies in the fact that a word is a noise with a meaning.

This fact makes for an enormous difference between literature and the other arts. A colour in a painting may be the colour of an object represented; a colour may even "mean" something; for example, the white of a white flag that a person depicted in a battle painting is holding up as a sign of surrender. But a colour, as a colour, has no meaning at all; and the same is true of musical tones. A pattern of musical tones may occasionally acquire a meaning (the first four tones of Beethoven's *Fifth Symphony* were used to symbolize victory in World War II), but when this happens it really has very little to do with the music, and in any case most music is appreciated without any such symbolism being present. But a noise, however pleasing to the ear, is only a noise and not a word unless it has an assigned meaning; and one must know what that meaning is in order to appreciate a poem or any other work of literature.

The translation problem. Because literature consists of conventional symbols, there exists in literature the problem of translation, which does not exist in the other arts. When one seeks to make a work of literature available to a wider audience than that composed of only the native speakers of the language in which the work was written, the process of translation must be resorted to, and, in this process, a great deal of the work's original character is lost.

Difficulties in translating poetry

In a poem there are (1) the sounds, (2) the dictionary meanings of the words, and (3) the connotations of the words—the manifold associations that they evoke (sensory, intellectual, and emotional) in the minds of readers. The sounds are the least important of the three, and many a great poem as sheer sound is hardly even pleasing. The finding of like dictionary meanings is usually a simple matter, and when there is a word that has no rough equivalent in the other language, it may be simply retained in the original language (for example, the German word *Weltanschauung,* meaning something like "world outlook," is often retained in English translations of German works). As for the associations that hover about a word, they may vary from language to language, so that if a work is translated rather literally, the associative values of the words are lost. Thus, "My God!" is a much stronger expletive in English than "mon dieu" is in French, so that if the French expression is translated into the English one, it is, though literally correct, quite unfaithful to the weaker emotive force of the French expression. Words can often be found in the second language that have a roughly equivalent associative value to the original one, but these will usually not provide a literal translation; thus, the translator is faced with the dilemma of being able to provide a literal-meaning translation or a translation that renders the spirit or "feel" of the original but not both.

The question of correspondence to actuality. The arts also differ from one another, according to their mediums, in whether the items in the medium correspond to items in the world. Objects with colours and shapes are represented on canvas, and objects with colours and shapes also exist in the outside world. Even when a painting is nonrepresentational, it consists of colours and shapes, which are items in the outside world (even though certain individual colours and shapes in the painting may not exist in the outside world). But the case with music is different: though the visual arts may (to varying degrees) convey the sights of nature, music does not convey the sounds of nature. Even when a work of music attempts to represent the sound of an iron foundry or the clattering of horses' hooves, it really does not sound like these things: musical instruments emit tones, and in nature are found largely noises, and between the two there is an enormous auditory difference. Some rhythms of nature can be duplicated by musical instruments but hardly the sounds themselves.

The medium of literature, words, is indeed man created; but of course this feature is far from unique to literature. Words were devised and employed in countless situations of daily life before they were ever embodied in literature; so in literature, as in visual art, a medium is being employed that existed before the art itself.

ART AS IMITATION (REPRESENTATION)

The view that "art is imitation" is at least as old as the Greek philosopher Plato (428–347 BC), and, although not widely held today, its long and distinguished history is evidence of its continuing hold on human beings as an account of the distinctive function of art. A terminological point, however, is in order here: in the interests of clarity, an artist should be spoken of as representing in his work the persons and things and scenes of the world but as imitating the work of other artists. Thus, "In this

painting he represents a barn and some wheat fields, and in his style he is imitative of Vincent van Gogh." This distinction will be employed here, with the result that these traditional theories of art will be spoken of as theories of representation rather than of imitation.

At some period in the history of art, aestheticians and critics wrote as if nature should be recorded by the artist with photographic fidelity. The invention of photography (which can do this better than any painter) could plausibly be said to have relieved the artist of any such responsibility. Still, art can represent reality: the representation of a house in a painting may not look exactly like a house—it cannot, since the real house is three-dimensional and the painting is two-dimensional—but it looks enough like one to enable everyone unhesitatingly to identify it as a house.

Distinction between depiction and portrayal

A distinction should be made between depiction and portrayal. A painting may be said to depict a house if it looks more like a house than like anything else. Thus, most persons unhesitatingly classify this as a man, that as a tree, and so on; only when the painter has distorted or abstracted so much that a thing looks somewhat like a wolf and also somewhat like a bobcat, do they hesitate in saying what the object represented is. A picture may depict a rather short man in a French general's uniform of the early 19th century; but it may, in addition, portray Napoleon. It portrays Napoleon if (1) the artist intended it to represent Napoleon (for example, if the title of the painting is "Napoleon") and (2) the painting does look like Napoleon to some degree at least—at any rate it contains no important characteristics known to be incompatible with those of Napoleon. Clearly, if it is a painting that depicts a tree in someone's yard, it cannot be considered a portrait of Napoleon, no matter how much the artist said he intended it to be one. Depiction subjects can ordinarily be recognized at once with a little knowledge of the world and the names of the things in it. Portrayal subjects require knowledge of whomever the artist intended to portray; even when that seems obvious, as in the case of Napoleon (who would be instantly recognized, unlike the portrait of a private in his army), the viewer would have to be told, by the title or otherwise, that not only does the painting depict a man in a French general's uniform but that it was intended by the artist to be a portrait of this particular man. Otherwise, how would the viewer know that it did not actually portray his double, or his stand-in? The word represent, as used in connection with art, can mean either "depict" or "portray."

Analysis of representation. Representation always involves a certain degree of abstraction; that is, the taking away of one characteristic or more of the original. Even a fairly realistic painting of a person, for example, lacks some features that characterize actual persons: the painting is two-dimensional, whereas every actual person is three-dimensional; the surface of a painting is paint, but not so the person; the actual person has very numerous pores and other marks on his face that are lacking (in whole or in part) in the painting, and so on. The depiction of a person in a painting is usually sufficient to enable human viewers to recognize the figure as a person—though it is apparently not sufficient for an animal, who sees only a coloured canvas where people see on the coloured canvas a representation. When the degree of abstraction is so great that it is no longer possible to recognize this shape as a human shape or as the shape of any identifiable object, the painting is then spoken of as non-representational. (In popular parlance such paintings are called abstract; but this is misleading, for abstraction is a matter of degree, and, as has just been shown, all depictions are necessarily abstract—that is, abstracted from reality to some degree.) The actual object with all its millions of qualities is at one end of the spectrum, and the painting so abstracted that a depiction subject is unrecognizable is at the other end; between the two extremes lie all the possible degrees of abstraction.

Representation in literature

Literature can be representational but not in the same way as visual art. It is quite natural to say that in a novel or drama a number of characters and actions are represented. The representation is, of course, not a visual one; it is representation through language. The painter portrays Napoleon by making a portrait of him; the writer does so by describing him in words. The writer, unlike the painter, can also depict action. Not all literature, of course, is representational in this way: a sonnet may contain no characters at all and no action, consisting solely of an expression of feeling by some unspecified speaker.

Any of the mixed arts that include words as part of their medium, such as drama or film, can be, like literature, representational. Indeed, they have a further advantage: they can depict action not only through words but also by showing the characters and exhibiting the action before the spectator. These arts are visual as well as verbal, and since they are not limited to one moment in time, as painting and sculpture are, they are temporal arts as well as spatial. These mixed arts, then, can be doubly representational.

Is it possible for music, too, to be representational? Music cannot visually show characters or objects, nor can it describe them in words; can it "depict them in tones"? Program notes at concerts usually assume without question that it can. The audience is told about the tone poem *Don Quixote,* by the Austrian composer Richard Strauss, "The composer has given us a musical representation of the Don's adventures. The 17th-century Spanish writer Miguel de Cervantes has described them in words, and Strauss has done so in tones." But the claim to representation in music is, to say the least, quite dubious. Without the title, with the music alone, would there be any clue that the music was supposed to be "about" the adventures of Don Quixote? True, there is a passage that resembles the bleating of sheep sufficiently for that much to be guessed; but even to conjecture that this passage is a representation of sheep bleating is a far cry from being able to reconstruct the entire story. Suppose that Strauss had left every note in the score just as it was but changed the title; would the piece then have been a representation of something else? The very fact that this question can be asked shows how different music is from visual art: if a painter has drawn a house but indicated in the title that it was supposed to be a tree, the viewer could still say, on the basis of what he saw in the picture, that it was not a tree but a house. But in music the listener is never in this situation: if he says that this series of tones represents the adventures of Don Quixote, he says this because of the title Strauss used. If the composer had given it no programmatic title, one listener might think of one represented subject, another a different one, and a third none at all, and there would be no way of showing who was right or even whose opinion was to be preferred. The conclusion seems to be that music by itself—without title, without words, without depicted action (as in a combination of music and drama such as opera)—is incapable of representing anything. There is simply a series of musical tones that may suggest differing associations, programmatic or otherwise; but the musical tones by themselves cannot be said to represent anything at all.

This might be objected to as an overstatement. If a picture can represent a house by looking more like a house than anything else, cannot a work of music represent the sea by sounding more like the sea than any alternative? And is this not the case in, for example, the French composer Claude Debussy's tone poem *La Mer?* Even this, however, is highly questionable; almost no one guesses the title to Debussy's tone poem without first knowing what it is; it may seem obvious enough after the composer has channelled the listener's response by means of his title but not beforehand. And surely this is because the sounds in the tone poem do not sound more like the sea than like anything else: the tone poem consists, after all, of a series of complex musical tones, emitted by violins, cellos, clarinets, flutes, trumpets, and so on; and it would be difficult indeed for these sounds, which are musical tones, to sound very much like the sea, whose sounds consist after all of a series of complex noises. There is no great similarity between any one series of musical tones and any one series of nature's noises. Hence, the first cannot be said to constitute a representation of the second.

The matter is even more obvious in the case of those

Musical attempts to represent the soundless

numerous programmatic titles in which the supposedly represented subject contains no sounds at all. Debussy's *Reflets dans l'eau* (*Reflections in the Water*) is taken by some as a musical representation of reflections in the water. But reflections in the water emit no sounds at all, not even noises. No one, then, could say that the sounds in Debussy's piano composition resemble the sounds of reflections in water. The resemblance, if there is any, is much more remote: it may be that the feeling obtained when Debussy's composition is heard is somewhat like the feeling that arises when reflections in water are seen. This is highly improbable without knowledge of the title, but at most it would provide a mood resemblance, which is far removed from a representation by music of things in the world. The conclusion seems inescapable that music is not to be classified as a representational art, at least not in the same straightforward meaning of "representation" that applies to the other arts.

So much, then, for the capacities of the various arts as far as representation is concerned. But the question remains: in those arts that are properly called representational, what should be the nature of the representation?

That art should be an outright duplication (incorrectly called "imitation") of reality is a view that was put forward by the French novelist Émile Zola in his book *Le Roman expérimental* (*The Experimental Novel*) and has been occasionally held (though not practiced) by painters reacting against Romanticism, such as the 19th-century French artist Gustave Courbet. Zola advocated a novel that resembled a scientific investigation into reality. Plot was to be of no importance, rather an aspect of reality was to be examined searchingly, and from this the story would unfold without imaginative effort. Persons or groups of persons would be depicted, and from them the action would evolve.

Objections to Zola's theory of duplicating reality

It would be impossible, of course, to carry out such an ideal of art as "report" and undesirable even if it were possible. First, the author or painter must select a subject; and, within the subject, he must select which details to treat, for he cannot in a hundred lifetimes describe them all: since every object and event has an indefinitely large array of qualities, there is no point at which a description of it would be completed. Besides, the very language used (no matter how neutral a description is attempted) will colour the account. Even if the words were colourless, the mode of putting them together would yield a style, which would colour the account once again. Indeed, should such an ideal be achieved (as in the verbatim transcription of an actual trial) it would be the deadliest possible bore.

Art, even representational art, is not a reproduction of reality; it is a transformation of reality. How, specifically, is reality transformed in being represented in art? There is probably no general satisfactory answer to this question. Each art, each style of art, and each work of art transforms reality in its own way—the 19th-century French painters Paul Cézanne in one way, Pierre-Auguste Renoir in another; the 19th-century Russian writers Fyodor Dostoyevsky in one way, Leo Tolstoy in another. No set of rules can lead to predictions as to what transformation of reality will be conceived in the mind of the next creative artist. Reality is the common base, but each artist deals with it in his own unique way.

Subject matter. Do all works of art have a subject matter? The answer to this depends on what is meant by the term subject matter, which signifies basically what the work is about. There are several senses of being "about" that may be referred to:

1. "What is the subject matter of the *Odyssey* by the ancient Greek poet Homer?" The most natural answer would be: "The wanderings of Odysseus." This is the "representational content" of the work. A person who read it simply for the story could easily give this answer. There is contained in the work itself an account of the wanderings of a character named Odysseus, who has no counterpart in the outside world (that is, he is a fictional character) but who does resemble people in the outside world in that he is a man, he is away from home, he is beset by many vicissitudes, and so on. If the subject matter were stated in greater detail, the result would be an account of the plot.

Does painting have subject matter in this sense? If "subject matter" is taken to mean representational content, the answer is often yes; representations of people and trees and the like are easily identifiable in the painting. Sometimes the subject could not be ascertained except for the title, which is not strictly a part of the painting (the title is not a piece of visual art, consisting as it does of words). But quite clearly it cannot be said that the subject matter of the painting is whatever the title says it is: if the title says "Clouds" and the painting is obviously a still life with pears and grapes, it can hardly be said that the painting has clouds as its subject matter simply because the title says so. It was long said that the painting called "Sacred and Profane Love" by the 16th-century Italian artist Titian had sacred and profane love as its subject matter; but Titian himself gave it no such title—the title was added more than a century later by someone else. It cannot be said that the subject matter is what whatever title it came to be given says it is (it could be given numerous incompatible titles); the subject matter, in this sense, must be evident from—or at least not incompatible with—what is seen in the painting itself. The same remarks apply to music: music cannot be "about" anything that the composer happens to seize on as a title; if it were, and he changed the title without changing a note of the music, would the subject matter of the music have changed? Strong reasons already have been given to doubt that music can have a subject matter at all.

2. In addition to a subject matter in the sense of representational content, a work of art may have subject matter in the sense of theme, or underlying idea. The subject matter in the sense of representational content of *Jude the Obscure*, by English novelist and poet Thomas Hardy, is the intellectual ambitions of the main character, Jude, and the vicissitudes befalling him along the way; but the subject matter in the sense of theme or underlying idea is man's struggle to realize his ambitions; and the thesis (implicit message) of the novel is (or may be) that man's highest ambitions are doomed to frustration. Not all novels and dramas and poems have a theme or thesis; they may simply tell a story and nothing more. Works of literature do sometimes operate in both dimensions: a story and a theme or thesis underlying the story. Works of representational visual art may also do so: one might allege that the theme of "Guernica," by the 20th-century Spanish artist Pablo Picasso, is the horror of war. Again, it is doubtful whether works of music can be said to have a theme at all (indeed, the term theme applied to music has a very different meaning: it refers to a set of tones around which variations or developments are composed). If someone were to say that the theme of a work of music was the human condition or man's fear of death, how would such a view be supported? Since music is not representational, what musical passages would he take as evidence for his theory, and why? A composer might indeed title his work "The Human Condition," but this would no more show that the music was about the human condition (had the human condition as its subject matter) than giving the programmatic title "Clouds" to a musical composition shows that the composition actually is about clouds.

Thematic subjects

Symbols in art. Works of art may not only have subject matter, they may also contain symbols. Certain elements in a work of art may represent, say, a whale; but the whale thus represented may be (as it is in *Moby Dick* by the 19th-century U.S. writer Herman Melville) a symbol of evil. In Tolstoy's *Anna Karenina* is represented a gallery of characters dominated by Anna herself, and a tremendous number of actions in which these characters engage; but there is a constantly recurring item in the representational content, namely, the train. Time and again the train causes or accompanies frustration, disaster, betrayal, and other evils—so much so that before the novel is ended it becomes apparent that the train here is a symbol of the iron forces of material progress toward which Tolstoy had such great antipathy.

What is it that makes an item in a work of art a symbol? It is something represented in the work of art—an object, an action, or a pattern of objects and actions, or even (less frequently) simply a nonrepresentational item such

as a colour or a line—that does the symbolizing; what is symbolized is a characteristic, such as evil or progress or courage. But by virtue of what does the first (A) become a symbol of the second (B)?

Conventional and natural symbols

The answer is not the same for all symbols, since some are conventional and some are natural. The cross is a symbol of Christianity, and it is a conventional symbol of suffering; in order for it to become a symbol, people had to adopt or accept the cross as standing for suffering. Other symbols are natural—the sun as the symbol of life and strength, a river as the symbol of eternal change and flowing, and so forth; in these cases there was no agreement (convention) as to what would stand for what, for the relation is too obvious—the symbolism is much the same in the tradition of all nations and civilizations.

Various symbols have, to varying degrees, elements of both the conventional and the natural: the eagle on the standard of the United States of America symbolizes strength—this is natural, because the eagle is strong, and conventional, because the eagle was officially adopted as the symbol of the United States. In the case of many symbols, the natural relation between symbol and thing symbolized is not strong enough by itself to achieve the symbolism, and the conventional relation was entered into in order to effect a symbolism that already had a certain natural basis. When A is a symbol of B, there must be some relation, either natural or conventional, between the symbol and what is symbolized. When the natural relation is strong, the conventional is minimal or nonexistent, and vice versa.

But there is another element to symbolism if it is to function with full-bodied effectiveness in a work of art, and that is what has been called a "vital basis" provided by history and tradition. In the U.S. flag there is very little natural basis for symbolism; it is almost wholly conventional. But the fact that it has been saluted for so long, that it has been present on battlefields, that the bodies of military dead have been wrapped in it, and so on—that is, the fact that it has acquired a life, a history—have given it an "affect" (human emotion or mood) it did not have before. Melville and Tolstoy not only represented the whale and the train: they invested these representations with such strong emotional affect that the reader is much more inclined than he would be otherwise to say that these objects are the symbols of the qualities he ascribes to them.

When some item in a work of art is construed as a symbol, it is always infused with these vital qualities. Clearly, not every case of one thing standing for something else is a symbol in this sense: a carrot in a painting does not, just by itself, symbolize growth. But barbed-wire fences do symbolize tyranny, not only because many prisoners have been enclosed in barbed-wire fences (this is the natural basis for the symbolism) but also because of concentration camps and the countless tragic events of recent history that have provided the "vital basis" for the symbolism.

Meaning. Do works of art have meaning? The answer depends once again on how the question is construed: the word "meaning" is an equivocal term that can itself mean many different things.

Clouds mean rain, a falling barometer means that a storm is coming, a twister in the sky means an approaching tornado—that is, the one is a sign of the other; these relations exist in nature and were discovered, not invented, by man. On the other hand, a bell ringing means the end of class, this note on the score means that D sharp is to be played on a certain instrument, and the word cat to someone who knows English means a certain species of domesticated quadruped; these relations are conventional, established by man. But both the natural and the conventional items are examples of meaning in its most general sense—one thing (A) standing for another (B).

Meaning in literature

Since the medium of literature is words, and words are conventional vehicles of meaning, literature has meaning in a way that the other arts do not, since every word, to be a word at all, must have a meaning. In the sense in which the word cat means something, middle C and an ellipse do not have meanings at all.

When the question is asked about meaning in art, however, it is not usually the individual ingredients in it that are being referred to: if it were, the answer could simply be, "Yes, this word has a meaning and that word has; so does the sentence as a whole; and items in paintings sometimes have meaning; for example, the halo over the Madonna's head symbolizes holiness." What is being asked is whether the work of art as a whole has a meaning. But what does the question itself mean? Several different things can be meant: (1) The inquirer may be asking, "What is it about?" in which case the question is about subject matter, already discussed. (2) He may be asking, "What is its theme?"—for example, is the motion picture *He Who Must Die* really a parable about the life of Christ? (3) He may be asking, "What is its thesis?" For example, what is the message of the Anglo-Irish author Jonathan Swift to the reader in "A Modest Proposal"? (4) The inquiry may be about the effects of a work of art on the recipient—either what these effects are or what they could or should be. In this sense, all works of art have meaning, since they all have effects (whether there is one type of effect that a given work of art should have is another question). This is, however, an extremely misleading use of the word "meaning." Indeed, the entire discussion of "meaning in art" is a most confusing one—and the fault does not lie in art, but in the human users of words. Endless unnecessary mysteries can be created by using such nebulous words as "meaning" as if they were simple, straightforward, and susceptible to one interpretation. It would contribute greatly to the clarification of discussions of philosophy of art if the word meaning were not used in them at all but some conception clearer and more specific.

ART AS EXPRESSION

The view that "art is imitation (representation)" has not only been challenged, it has been moribund in at least some of the arts for more than a century. It has largely been replaced by the theory that art is expression. Instead of reflecting states of the external world, art is held to reflect the inner state of the artist. This, at least, seems to be implicit in the core meaning of "expression": the outer manifestation of an inner state. Art as a representation of outer existence (admittedly "seen through a temperament") has been replaced by art as an expression of man's inner life.

The terms "express" and "expression"

But the terms "express" and "expression" are ambiguous and do not always denote the same thing. Like so many other terms, "express" is subject to the process–product ambiguity: the same word is used for a process and for the product that results from that process. "The music expresses feeling" may mean that the composer expressed his feeling in writing the music or that the music when heard is expressive (in some way yet to be defined) of human feeling. Based on the first sense are theories about the creation of art; founded on the second are theories about the content of art and the completion of its creation.

Expression in the creation of art. The creation of a work of art is the bringing about of a new combination of elements in the medium (tones in music, words in literature, paints on canvas, and so on). The elements existed beforehand but not in the same combination; creation is the re-formation of these pre-existing materials. Pre-existence of materials holds true of creation quite apart from art: in the creation of a scientific theory or the creation of a disturbance. It applies even to creation in most theologies, except some versions of Christian theology, in which creation is *ex nihilo;* that is, without pre-existing matter.

That creation occurs in various art mediums is an obvious truth. But once this is granted, nothing has yet been said about expression; and the expressionist would say that the foregoing statement about creation is too mild to cover what he wants to say about the process of artistic creation. The creative process, he wants to say, is (or is also) an expressive process; and for expression something more is necessary than that the artist be creating something. Great care must be taken at this stage: some say that the creation of art is (or involves) self-expression; others say that it is the expression of feeling, though not necessarily of one's own feeling (or perhaps that and something more, such as the feeling of one's race, or of one's nation, or of all men);

others say that it is not necessarily limited to feelings, but that ideas or thoughts can be expressed, as they clearly are in essays. But the distinctively expressionist view of artistic creation is the product of the Romantic movement, according to which the expression of feelings constitutes the creation of art, just as philosophy and other disciplines are the expression of ideas. It is, at any rate, the theory of art as the expression of feelings (which here shall be taken to include emotions and attitudes) that has been historically significant and developed: art as specially connected with the life of feeling.

When a person is said to be expressing a feeling, what specifically is he doing? In a perfectly ordinary sense, expressing is "letting go" or "letting off steam": man expresses his anger by throwing things or by cursing or by striking the person who has angered him. But, as many writers have pointed out, this kind of "expressing" has little to do with art; as the U.S. philosopher John Dewey said, it is more of a "spilling over" or a "spewing forth" than expression. In art at least, expression requires a medium, a medium that is recalcitrant and that the artist must bend to his will. In throwing things to express anger, there is no medium—or, if the man's body is called the medium, then it is something he does not have to study to use for that purpose. It is still necessary to distinguish a "natural release" from an expression. If poetry were literally "the spontaneous overflow of powerful feelings," as William Wordsworth said, it would consist largely of things like tears and incoherent babblings. If artistic creation can plausibly be said to be a process of expression, something different from and more specific than natural release or discharge must be meant.

Emotional expression in art

One view of emotional expression in art is that it is preceded by a perturbation or excitement from a vague cause about which the artist is uncertain and therefore anxious. He then proceeds to express his feelings and ideas in words or paint or stone or the like, clarifying them and achieving a release of tension. The point of this theory seems to be that the artist, having been perturbed at the inarticulateness of his "ideas," now feels relieved because he has "expressed what he wanted to express." This phenomenon, indeed a familiar one (for everyone has felt relieved when a job is done), must still be examined for its relevance. Is it the emotion being expressed that counts or the relief at having expressed it? If the concern here is with art as therapy or doing art to provide revelations for a psychiatrist, then the latter is what counts; but the critic or consumer of the art is surely not concerned with such details of the artist's biography. This is an objection to all accounts of expression as process: how is any light at all cast upon the work of art by saying that the artist went through any expressive process or through any process whatever in the genesis of it? If the artist was relieved at the end of it, so much the better for him; but this fact is as aesthetically irrelevant as it would be if he had committed suicide at the end of it or taken to drink or composed another work immediately thereafter.

Another problem should be noted: assuming that the artist does relieve his oppressed state of mind through creating, what connection has this with the exact words or score or brushstrokes that he puts on paper or canvas? Feelings are one thing, words and visual shapes and tones are quite another; it is these latter that constitute the art medium, and in it that works of art are created. There is doubtless a causal connection between the feelings of the artist and the words he writes in his poem; but the expression theory of creation talks only about the artist's feelings, while creation occurs within the art mediums themselves; and to speak only of the former is not to tell anything about the work of art—anything, that is, that would be of interest other than to the artist's psychiatrist or biographer. Through what paroxysms of emotion the artist passed does not matter anymore, insofar as one's insight into his work is concerned, than knowing that a given engineer had had a quarrel with his wife the night before he began building his bridge. To speak of anything revelatory of works of art, it is necessary to stop talking about the artist's emotions and talk about the genesis of words, tones, and so on—items in the specific art mediums.

Distinction between art and craft

The expressionists have indeed brought out and emphasized one important distinction: between the processes involved in art and in craft. The activity of building a bridge from the architect's blueprint or constructing a brick wall or putting together a table just like a thousand others the artisan has already made is a craft and not an art. The craftsman knows at the beginning of the processes exactly what sort of end product is wanted: for example, a chair of specific dimensions made of particular materials. He knows at the beginning how much material it will take to do the job, which tools, and so forth, and, if he does not know these things, he is not a good (efficient) craftsman. But the creative artist cannot work in this manner: "The artist doesn't know what he is going to express until he has expressed it" is a watchword of the expressionist. He cannot state in advance what the completed work of art will be like: the poet cannot say what words will constitute the completed poem or how many times the word "the" will occur in it or what the order of the words will be—when he knows that, he has completed the creation of the poem, and until then he cannot say. Nor could he set about working with such a plan: "I shall compose a poem that contains the word 'the' 563 times, the word 'rose' 47 times," and so on. What distinguishes art from craft is that the artist, unlike the craftsman, "does not know the end in the beginning"; he cannot state until his work is finished what the completed product will be like—if he could, he would not have to undergo the "divine agonies" of creation in order to produce it.

The distinction seems valid enough, but whether it supports the expressionist's view is more dubious; for it can be held regardless of the attitude assumed toward the theory of expression. The open-ended process described as art rather than craft characterizes all kinds of creation: of mathematical hypotheses and of scientific theory, as well as art. What distinguishes creation from all other things is that it results in a new combination of elements, and it is not known in advance what this combination will be. Thus, one may speak of creating a work of sculpture or creating a new theory, but rarely of creating a bridge (unless the builder was also the architect who designed it, and then it is to the genesis of the idea for the bridge, not to its execution, that the word creation applies). This, then, is a feature of creation; it is not clear that it is a feature of expression (whatever is being done in expressing that is not already being done in creating). Is it necessary to talk about expression, as opposed to creation, to bring out the distinction between art and craft?

There does not seem to be any true generalization about the creative processes of all artists nor even of great artists. Some follow their "intuitions," letting their artistic work grow "as the spirit moves" and being comparatively passive in the process (that is, the conscious mind is passive, and the unconscious takes over). Others are consciously active, knowing very much what they want in advance and figuring out exactly how to do it (for example, the 19th-century U.S. writer Edgar Allan Poe in his essay "The Philosophy of Composition"). Some artists go through extended agonies of creation (the 19th-century German composer Johannes Brahms, weeping and groaning to give birth to one of his symphonies), whereas for others it seems to be comparatively easy (Mozart, who could write an entire overture in one evening for the next day's performance). Some artists create only while having physical contact with the medium (for example, composers who must compose at the piano, painters who must "play about" in the medium in order to get painterly ideas), and others prefer to create in their minds only (Mozart, it is said, visualized every note in his mind before he wrote the score). There appears to be no true generalization that can be made about the process of artistic creation—certainly not that it is always a process of expression. For the appreciation of the work of art, no such uniformity, of course, is necessary, greatly though it may be desired by theorists of artistic creation.

Problems in drawing conclusions about the creation of art

The main difficulties in the way of accepting conclusions about the creative process in art are (1) that artists differ so much from one another in their creative processes that no generalizations can be arrived at that are both true and interesting or of any significance; and (2) that in the present stage of psychology and neurology very little is known about the creative process—it is surely the most staggeringly complex of all the mental processes in man, and even his simpler mental processes are shrouded in mystery. In every arena hypotheses are rife, none of them substantiated sufficiently to compel assent over other and conflicting hypotheses. Some say—for example, Graham Wallas in his book *The Art of Thought*—that in the creation of every work of art there are four successive stages: preparation, incubation, inspiration, and elaboration; others say that these stages are not successive at all but are going on throughout the entire creative process, while still others would produce a different list of stages. Some say that the artist begins with a state of mental confusion, with a few fragments of words or melody gradually becoming clear in his mind and the rest starting from there, working gradually toward clarity and articulation; whereas others hold that the artist sets himself a problem, which he gradually works out during the process of creation, but his vision of the whole guides his creative process from its inception. The first view would be a surprise to the dramatist who sets himself from the beginning to write a drama in five acts about the life and assassination of Julius Caesar; and the second would be a surprise to artists like the 20th-century English artist Henry Moore, who has said he sometimes begins a drawing with no conscious aim but only the wish to use pencil on paper and make tones, lines, and shapes. Again, as to psychological theories about the unconscious motivations of artists during creation, an early Freudian view is that the artist is working out in his creation his unconscious wish fulfillments; a later Freudian view is that he is engaged in working out defenses against superego charges, "proving something to himself." Views based on the ideas of the 20th-century Swiss psychologist Carl Jung reject both these alternatives, substituting an account of the unconscious symbol-making process. Until a great deal more is known about the empirical sciences that bear on the issue, there is little point in attempting to defend one view of artistic creation against another.

The expressive product. Although talk about expression as a process is hedged with difficulties and in any case seems irrelevant to the philosophy of art (as opposed to the psychology of art), there is another way in which talk about expression may be both true and important to the philosophy of art. Mention is made about expressive properties as belonging to works of art: for example, it is said that a certain melody expresses sadness, that there is a feeling of great calm expressed in a particular painting, or that tension is expressed in the thrusts of a tower or the development of the plot of a novel or drama.

The question arises at once of what it means to say such things. Melodies and sentences are not joyous or tense or melancholy; only persons have these qualities. The artist can have them, but how can the work of art? Clearly, to speak of a work of art as having emotions, if it is not to be utter nonsense, must be metaphorical. But what is the meaning of this metaphor? What does it mean to say that the music expresses sadness, if not in the sense of process; *i.e.,* that in writing it the composer expressed his sadness?

Metaphorical expression of emotion

The music is heard, the painting is seen; each presents itself to the senses. But there is much more involved in music than simply hearing (or even listening to) the sounds and in visual art than simply seeing (or even looking at) the colours and shapes. Even very simple combinations of sounds and shapes and colours seem to express certain qualities of life: a curved line, it is said, is graceful or sprightly; the drooping willow tree is sad, as are certain passages in music. It is virtually impossible for most persons to view art as a series of sensory stimuli only. Even when a picture contains no story, no plot, no program, the viewer "reads into the script": he attributes to works of art qualities of human moods, feelings, emotions—in short, "affects." It would be safe to say that in all art, every percept is suffused with affect. The problem is: What is it that makes certain percepts expressive of certain affects?

The simplest answer, that "The melody is sad" means no more or less than "Hearing the melody makes me (or other listeners, or most listeners) feel sad," is surely inadequate. (This would be a theory of evocation, not of expression.) "The music expresses whatever feelings it arouses in me when I hear it." But often the listener does not feel emotions at all (he may imagine them), or, if he does, he feels very different emotions from the ones he believes to be expressed in the music. He may consider the rondo delirious with joy, but if he is grief stricken on a given day he hears it without feeling joy; and if he has heard the same rondo 30 times that day he feels only boredom or fatigue, while still believing that the piece is expressive of joy. Nor is it an adequate analysis to say that "The melody expresses joy" means "I am disposed to (or inclined to) feel joy when I hear it," for many people seem to recognize joy as a quality of the music without feeling it at all: or they may imagine it or just recognize the emotion without feeling it or believe that what they hear sounds the way joy feels—or any of a number of other accounts.

The true analysis of expressiveness in art must be more complex than this: it is not that the melody evokes emotion X, but that emotion X is somehow embodied in the music. But this leads back again to the question, how can an emotional quality be in a work of art? There is no single answer to this question that would be accepted by all philosophers of art, but most accounts begin by noting certain similarities, or analogies, between features of music and features of human feeling; so that when X (a passage of music, for example) is said to express Y (a state of feeling), there are certain similarities (for example, of structure) between X and Y. The physical accompaniments to a mood, say, of restlessness, such as rapid breathing and drumming fingers, have their musical equivalents: trills, quavers, increases in tempo, and the like.

Emotional qualities embodied in art

When a listener says that a certain melody is sad, he is saying that the music literally has certain qualities A, B, C, D that can be perceived in the music. Slowness is surely one such quality (the same melody played fast would not be called sad); another is the absence of large intervals between tones; another is that the sounds tend to be hushed rather than, for example, strident; another is that the tendency of the musical movement is downward rather than rising. When a listener says that the music is sad, he is saying that it has these qualities.

But why these qualities rather than others? Why is it said that the music is sad when it has A, B, C, D rather than when it has M, N, O, and P? Because A, B, C, D are the qualities that also characterize people when they are sad, such as slowness and soft and low speaking voices. If this theory or anything like it is true, it explains how emotional characteristics can be attributed to works of art—why it can be said that the melody is sad, that the horizontal lines in a painting make it calm (horizontal being the position of rest and peace, sleep and maximum relaxation, and from which one does not fall), why the lines in a painting are droopy (they are lines similar in shape to that of, say, an old woman with hunched shoulders), and so on. These qualities, it should be noted, are qualities of the work of art, not of the artist (whether he was sad when he wrote sad music is a separate question, to which the answer may sometimes be no) and not of the listener or observer (the melody is sad even if I am not sad when I hear it). They are, so to speak, embodied in the music, quite independently of the state of the artist or of the observer, or listener, though of course it requires the presence of a listener to recognize them and be moved by them.

Talk about artistic expressiveness, then, can be justified. But there is no need to resort to the language of expression in order to state it: instead of saying, "The music expresses sadness," it can simply be said, "The music is sad." But regardless of the terminology employed, it is important to have justified this conclusion.

ART AS FORM

The formalist position. Against all the foregoing accounts of the function of art stands another, which belongs

distinctively to the 20th century—the theory of art as form, or formalism. The import of formalism can best be seen by noting what it was reacting against: art as representation, art as expression, art as a vehicle of truth or knowledge or moral betterment or social improvement. Formalists do not deny that art is capable of doing these things, but they believe that the true purpose of art is subverted by its being made to do these things. "Art for art's sake, not art for life's sake" is the watchword of formalism. Art is there to be enjoyed, to be savoured, for the perception of the intricate arrangements of lines and colours, of musical tones, of words, and combinations of these. By means of these mediums it is true that objects in the world can be represented, scenes from life depicted, and emotions from life expressed; but these are irrelevant to the principal purpose of art—indeed, art is much less adapted to the telling of a story or the representation of the world than it is to the presentation of colours, sounds, and other items in the art medium simply for their own sake.

"Art for art's sake"

Most people who claim to enjoy paintings, for example, enjoy them not as presentations but as representations of things and situations in life; and thus their response is not of a kind that is unique to art, but one that takes them back to the emotions of life, from which they came. They could use art to take them into a realm of pure form unknown to anyone who is unacquainted with art; but instead they use it to direct them back to the feelings and situations of life. Thus, according to the formalists, these viewers miss the opportunity of being taken into a fresh world of purely aesthetic experience and get from a work only what they bring to it: familiar experiences and emotions they employ the work to recall.

What, then, should be brought from life to art? Knowledge of life's struggles and emotions? Knowledge at least of what people are like, and what visual objects look like? Not even these things, for even they get in the way. Representation is not bad in itself, it is merely irrelevant. Only if the representation is satisfactory as form and contributes to the general abstract design can it be said to matter aesthetically.

Most formalists have directed their attention primarily to visual art. The prerequisite for appreciating this, they believe, is a sense of form and colour and a knowledge of three-dimensional space (the last required because otherwise a cube, for example, would appear in a painting as a flat pattern and would be unable to play the architectural role intended for it). Armed with this bit of knowledge from life, they have all they need (as far as knowledge of the world outside art is concerned) for appreciating visual art. Armed with more than this, they would find their attention drawn away from the sublimities of art to the more approachable concerns of humanity (such as representation). Shorn of this extraneous knowledge and coming to painting with eyes innocent of extraneous concepts, the viewer could then be in a position to look at what painting presents directly to his vision—complex arrangements of forms and colours—which, for reasons thus far unexplained, have the capacity to move the recipient deeply with emotions utterly alien to the emotions of life.

The formalist's account of music runs along similar lines: not only representation (program music) is excluded but so is the entire realm of human feeling—not the feeling that the contemplation of pure form can give but only the feelings of life, such as love or terror. As to literature and all those arts in which words play a part (song, opera, drama, cinema), the formalist position would seem to be impossible; and indeed formalists have seldom attempted to extend their theory to literature. The medium of literature is words and sentences; and words and sentences are not merely noises but noises with meanings; and these meanings inevitably have to do with the objects, actions, qualities, and situations of life. As sound, literature is a poor thing—as a complex of meanings, it can be profound and beautiful. Take these away, and literature could cease to be. Literature does have formal properties—a drama can be as tightly knit as a fugue or a symphony—but the appreciation of it can never consist of these formal properties alone; and the reason lies in the very nature of the medium. It is the role of words to indicate images and meanings and emotions, and these are the stuff of life—therefore verbal art is inescapably humanistic. Whatever sensory beauty words have in their sounds is slight and secondary; not so with painting, where colour, line, and space have a beauty all their own and need stand for nothing outside themselves to satisfy the eye.

Inapplicability of the formalist position

Formal principles in art. What, then, are the specific qualities in works of art that the formalist is seeking? Most formalists have held that a partial account can be given of these but that, in the end, the presence of the qualities must be felt intuitively and cannot be described. Accounts of formal qualities in works of art go back as far as Aristotle's *Poetics,* written in the 4th century BC, and usually include (though sometimes in different terminology) the following as principal ingredients:

Organic unity. A work of art must have what Aristotle called "a beginning, a middle, and an end"; it must be unified, it must "hang together" as one entity. Everything, of course, has some degree of unity or other; even a collection of things, such as a woodpile, has some unity inasmuch as it can correctly be called one thing: it is a collection, but it is a single collection. But the unity desired in works of art is much greater than this: it is more like the unity of the higher organisms in which every part functions not independently of the others but interdependently with them; and it is this interdependency of the parts that constitutes an organic unity. Take away one part, and the remainder of the parts fail to function as before. This is only approximately true of organisms: without a heart or a brain a person could not continue to exist, and the activity of the other organs would cease; but without an ear or a toe they surely would. Philosophers of art have often noted that the purest examples of organic unity in the universe are not organisms but works of art: here the interdependency of parts often achieves a state of such perfection that it could often be said, of a melody or a sonnet, that if this note (or word) were not there, in just the place that it is, the effect on the entire remainder of the melody or poem would be disastrous.

Complexity, or diversity. This principle is the natural accompaniment of the first one. A blank wall has unity but no variety and is not long worth contemplating. Nor is there any triumph in achieving unity at so small a price. The work of art must hold in suspension (as it were) a great diversity of elements and unify them—the greater the complexity that is integrated into a unity, the greater the achievement. This fact is so universally recognized that the two criteria are often stated as one, unity-in-diversity, or variety-in-unity.

Unity-in-variety in works of art

Very many great works of art are much less than perfect organic unities (which is another way of saying that unity-in-variety is not the only criterion for excellence in works of art). Particularly in long poems or novels or operas, some parts are clearly more important than the others, though contributing to the whole, and some parts may be simply "padding." One could hardly allege that the entire ancient Greek epic the *Iliad* is an organic unity and that if (for example) the catalog of ships were removed the entire epic would be ruined (some even say it would be improved). In Dostoyevsky's novels there are whole chapters that are unnecessary from the point of view of relevance to the rest of the story, and, aesthetically (though perhaps not in other ways), these are a pure excrescence. In most works of art there are high spots and low spots, and there is a great deal of elasticity as to what could follow what. But in spite of this, unity-in-variety is quite universally recognized as a criterion for artistic excellence. If a drama consisted of two plots that never connected with one another, even at the end, such a play would be condemned at once for lack of unity; and a work of art is never praised for being disjointed or disunified, though it might be praised in spite of being disunified.

Theme and thematic variation. In many works of art there is a dominant theme, or motif, which stands out and upon which the other portions are centred. This theme is then varied in different ways in other portions of the work. This is a special case of unity-in-variety: if every line in a work of music or literature were entirely novel and different from the other ones, there would be enor-

mous diversity but no unifying connecting links; whereas if there were simply a repetition of the initial theme or of entire sections of the work (as sometimes happens when a composer does not know how to develop the thematic material with which he has begun) there is unity but no variety. Both unity and variety are preserved by having central themes, with other material that is related to them (unity) but not identical with them (variety).

Development, or evolution. In works of temporal art, each part develops or evolves into the next, each part being necessary to the succeeding part, so that if an earlier part were altered or deleted, all the subsequent parts would have to be altered in consequence. If a portion of Act IV could be interchanged with a portion of Act II without loss of effect, the principle of development has not been observed, for then the material occurring in between would not have made any difference.

Balance. The arrangement of the various parts should be balanced, usually in contrasting ways (the adagio movement coming between two faster movements, for example). In painting, there should be a balance between the right and left halves of the canvas. The many ways, other than simple mechanical symmetry ("for every item on the left there should be an item on the right," which soon becomes monotonous), in which a painting may have variety and yet retain balance are too complex to be discussed other than in a book of art criticism. But in its simplest essentials, the principle is acknowledged by everyone: the housewife who places all the furniture on one half of the living room while leaving the other half empty finds the arrangement aesthetically displeasing because the room lacks balance.

Generalizations based on formal principles

There are many descriptions of principles of form in art, which differ from one another in their terminology more than in their final outcome. In general, however, few if any of these principles would be denied (only the detail of their formulations might be) by most philosophers of art. Why certain principles of form are found satisfactory and others are not is a fascinating psychological question, leading back to a discussion (necessarily vague in the present state of knowledge) of the nature of the human organism and the discriminatory powers of the human mind. But however obscure the explanation, the facts of the case seem clear enough: certain formal principles in art must be observed, and, to the degree that they are ignored or violated, aesthetic catastrophe occurs: the work of art cannot evoke our interest or sustain it long once it is initiated.

Though a discussion of these formal principles is helpful particularly to those who have little native sense of form and who want "to know what to look for" in art, they are sufficiently vague so that many critics can agree on a formal principle, such as unity, and yet disagree on the degree to which a specific work possesses it.

In addition, these principles are far from complete: a work of art can possess unity and the other requirements in high degree and yet be unsuccessful even as form. The requirements listed seem only to have skimmed the surface. Yet what more is required to distinguish a formally correct but dull work of art from a brilliant one seems to defy precise analysis. Moreover, the majority of critics who have assented to these principles are not formalists: they have acknowledged and even insisted on the great importance of form in works of art, but they have not alleged, as formalists do, that these principles constitute the sole criteria of excellence in works of art. They have held that the fulfillment of formal criteria counts as a necessary condition for artistic excellence but not a sufficient condition.

PRAGMATIC THEORIES OF ART

There are theories of art that differ from one another in what they allege to be the real purpose or function of art but are at one with each other in the belief that art is a means to some end, whether that end be the titillation of the senses or the communization of the nations of the world or the conversion of mankind to belief in God or the improved moral beliefs or moral tone of the reader or viewer. In every case, the work of art is considered as a means to some end beyond itself, and hence what counts in the final analysis is not the nature of the work of art itself but its effects upon the audience—whether those effects be primarily sensory, cognitive, moral, religious, or social.

Hedonistic theories of art. According to one kind of theory, the function of art is to produce just one kind of effect upon its audience: pleasure. It may also inform or instruct, represent or express, but first and foremost it must please. The more pleasure it gives, the better the art.

If the theory is left in this simple form, it yields the result that glossy and superficial works and those containing nothing difficult or obscure are the best works of art: thus, on the hedonistic account, *King Lear* might come out far behind Henry Wadsworth Longfellow's *The Song of Hiawatha,* or Joyce Kilmer's "Trees," in view of the difficulty of comprehending Shakespeare by many people and the pleasant, easy lilting quality of Longfellow's poem; and similarly a simple ditty might come out ahead of Bach's *Mass in B Minor.* True, Shakespeare and Bach might produce more pleasure in the long run since their works have endured through more centuries. But on the other hand, the simple works can be apprehended and enjoyed by vastly more people.

The meaning of "aesthetic pleasure"

In any case, the theory has often been amended to read "aesthetic pleasure" rather than simply "pleasure"—thus placing great importance on exactly how the term "aesthetic" is to be defined. The definition of this troublesome term is beyond the scope of this article (see AESTHETICS); it will simply be said here that no quick and easy way of distinguishing aesthetic pleasures from other pleasures will suffice for the task at hand. If it is said, for example, that aesthetic pleasure consists in satisfaction taken in the contemplation of sensuous particulars (tones, colours, shapes, smells, tastes) for their own sake—that is, for no further end and without ulterior motive—then one confronts the fact that as much pleasure may be taken in single smells and tastes for their own sakes, without any reference beyond them, as may be taken in the most complex works of art. For that matter, pleasure in playing a game (one not played for money) is pleasure in doing something for its own sake, as is the pleasure of robbing a house if it is done not for money but for "kicks." If something is found pleasurable, ordinarily the pleasure is what one wants from it, not something else beyond it.

Moreover, if it is said that a work of art should be a means toward pleasure, that is treading suspiciously near to the opposed view that art should not be a means to an end but an end in itself. If someone says, "Why do you go jogging every morning for three miles? Because you feel the exercise is good for you?" and another person answers, "No, not that at all, I just enjoy doing it," this would ordinarily and quite sensibly be taken as saying that he did not do exercise as a means toward an end but as an end in itself. If something is done just because it is enjoyed, in common parlance this would be taken to be "doing it as an end in itself"; and if one objected, "No, I'm not doing it as an end in itself, I'm doing it as a means toward the enjoyment I'll get out of it," his reply would be considered sophistical, for doing it for enjoyment's sake is precisely what is ordinarily meant (or one thing that is ordinarily meant) by the statement that a thing is being done for its own sake.

In any case, the effect of great works of art upon a reader or viewer or listener can hardly be described as merely hedonistic. No one would presumably wish to deny that art can and should give us pleasure; but few would wish to assert that pleasure is all that it should give us. If one were to ask, "How did viewing Picasso's 'Guernica' affect you?" and the reply was, "I found it pleasant," we would conclude that his reaction to the painting was, to say the least, inadequate. Great art may please; it may also move, shock, challenge, or change the lives of those who experience it deeply. Pleasure is only one of many kinds of effects it produces.

Art as a means to truth or knowledge. One of the things that has been alleged to be the purpose of art is its cognitive function: art as a means to the acquisition of truth. Art has even been called the avenue to the highest

knowledge available to man and to a kind of knowledge impossible of attainment by any other means.

Knowledge in the most usual sense of that word takes the form of a proposition, knowing that so-and-so is the case. Thus, it can be learned from sense observation that the sun is setting, and this is knowledge. Is knowledge acquired in this same sense from acquaintance with works of art? There is no doubt that there are some propositions (statements) that can be made after acquaintance with works of art that could not be made before: for example, that this performance of Beethoven's *Eroica Symphony* was 47 minutes long, that this painting predominates in green, that this piece of sculpture originated around 350 BC. The question is whether there is anything that can be called truth or knowledge (presumably knowledge is of truths, or true propositions) that can be found in works of art.

The cognitive function of literature

Literature is surely the most obvious candidate; for literature consists of words, and words are combined into sentences, and sentences (at least declarative sentences) are used to convey propositions; that is, to make assertions that are either true or false. And works of literature do certainly contain many true statements: a novel about the French Revolution conveys facts about the series of events; in a verse of the English scholar and poet A.E. Housman (1859–1936), it is said that "The tears of all that be/ Help not the primal fault." Since literature contains statements, it would be surprising indeed if at least some of them were not true.

But the relevance of this fact to literature as an art is extremely dubious. If an 18th-century novel gives a true picture of English country life of that time, this makes it useful to read as history; does it also make it a better novel? Many, at any rate, would say that it does not: that a tenth-rate novel might give more facts about 18th-century life than a first-rate novel of the same century. For that matter, many of the propositions in a novel are, taken at face value, false; it is false, for example, that there was a foundling named Tom Jones who had an uncle named Squire Western. The thousands of pages of description in novels of fictional characters, ascribing to them thoughts and actions, are all false, since these characters never actually existed. Yet this fact in no way impugns their value as literature. Shakespeare, in *The Winter's Tale,* sets part of the action on the seacoast of Bohemia: but the fact that Bohemia has no seacoast does not damage *The Winter's Tale* as literature, though it would as geography. The fact that Milton used the outdated Ptolemaic astronomy does not make *Paradise Lost* less valuable, nor does the nonexistence of the lands described in *Gulliver's Travels* in any way diminish Swift's work. There is no doubt, then, that works of literature can contain true statements and false ones; but it is tempting to ask, what does their truth or falsity matter? Literature is not astronomy or geography or history or any branch of knowledge, particular or general.

Many would hold that the above statements are indeed irrelevant, as are any that encroach upon the domain of science; but, they would add, there are other assertions that matter a great deal: for example, the statements in which a world view is presented in a poem or drama or novel. The main burden of the ancient Latin poet Lucretius' *De rerum natura* ("On the Nature of Things") is a presentation of the materialism of the Greek philosopher Democritus; and an embodiment of the world view of medieval Catholicism is the very warp and woof of Dante's *Divine Comedy*—and such considerations (it would be contended) are relevant to these works as literature.

In reply, however, it might be said that while it is true that these world views must be understood and taken into consideration in the reading of these poems and that they cannot be understood or appreciated without knowing them, the truth or falsity of these views still does matter aesthetically. If Lucretius' view is true, then Dante's must be false, and vice versa, since they are incompatible; but in order to appreciate the poem it is not necessary to know which (if either) is true. Appreciating art, unlike taking a stand for or against a cause in life, does not require a yes or no to statements. It requires only that the viewer look and appreciate, that he experience as richly and fully as possible the feeling and attitudes involved in the world view that is presented. Philosophers and scientists are concerned with whether the Democritean materialism of Lucretius is true; appreciators of art are concerned only to capture the feeling appropriate to the world view in question.

Explicit and implicit statements in literature

Many statements in works of literature are not explicitly made at all but are implicit: Hardy never tells in his novels what his world view is, but it emerges rather clearly before the reader is halfway through any of them. Probably the most important points made in works of literature that contain a central thesis are implicit rather than explicit. How, in that case, can it be determined what thesis it is that is implied? In a court of law, if someone says, "He didn't say it exactly, he just implied it," the judge would be likely to rule that this was insufficient evidence of slander, since the person did not actually say it. Still, many statements in daily life are not stated but implied—in the sense that they are intended; the trouble lies in proving that the speaker intended them, since no one else is in a comparable position to say what his intentions were; and, in the case of deceased authors, there is no evidence of their intentions other than what they said. One is doubtless on safer ground, therefore, saying that many statements are implied in the sense that they are suggested (whether the speaker intended to do so or not) by the tone of voice and the juxtaposition of the words used. Thus, "They had children and got married" suggests, though it does not state, that they had the children before they were married; any normal user of the English language would tend to construe it thus. And it is surely no overstatement to say that Swift's *Gulliver's Travels* suggests that the author was misanthropic or that the novels of the French author Marcel Proust (1871–1922) suggest a pessimistic view of love and other human relationships close to that of the German philosopher Arthur Schopenhauer (1788–1860). A serious reader of literature will become increasingly sensitive to what is suggested in the works he is reading.

But, once again, the importance of the suggested statements, even when they are true, in no way shows that they must be accepted as true by the reader if he is to value them as works of art. Is the sincere Roman Catholic who finds Dante's world view congenial and Lucretius' repellent committed to saying that Dante's is the better poem? If so, he may be accused of confusing his moral and theological judgments with his aesthetic ones. Still, it should be noted that there are some critics who believe that if two works of literature are both equal in excellence on all counts, yet one presents a true view of reality and the other fails to, the one presenting a true view is better—better even as a work of art—than is the other one.

The notion of "truth to human nature" in literature

There is, however, another way of talking about truth in literature that is not or is not as obviously connected to propositions. A characterization in a novel or drama is spoken of as being true to human nature, true to the way people actually speak or behave or feel. No matter that Becky Sharp—in the English novelist William Thackeray's *Vanity Fair*—is a fictional character, it would be said, as long as she is depicted as a person of a certain type would behave, she is being depicted truly; truth in fiction does not mean truth of the statements (for the statements in Thackeray's novel describing her are false), but truth to human nature.

But what exactly does "truth to human nature" mean? The criterion is as old as Aristotle, who wrote that poetry is more true than history because it presents universal truths whereas history gives only particular truths and that poetry (dramatic fiction) shows how a person of this or that kind probably or necessarily would behave (or think, or feel). This criterion, however, is too vague as it stands: what is probable or plausible behaviour in one person is not in another, and what is probable in one set of circumstances is not so in another. The test of truth to human nature would be roughly as follows: Would a person such as has been described thus far (in the novel or drama) behave (or think or feel or be motivated) in the way that the author depicts this character as behaving in the circumstances described? It is often very difficult to decide this question, because knowledge of human beings is insufficient or because the dramatist himself has not

provided enough clues. Still, once readers or critics are convinced that the character described would not have behaved as the novelist depicts him as behaving, they may criticize the characterization (at least with regard to this bit of behaviour or motivation) as implausible. If a character who has been described as spending years working toward a certain goal is represented by the novelist as abandoning it once he is within sight of it, the reader will have considerable reservations about this delineation unless the author has depicted the character as being unstable or masochistic or in some way as being the kind of person who might in these circumstances do this kind of thing. It is true that there are people in the world who abandon their goals within sight of them after years of labour, but the conviction must be implanted that the character already presented by the novelist belongs to this classification or the behaviour will seem reasonless and unmotivated.

Is truth to human nature aesthetically relevant? That is, when present does it make the work of literature better and when absent or flawed does it make the work worse as literature? Here again there would be some difference of opinion, but a very large number of critics and aestheticians, in the tradition of Aristotle, would say that it matters aesthetically a great deal. The novelist does not have to be true to geography or history or astronomy, but he must be, as the 19th-century U.S. author Nathaniel Hawthorne said of all literary artists, true to the human heart. A literary artist may tamper with all the other truths with impunity, but not this one: his characters must be convincing, and they will not be convincing if they are not depicted as having anger, love, jealousy, and other human emotions that real people have and in pretty much the contexts in which real people have them. If a novelist's characters were not motivated in much the way that human beings are motivated, the reader would not even be able to understand them—they would be alien and unintelligible to him. Even when a writer (such as the Englishman Kenneth Grahame in *The Wind in the Willows*) depicts animals as central characters in novels, however much they may differ from human beings in external appearance, they must psychologically be presented as human beings—how else and in what other terms could their behaviour and their motivation be understood? Such, then, are the reasons for saying that whatever else a literary artist does, his depictions must be truthful to human nature.

Truth to human nature in arts other than literature

Can works of art other than literature possess truth to human nature? It would seem that in a limited degree they can. Motion pictures and operas and other mixed arts clearly can, but they employ words, and literature is a principal ingredient in them. But what of arts that employ no words at all? Painting and sculpture, not being temporal arts, cannot depict action, and action is all-important in the representation of human character. These arts contain depictions of persons (real or imaginary) only on a knife-edge of time. Still, sometimes something may be inferred even from a knife-edge. The late self-portraits of the 17th-century Dutch artist Rembrandt do seem to reveal an agonized yet sometimes serene inner spirit, suggesting that there are flashes of human insight to be found in depictions of human beings in visual art. As for musical art (music without the accompaniment of words), it contains nothing that could be called depiction, not even depiction on a knife-edge of time; and, if this is so, there can be no such thing here as true depiction or false depiction. Music may be expressive of human feelings, in the sense already described, but this is a far cry from saying that it contains depictions that are true to human nature.

Even if truth to human nature in the depiction of character is aesthetically relevant (which many would question), to say this is still far from saying that it is the only criterion for excellence in works of art, or even that this is the principal thing that art gives or its main excuse for being. To go so far would be to discount colour and form and expressiveness as criteria for excellence in art; and this virtually no one is willing to do. It would seem, then, that in no case is truth (even truth to human nature) necessary in works of art, seeing that entire genres of art, such as music, exist without it; and that even when it is present and when its presence increases the merit of a work of art (which again many would deny), it is only one virtue among many. Thus, the view that the purpose or function of art is to provide truth is quite surely mistaken; perhaps the person who wants truth and is indifferent to the presence of anything else had better turn to science or philosophy rather than to the arts.

Art as a means to moral improvement. To say that a work of art is aesthetically good or has aesthetic value is one thing; to say that it is morally good or has a capacity to influence people so as to make them morally better is another. Yet, though the two kinds of judgments differ from one another, they are not entirely unrelated. Three views on the relation of art to morality can be distinguished:

Moralism. According to this view, the primary or exclusive function of art is as a handmaiden to morality—which means, usually, whatever system of morality is adhered to by the theorist in question. Art that does not promote moral influence of the desired kind is viewed by the moralist with suspicion and sometimes with grudging tolerance of its existence. For art implants in people unorthodox ideas; it breaks the molds of provincialism in which people have been brought up; it disturbs and disquiets, since it tends to emphasize individuality rather than conformity; and works of art are often created out of rebellion or disenchantment with the established order. Thus, art may undermine beliefs and attitudes on which, it is thought, the welfare of society rests and so may be viewed with suspicion by the guardians of custom. When art does not affect people morally one way or the other (for example, much nonrepresentational painting), it is considered a harmless pleasure that can be tolerated if it does not take up too much of the viewer's time; but, when it promotes questioning and defies established attitudes, it is viewed by the moralist as insidious and subversive. It is viewed with approval only if it promotes or reinforces the moral beliefs and attitudes adhered to by the moralist.

Plato's moralistic view of art

Plato is the first champion in the Western world of the moralistic view of art—at least in *The Republic* and *Laws*. Plato admired the poets and was himself something of a poet; but, when he was founding (on paper) his ideal state, he was convinced that much art, even some passages in Homer, tended to have an evil influence upon the young and impressionable, and accordingly he decided that they must be banned. Passages that spoke ill or questioningly of the gods, passages containing excessive sexual passion (and all works that would today be described as pornographic), and even passages of music that were disturbing to the soul or the senses were all condemned to the same fate. Plato's concern here was with the purity of soul of the men who would become members of the council of rulers of the state; he was not concerned with censorship for the masses, but, since one could not predict which young people would pass the series of examinations required for membership in the council of rulers and since it was (and is) practically impossible to restrict access to works of art to a certain group, the censorship, he decided, would have to be universal. The objection might be raised, to be sure, that rulers to be should not be hothouse plants separated from the influences of the outside world and that they would be better off facing all of reality, including its evils. But Plato's view was that these influences should be kept from them during their formative years—that during this critical time, when the whole tenor of their lives was being shaped, art could be an influence for evil and had to be sacrificed in the interests of morality. In other dialogues of Plato, such as the *Ion* and the *Phaedrus*, when he was not concerned with building a state, he extolled the virtues of art and even held the artist to be divine (although madly divine); but when it came to a conflict between art and morality, it was art that would have to go.

Tolstoy's moralistic view of art

The most famous champion of the moralistic view of art in modern times is Tolstoy. Long after he had finished writing his novels, he fell under the influence of primitive (prechurch) Christianity, the principal tenet of which was the brotherhood of all mankind. This one idea became such an obsession with him that everything else, including the pursuit of art to which he had devoted his life,

became subordinate to it. Almost all the literature of his own time, including all his own novels, he condemned as inimical to the brotherhood of man by emphasizing class distinction and pitting one group of mankind against another. Even art that appealed primarily (in his opinion) to "upper class" tastes, such as the symphonies of Beethoven and the operas of Richard Wagner, both 19th-century German composers, were condemned as "false art." The art that remained after these colossal excisions included such items as folk songs that peasants might sing in the fields as they worked and pictures and stories either illustrating the tenets of primitive Christianity or fostering the spirit of Christianity by promoting the brotherhood of all mankind.

The moralistic view of art is still, on the whole, the unarticulated view of art held by the masses, particularly when they are under the sway of a dominant religious or political doctrine. Historically, Christianity has been suspicious of all art except those works that depicted some aspects of biblical history or that could be used to further the spread of Christian belief and practice (although this is no longer strictly true). It would probably be fair to say that the view of art held by the former Soviet government was a moralistic one: works of fiction and poems had to praise Communism or further its doctrines, and works of music had to be melodic and singable (Soviet composers such as Dmitry Shostakovich were often condemned by the official hierarchy as "too German" or "too materialistic"). Whenever a culture or nation is under the sway of a dominant view, whether moral or religious or political, the tendency of the rulers of that nation is to promote it at all costs—and one of the casualties in the process is art, at any rate that great body of art that is either indifferent or hostile to the reigning dogma.

Aestheticism. Diametrically opposed to the moralistic view is aestheticism, the view that, instead of art (and everything else) being the handmaiden of morality, morality (and everything else) should be the handmaiden of art. The proponents of this view hold that the experience of art is the most intense and pervasive experience available in human life and that nothing should be allowed to interfere with it. If it conflicts with morality, so much the worse for morality; and, if the masses fail to appreciate it or receive the experience it has to offer, so much the worse for the masses. The vital intensity of the aesthetic experience is the paramount goal in human life. If there are morally undesirable effects of art, they do not really matter in comparison to this all-important experience which art can give. When the son-in-law of the 20th-century Italian dictator Benito Mussolini waxed lyrical in his description of the beauty of a bomb exploding in the midst of a crowd of unarmed Ethiopians, he was carrying to its fullest extent the aestheticist's view of art.

Aesthetic values versus moral values

Few persons would wish to go so far. Even the most ardent lovers of art would stop short of saying that the value of art holds a monopoly over all other values. It may well be that the experience of works of art is the greatest experience available to human beings (though this, too, could be questioned), but at any rate it is not the only one available, and, this being the case, the others should be considered as well. There is a plurality of values; and aesthetic values, although far greater, admittedly, than most persons realize, are still just a few among many. It is therefore necessary to consider the relation of the values derived from art to the values derived from other things, such as the conduct of life apart from art: no one can devote every waking hour to the pursuit of art, even if for no other reason than the need for survival, and thus the values of such mundane things as food and shelter have also to be considered.

Mixed positions. The moralistic and aesthetic positions are extremes, and the truth is likely to be found somewhere between them. Indeed, art and morality are intimately related, and neither functions wholly without the other. But to trace the precise relations between art and morality is far from easy; for want of a better term, "interactionism" could be used to label the view that aesthetic and moral values each have distinctive roles to play in the world but that neither operates independently of the other.

It would be admitted, first of all, that works of literature (which will be examined first, since of all the arts the relation of literature to morality is most obvious) can teach valuable moral lessons through explicit presentation: the genre that has this as its aim is didactic literature, as exemplified by *Pilgrim's Progress* by the English Puritan John Bunyan and *Back to Methuselah* by the Irish dramatist George Bernard Shaw. But most works of literature do not exist to teach a moral lesson: possibly, Shakespeare did not write *Othello* merely to attack racial prejudice or *Macbeth* to prove that crime does not pay. Literature does teach but in a far more important way than by explicit preachment: it teaches, as John Dewey said, by being, not by express intent.

Expression of morality in literature

How does literature achieve this moral effect? It presents characters and situations (usually situations of difficult moral decision) through which the reader can deepen his own moral perspectives by reflecting on other people's problems and conflicts, which usually have a complexity that his own daily situations do not possess. He can learn from them without himself having to undergo in his personal life the same moral conflicts or make the same moral decisions. The reader can view such situations with a detachment that he can seldom achieve in daily life when he is immersed in the stream of action. By viewing these situations objectively and reflecting on them, he is enabled to make his own moral decisions more wisely when life calls on him in turn to make them. Literature can be a stimulus to moral reflection unequalled perhaps by any other, for it presents the moral choice in its total context with nothing of relevance omitted.

Perhaps the chief moral potency of literature lies in its unique power to stimulate and develop the faculty of the imagination. Through literature the reader is carried beyond the confines of the narrow world that most persons inhabit into a world of thought and feeling more profound and more varied than his own, a world in which he can share the experiences of human beings (real or fictitious) who are far removed from him in space and time and in attitude and way of life. Literature enables him to enter directly into the affective processes of other human beings, and, having done this, no perceptive reader can any longer condemn or dismiss en masse a large segment of humanity as "foreigners" or "wastrels," for a successful work of literature brings them to life as individuals, animated by the same passions as he is, facing the same conflicts, and tried in the same crucible of bitter experience. Through such an exercise of the sympathetic imagination, literature tends to draw all men together instead of setting them apart from one another in groups or types with convenient labels for each. Far more than preaching or moralizing, more even than the descriptive and scientific discourses of psychology or sociology, literature tends to unite mankind and reveal the common human nature that exists in everyone behind the facade of divisive doctrines, political ideologies, and religious beliefs.

This is not to say, of course, that those who read great works of literature are necessarily tolerant or sympathetic human beings. Reading literature alone is not a cure for human ills, and people who are neurotically grasping or selfish in their private lives will hardly cease to be so as a result of reading works of literature. Still, wide and serious reading of literature has an observable effect: people who do this kind of reading, no matter what their other characteristics may be, do tend to be more understanding of other people's conflicts, to have more sympathy with their problems, and to be able to empathize more with them as human beings than do people who have never broadened their horizons by reading literature at all. No one who has read great literature widely and for a considerable period, so as to make it an integral part of his life, can any longer share the same provincialism and be dominated by the same narrow prejudices that seem to characterize most people most of the time. Literature, perhaps more than anything else, exercises a leavening influence on the temper of a man's moral life. It looses him from the bonds of his own position in space and time; it releases him from exclusive involvement with his own struggles from day to day; it enables him to see his own local problems and

trials from the perspective of eternity—he can now view them as if from an enormous height.

To have moral effects, it is not necessary that a work of literature present a system of morality. Its moral potency is perhaps greatest when it presents not systems but human beings in action, so that through the exercise of the imagination the reader can see his own customs and philosophies as he sees theirs: as some among many of the countless adjustments and solutions to human problems that different circumstances and man's endlessly varied and resourceful nature have produced.

Works of literature, then, develop more than anything else the human faculty of the imagination; and the 19th-century English poet Percy Bysshe Shelley said that the imagination is the greatest single instrument of moral good. Perhaps this sounds like an absurd overstatement, but consider what morality is like without the imagination. Consider the average morality of a small community, relatively isolated from centres of culture and unacquainted with any artistic tradition. Its morality is rigid and circumscribed; the details of each member's personal life are hedged about with constant annoyances, and everyone's life is open to the prying eyes of others who are unfailingly quick to judge, with or without evidence. Outsiders are looked upon askance; people of a different religion, race, or culture are viewed with suspicion and distrust; and anyone who does not subscribe to whatever moral code is dominant in the community is condemned or ostracized. No doubt these people are sincere—they are dreadfully sincere, deadly sincere. But sincerity without enlightenment can be as harmful to the achievement of good as intelligence without wisdom when that intelligence is possessed by political leaders playing with hydrogen bombs. Generally speaking, the people of a small community have not known the leavening influence of literature. Their morality is rigid, cramped, and arid. If these same people had been exposed from early youth to great masterpieces of literature and had learned through them to appreciate the tremendous diversity of human mores and beliefs held by other groups, with the same degree of sincerity that they themselves possess, they would be less likely to be as harsh, intolerant, and rigid as they are.

People are usually inclined to separate art and morality into two hermetically sealed compartments. They talk as if morality were already complete and self-sufficient without art, and that art, if it is to be tolerated at all, can grudgingly be permitted, provided that it conforms to the moral customs of the time and place of those judging it. But this view is surely to conceive the relation between art and morality in far too one-sided a manner. If art must take cognizance of morality, equally morality must take cognizance of art. Almost everything that is alive and imaginative about morality comes from the leavening influence of art.

Influence of ancient dramatists

To consider examples from ancient Greece alone, what would morality be today without the influence of the dramatists Aeschylus and Sophocles, without Socrates as described in Plato's dialogues, even without the historians Herodotus and Thucydides with their quiet humour, gentle prodding skepticism, and tolerance for other customs and views? It is through great works of art that the most vivid conceptions of various ways of life are obtained. What is it about other times and places that people most remember? Is it their political squabbles, their wars, their economic upheavals? These events are known in general to intelligent laymen and in detail to historians, but even then such events do not usually make much of a dent in peoples' personal lives in the way that art does. What is alive today about ancient Greece is its sculpture, its poetry, its epic and drama; what is alive today about the Elizabethan period, even more than the defeat of the Spanish Armada and the reign of Queen Elizabeth I, is its poetic drama, with its vivid characterizations and boundless energy. Other civilizations and cultures may be sources of facts and theories that enlighten modern understanding, but what enables contemporary man to share directly their feelings and attitudes toward life is not their politics nor even their religion but their art. Art alone is never out of date. Science is cumulative; even the science textbooks of ten years ago are now discarded as obsolete; the science of the ancient Greeks and the Elizabethans is studied today primarily for its historical value. But great art is never obsolete; it can still present to modern man its full impact, undiminished by time. Shakespeare will not be out of date as long as human beings continue to feel love, jealousy, and conflict in a troubled world. A biblical statement might be paraphrased and applied to past cultures: "By their arts shall ye know them." The artists whose works are now revered may have died unsung; most of them, even those who were appreciated during their lifetimes, were considered far less important than the latest naval victories or the accession of the current king; yet today these things have all passed into history, but art survives with undiminished vigour. The art of the past molds in countless ways the attitudes, responses, and dispositions of modern man's daily life. Most of what is perceptive and imaginative in morality owes its origin to art, and, when morality loses contact with the tradition of art, it becomes dead and sterile. Yet, in spite of this, some people tell us that art is merely the salve of morality, to soothe its stringency.

Already, in the preceding paragraph, mention has begun of arts other than literature. How, it could be asked, can they have any moral effects on those who view them or listen to them? Yet there are effects of these arts on the observer that, in a broad sense, are moral (as opposed to nonmoral) and that account for the attempts of many people to censor them.

Aristotle's theory of catharsis

Historically, the most famous supposition about the moral effect of art on its audience is Aristotle's theory of catharsis; Aristotle applied the theory to tragedy only, but many since his day have applied it to art in general. According to this view, art acts as an emotional cathartic and achieves a "purgation of the emotions." Certain emotions man would be better off without (Aristotle limited them to pity and fear, but they could easily be extended) are generated during the course of daily life. Art is the principal agency that should help to dispel these emotions. By observing works of art (witnessing a drama, listening to a powerful symphony, looking at certain works of sculpture or painting) the recipient can work off these emotions rather than let them fester inside him or take them out in unpleasant ways on his fellow men. Art siphons off these disturbing inner states rather than letting them grow rancid within man.

As it stands, this view is undoubtedly somewhat crude, especially in the light of modern psychology; and fault could be found in many respects with the Aristotelian doctrine of catharsis. Yet the experience of reading, viewing, or listening to a work of art does give a peculiar release, a feeling of freedom from inner turbulence. The mere act of plunging, for a few hours, into an entirely different world when attending a play or a concert is often enough to transform, however temporarily, the tone of people's daily lives. It is not merely that for a few hours they can forget their troubles—any form of entertainment, however worthless, might do this. It is not merely that art provides a break or interruption in the course of people's lives at the end of which they are exactly what they were before. It is that through the aesthetic process itself, in the very act of concentrating energies on an art object of great unity and complexity and depth, a kind of inner clarification is achieved that was not present before.

It is not true, therefore, that reading novels of crime and detection leads people to indulge in a life of crime; on the whole, those who read such novels are law-abiding people, and, if anything, the reading of such novels is a substitute for aggressive activity (it is aggression vicariously experienced) rather than an incitement to it. Nor do works of art of a licentious nature usually incite people to rape or adultery; far from acting as incitements to action, they are safety valves against action by providing a kind of substitute gratification. It has been said, for example, that Shakespeare's *Antony and Cleopatra* is an immoral work because it celebrates the passionate surrender to an illicit love and the victory of this love over practical, political, and moral concerns. But is there any evidence that people who read this play will behave like the lovers in question

because they read the play? On the contrary: it could be argued that reading the play has an instrumental value in that it presents another example of a complex moral situation, the perusal of which provides many avenues for moral reflection, and that the play also possesses the intrinsic value of acute characterization, dramatic power, and poetry whose imagery and intensity are among the most splendid in the English language. Again, it is said that American youth has been demoralized by such 20th-century U.S. writers as Ernest Hemingway and William Faulkner in that these writers set an example of bad behaviour. But to say that they are capable of demoralizing an entire generation is certainly to attribute to them too much moral power, especially over people who have never even heard of them. Even among those who do read serious literature, the effects are probably more beneficial than harmful: through books the horizons of such readers have been expanded to include other ways of life than they would have previously known.

The essential morality of the aesthetic experience

Quite apart from the ultimate effect of a work of art on a man's emotions, it would appear that the very act of experiencing the work may itself have a moral effect. If he is really concentrating on the details of a work of art and not just passively letting it play upon his senses, this effect—the heightening of his sensibilities and the refining of his capacities for perceptual discrimination—will make him more receptive to the world around him, thus raising the tone of his daily life and making his experience of the world richer than before.

Most of what passes for aesthetic appreciation does not begin to have this effect; but its failure is only because it is not aesthetic appreciation at all—it is a kind of tired reverie rather than an intense absorption in the aesthetic object. Most people, when they hear music, simply allow themselves to be inundated by the sheer flow of sound. Such people do not actively listen to the music and are not even aware of the most elementary kind of ebb and flow occurring within it; they only receive it passively, perhaps using it as a springboard for a private reverie or an emotional debauch of their own. Music has for them not an aesthetic effect but an anesthetic effect. It is not just hearing music that will have the required effect. The aesthetic experience, which involves nothing less than a total concentration on the perceptual details of the aesthetic object, is an experience that heightens consciousness, exercises man's capacity for perceptual awareness and discrimination, and helps him come alive to the sight and texture of the world around him. After a viewer has seen an exhibition of landscapes by Cézanne, the entire world may seem to him to have changed its structure and complexion: it may, indeed, take on the look of Cézanne's landscapes. And is not anything that increases awareness and subtlety of discernment and discrimination a potentially moral agent? Art provides the most intense, concentrated, and sharply focussed of the experiences available to man. Because of this, art can have an enormous influence on the tenor of a person's life, more influential no doubt than any particular system of morality. In its ability to do this, it has an effect on man's life that, in an extended sense at least, can surely be called moral. Morality transcends particular systems of morality; and art, by being for many persons the dominant influence in their lives, thus transcends them also. (Jo.Ho.)

Philosophy of logic

The term logic comes from the Greek word *logos*. The variety of senses that *logos* possesses may suggest the difficulties to be encountered in characterizing the nature and scope of logic. Among the partial translations of *logos*, there are "sentence," "discourse," "reason," "rule," "ratio," "account" (especially the account of the meaning of an expression), "rational principle," and "definition." Not unlike this proliferation of meanings, the subject matter of logic has been said to be the "laws of thought," "the rules of right reasoning," "the principles of valid argumentation," "the use of certain words labelled 'logical constants'," "truths (true propositions) based solely on the meanings of the terms they contain," and so on.

LOGIC AS A DISCIPLINE

Valid argument and matters of fact

Nature and varieties of logic. It is relatively easy to discern some order in the above embarrassment of explanations. Some of the characterizations are in fact closely related to each other. When logic is said, for instance, to be the study of the laws of thought, these laws cannot be the empirical (or observable) regularities of actual human thinking as studied in psychology; they must be laws of correct reasoning, which are independent of the psychological idiosyncrasies of the thinker. Moreover, there is a parallelism between correct thinking and valid argumentation: valid argumentation may be thought of as an expression of correct thinking, and the latter as an internalization of the former. In the sense of this parallelism, laws of correct thought will match those of correct argumentation. The characteristic mark of the latter is, in turn, that they do not depend on any particular matters of fact. Whenever an argument that takes a reasoner from p to q is valid, it must hold independently of what he happens to know or believe about the subject matter of p and q. The only other source of the certainty of the connection between p and q, however, is presumably constituted by the meanings of the terms that the propositions p and q contain. These very same meanings will then also make the sentence "If p, then q" true irrespective of all contingent matters of fact. More generally, one can validly argue from p to q if and only if the implication "If p, then q" is logically true—*i.e.*, true in virtue of the meanings of words occurring in p and q, independently of any matter of fact.

Logic may thus be characterized as the study of truths based completely on the meanings of the terms they contain.

In order to accommodate certain traditional ideas within the scope of this formulation, the meanings in question may have to be understood as embodying insights into the essences of the entities denoted by the terms, not merely codifications of customary linguistic usage.

The following proposition (from Aristotle), for instance, is a simple truth of logic: "If sight is perception, the objects of sight are objects of perception." Its truth can be grasped without holding any opinions as to what, in fact, the relationship of sight to perception is. What is needed is merely an understanding of what is meant by such terms as "if–then," "is," and "are," and an understanding that "object of" expresses some sort of relation.

The logical truth of Aristotle's sample proposition is reflected by the fact that "The objects of sight are objects of perception" can validly be inferred from "Sight is perception."

The various branches of logic

Many questions nevertheless remain unanswered by this characterization. The contrast between matters of fact and relations between meanings that was relied on in the characterization has been challenged, together with the very notion of meaning. Even if both are accepted, there remains a considerable tension between a wider and a narrower conception of logic. According to the wider interpretation, all truths depending only on meanings belong to logic. It is in this sense that the word logic is to be taken in such designations as "epistemic logic" (logic of knowledge), "doxastic logic" (logic of belief), "deontic logic" (logic of norms), "the logic of science," "inductive logic," and so on. According to the narrower conception, logical truths obtain (or hold) in virtue of certain specific terms, often called logical constants. Whether they can be given an intrinsic characterization or whether they can be specified only by enumeration is a moot point. It is generally agreed, however, that they include (1) such propositional connectives as "not," "and," "or," and "if–then" and (2) the so-called quantifiers "$(\exists x)$" (which may be read: "For at least one individual, call it x, it is true that") and "$(\forall x)$" ("For each individual, call it x, it is true that"). The dummy letter x is here called a bound (individual) variable. Its values are supposed to be members of some fixed class of entities, called individuals, a class that is variously known as the universe of discourse, the universe presupposed in an interpretation, or the domain of individuals. Its members are said to be quantified over in "$(\exists x)$" or "$(\forall x)$." Furthermore, (3) the concept of identity (expressed by $=$) and (4) some notion of predication

(an individual's having a property or a relation's holding between several individuals) belong to logic. The forms that the study of these logical constants take are described in greater detail in the article LOGIC, in which the different kinds of logical notation are also explained. Here, only a delineation of the field of logic is given.

When the terms in (1) alone are studied, the field is called propositional logic. When (1), (2), and (4) are considered, the field is the central area of logic that is variously known as first-order logic, quantification theory, lower predicate calculus, lower functional calculus, or elementary logic. If the absence of (3) is stressed, the epithet "without identity" is added, in contrast to first-order logic with identity, in which (3) is also included.

Borderline cases between logical and nonlogical constants are the following (among others): (1) Higher order quantification, which means quantification not over the individuals belonging to a given universe of discourse, as in first-order logic, but also over sets of individuals and sets of *n*-tuples of individuals. (Alternatively, the properties and relations that specify these sets may be quantified over.) This gives rise to second-order logic. The process can be repeated. Quantification over sets of such sets (or of *n*-tuples of such sets or over properties and relations of such sets) as are considered in second-order logic gives rise to third-order logic; and all logics of finite order form together the (simple) theory of (finite) types. (2) The membership relation, expressed by $\in$, can be grafted on to first-order logic; it gives rise to set theory. (3) The concepts of (logical) necessity and (logical) possibility can be added.

This narrower sense of logic is related to the influential idea of logical form. In any given sentence, all of the nonlogical terms may be replaced by variables of the appropriate type, keeping only the logical constants intact. The result is a formula exhibiting the logical form of the sentence. If the formula results in a true sentence for any substitution of interpreted terms (of the appropriate logical type) for the variables, the formula and the sentence are said to be logically true (in the narrower sense of the expression).

Features and problems of logic. Three areas of general concern are the following.

Logical semantics. For the purpose of clarifying logical truth and hence the concept of logic itself, a tool that has turned out to be more important than the idea of logical form is logical semantics, sometimes also known as model theory. By this is meant a study of the relationships of linguistic expressions to those structures in which they may be interpreted and of which they can then convey information. The crucial idea in this theory is that of truth (absolutely or with respect to an interpretation). It was first analyzed in logical semantics around 1930 by the Polish-American logician Alfred Tarski. In its different variants, logical semantics is the central area in the philosophy of logic. It enables the logician to characterize the notion of logical truth irrespective of the supply of nonlogical constants that happen to be available to be substituted for variables, although this supply had to be used in the characterization that turned on the idea of logical form. It also enables him to identify logically true sentences with those that are true in every interpretation (in "every possible world").

Truth as defined in model theory

The ideas on which logical semantics is based are not unproblematic, however. For one thing, a semantical approach presupposes that the language in question can be viewed "from the outside"; *i.e.,* considered as a calculus that can be variously interpreted and not as the all-encompassing medium in which all communication takes place (logic as calculus versus logic as language).

Furthermore, in most of the usual logical semantics the very relations that connect language with reality are left unanalyzed and static. Ludwig Wittgenstein, an Austrian-born philosopher, discussed informally the "language-games"—or rule-governed activities connecting a language with the world—that are supposed to give the expressions of language their meanings; but these games have scarcely been related to any systematic logical theory. Only a few other attempts to study the dynamics of the representative relationships between language and reality have been made. The simplest of these suggestions is perhaps that the semantics of first-order logic should be considered in terms of certain games (in the precise sense of game theory) that are, roughly speaking, attempts to verify a given first-order sentence. The truth of the sentence would then mean the existence of a winning strategy in such a game.

Limitations of logic. Many philosophers are distinctly uneasy about the wider sense of logic. Some of their apprehensions, voiced with special eloquence by a contemporary Harvard University logician, Willard Van Quine, are based on the claim that relations of synonymy cannot be fully determined by empirical means. Other apprehensions have to do with the fact that most extensions of first-order logic do not admit of a complete axiomatization; *i.e.,* their truths cannot all be derived from any finite—or recursive (see below)—set of axioms. This fact was shown by the important "incompleteness" theorems proved in 1931 by Kurt Gödel, an Austrian (later, American) logician, and their various consequences and extensions. (Gödel showed that any consistent axiomatic theory that comprises a certain amount of elementary arithmetic is incapable of being completely axiomatized.) Higher-order logics are in this sense incomplete and so are all reasonably powerful systems of set theory. Although a semantical theory can be built for them, they can scarcely be characterized any longer as giving actual rules—in any case complete rules—for right reasoning or for valid argumentation. Because of this shortcoming, several traditional definitions of logic seem to be inapplicable to these parts of logical studies.

Incompleteness of higher-order logics

These apprehensions do not arise in the case of modal logic, which may be defined, in the narrow sense, as the study of logical necessity and possibility; for even quantified modal logic admits of a complete axiomatization. Other, related problems nevertheless arise in this area. It is tempting to try to interpret such a notion as logical necessity as a syntactical predicate; *i.e.,* as a predicate the applicability of which depends only on the form of the sentence claimed to be necessary—rather like the applicability of formal rules of proof. It has been shown, however, by Richard Montague, an American logician, that this cannot be done for the usual systems of modal logic.

Logic and computability. These findings of Gödel and Montague are closely related to the general study of computability, which is usually known as recursive function theory (see MATHEMATICS, FOUNDATIONS OF: *The crisis in foundations following 1900: Logicism, formalism, and the metamathematical method*) and which is one of the most important branches of contemporary logic. In this part of logic, functions—or laws governing numerical or other precise one-to-one or many-to-one relationships—are studied with regard to the possibility of their being computed; *i.e.,* of being effectively—or mechanically—calculable. Functions that can be so calculated are called recursive. Several different and historically independent attempts have been made to define the class of all recursive functions, and these have turned out to coincide with each other. The claim that recursive functions exhaust the class of all functions that are effectively calculable (in some intuitive informal sense) is known as Church's thesis (named after the American logician Alonzo Church).

One of the definitions of recursive functions is that they are computable by a kind of idealized automaton known as a Turing machine (named after Alan Mathison Turing, a British mathematician and logician). Recursive function theory may therefore be considered a theory of these idealized automata. The main idealization involved (as compared with actually realizable computers) is the availability of a potentially infinite tape.

The theory of computability prompts many philosophical questions, most of which have not so far been answered satisfactorily. It poses the question, for example, of the extent to which all thinking can be carried out mechanically. Since it quickly turns out that many functions employed in mathematics—including many in elementary number theory—are nonrecursive, one may wonder whether it follows that a mathematician's mind in thinking of such functions cannot be a mechanism and whether the possibly nonmechanical character of mathematical thinking

may have consequences for the problems of determinism and free will. Further work is needed before definitive answers can be given to these important questions.

ISSUES AND DEVELOPMENTS IN THE PHILOSOPHY OF LOGIC

In addition to the problems and findings already discussed, the following topics may be mentioned.

Analytical truth

Meaning and truth. Since 1950, the concept of analytical truth (logical truth in the wider sense) has been subjected to sharp criticism, especially by Quine. The main objections turned around the nonempirical character of analytical truth (arising from meanings only) and of the concepts in terms of which it could be defined—such as synonymy, meaning, and logical necessity. The critics usually do not contest the claim that logicians can capture synonymies and meanings by starting from first-order logic and adding suitable further assumptions, though definitory identities do not always suffice for this purpose. The crucial criticism is that the empirical meaning of such further "meaning postulates" is not clear.

Logical semantics of modal concepts. In this respect, logicians' prospects have been enhanced by the development of a semantical theory of modal logic, both in the narrower sense of modal logic, which is restricted to logical necessity and logical possibility, and in the wider sense, in which all concepts that exhibit similar logical behaviour are included. This development, initiated between 1957 and 1959 largely by Stig Kanger of Sweden and Saul Kripke of the U.S., has opened the door to applications in the logical analysis of many philosophically central concepts, such as knowledge, belief, perception, and obligation. Attempts have been made to analyze from the viewpoint of logical semantics such philosophical topics as sense-datum theories, knowledge by acquaintance, the paradox of saying and disbelieving propounded by the British philosopher G.E. Moore, and the traditional distinction between statements *de dicto* ("from saying") and statements *de re* ("from the thing"). These developments also provide a framework in which many of those meaning relations can be codified that go beyond first-order logic, and may perhaps even afford suggestions as to what their empirical content might be.

Intensional logic. Especially in the hands of Montague, the logical semantics of modal notions has blossomed into a general theory of intensional logic; *i.e.,* a theory of such notions as proposition, individual concept, and in general of all entities usually thought of as serving as the meanings of linguistic expressions. (Propositions are the meanings of sentences, individual concepts are those of singular terms, and so on.) A crucial role is here played by the notion of a possible world, which may be thought of as a variant of the logicians' older notion of model, now conceived of realistically as a serious alternative to the actual course of events in the world. In this analysis, for instance, propositions are functions that correlate possible worlds with truth-values. This correlation may be thought of as spelling out the older idea that to know the meaning of a sentence is to know under what circumstances (in which possible worlds) it would be true.

Logic and information. Even though none of the problems listed seems to affect the interest of logical semantics, its applications are often handicapped by the nature of many of its basic concepts. One may consider, for instance, the analysis of a proposition as a function that correlates possible worlds with truth-values. An arbitrary function of this sort can be thought of (as can functions in general) as an infinite class of pairs of correlated values of an independent variable and of the function, like the coordinate pairs (x, y) of points on a graph. Although propositions are supposed to be meanings of sentences, no one can grasp such an infinite class directly when understanding a sentence; he can do so only by means of some particular algorithm, or recipe (as it were), for computing the function in question. Such particular algorithms come closer in some respects to what is actually needed in the theory of meaning than the meaning entities of the usual intensional logic.

Possible worlds and truth-values

This observation is connected with the fact that, in the usual logical semantics, no finer distinctions are utilized in semantical discussions than logical equivalence. Hence the transition from one sentence to another logically equivalent one is disregarded for the purposes of meaning concepts. This disregard would be justifiable if one of the most famous theses of Logical Positivists were true in a sufficiently strong sense, viz., that logical truths are really tautologies (such as "It is either raining or not raining") in every interesting objective sense of the word. Many philosophers have been dissatisfied with the stronger forms of this thesis, but only recently have attempts been made to spell out the precise sense in which logical and mathematical truths are informative and not tautologous.

Problems of ontology. Among the ontological problems—problems concerning existence and existential assumptions—arising in logic are those of individuation and existence.

Individuation. Not all interesting interpretational problems are solved by possible-world semantics, as the developments earlier registered are sometimes called. The systematic use of the idea of possible worlds has raised, however, the subject of cross identification; *i.e.,* of the principles according to which a member of one possible world is to be found identical or nonidentical with one of another. Since one can scarcely be said to have a concept of an individual if he cannot locate it in several possible situations, the problem of cross-identification is also one of the most important ingredients of the logical and philosophical problem of individuation. The criticisms that Quine has put forward concerning modal logic and analyticity (see above) can be deepened into questions concerning methods of cross identification. Although some such methods undoubtedly belong to everyone's normal unarticulated conceptual repertoire, it is not clear that they are defined or even definable widely enough to enable philosophers to make satisfactory sense of a quantified logic of logical necessity and logical possibility. The precise principles used in ordinary discourse—or even in the language of science—pose a subtle philosophical problem. The extent to which special "essential properties" are relied on in individuation and the role of spatio-temporal frameworks are moot points here. It has also been suggested that essentially different methods of cross identification are actually used together, some of them depending on impersonal descriptive principles and others on the perspective of a person.

Existence and ontology. Because one of the basic concepts of first-order logic is that of existence, as codified by the existential quantifier "$(\exists x)$," one might suppose that there is little room left for any separate philosophical problem of existence. Yet existence, in fact, does seem to pose a problem, as witnessed by the bulk of the relevant literature. Some issues are relatively easy to clarify. In the usual formulations of first-order logic, for instance, there are "existential presuppositions" present to the effect that none of the singular terms employed is without a bearer (as "Pegasus" is). It is a straightforward matter, however, to dispense with these presuppositions. Though this seems to involve the procedure, branded as inadmissible by many philosophers, of treating existence as a predicate, this can nonetheless be easily done on the formal level. Given certain assumptions, it may even be shown that this "predicate" will have to be "$(\exists x)\ (x = a)$" (for "*a* exists"—literally, "There exists an x such that x is a") or something equivalent. Furthermore, the logical peculiarities of this predicate seem to explain amply philosophers' apparent denial of its reality.

Entities to which a theory is committed

The interest in the notion of existence is connected with the question of what entities a theory commits its holder to or what its "ontology" is. The "predicate of existence" just mentioned recalls Quine's criterion of ontological commitment: "To be is to be a value of a bound variable"—*i.e.,* of the x in $(\forall x)$ or in $(\exists x)$. According to Quine, a theory is committed to those and only those entities that in the last analysis serve as the values of its bound variables. Thus ordinary first-order theory commits one to an ontology only of individuals (particulars), whereas higher order logic commits one to the existence of sets—*i.e.,* of collections of definite and distinct entities (or, alternatively, of properties and relations). Likewise, if

bound first-order variables are assumed to range over sets (as they do in set theory), a commitment to the existence of these sets is incurred.

The doctrine that an ontology of individuals is all that is needed is known as (the modern version of) nominalism. The opposite view is known as (logical) realism. Even those philosophers who profess sympathy with nominalism find it hard, however, to maintain that mathematics could be built on a consistently nominalistic foundation.

The precise import of Quine's criterion of ontological commitment, however, is not completely clear. Nor is it clear in what other sense one is perhaps committed by a theory to those entities that are named or otherwise referred to in it but not quantified over in it. Questions can also be raised concerning the very distinction between what in modern logic are usually called individuals ("particulars" would be a more traditional designation) and such universals as their properties and relations; and these questions can be combined with others concerning the "tie" that binds particulars and universals together in predication.

An interesting approach to these problems is the distinction made by Gottlob Frege, a pioneer of mathematical logic in the late 19th century, between individuals—he called them objects—and what he called functions (which in his view include concepts) and his doctrine of the unsaturated character of the latter, according to which a function (as it were) contains a gap, which can be filled by an object. Another approach is the "picture theory of language" of Wittgenstein's *Tractatus Logico-Philosophicus,* according to which a simple sentence presents a person with an isomorphic representation (a "picture") of reality as it would be if the sentence were true. According to this view (which was later given up by Wittgenstein), "a sentence [or proposition, *Satz*] is a model of reality such as we think of it as being."

Alternative logics. The natures of most of the so-called nonclassical logics can be understood against the background of what has here been said. Some of them are simply extensions of the "classical" first-order logic—*e.g.,* modal logics and many versions of intensional logic. The so-called free logics are simply first-order (or modal) logics without existential presuppositions.

Intuitionistic logic

One of the most important nonclassical logics is intuitionistic logic, first formalized by the Dutch mathematician Arend Heyting in 1930. It has been shown that this logic can be interpreted in terms of the same kind of modal logic serving as a system of epistemic logic. In the light of its purpose to consider only the known, this isomorphism is suggestive. The avowed purpose of the intuitionist is to consider only what can actually be established constructively in logic and in mathematics—*i.e.,* what can actually be *known.* Thus, he refuses to consider, for example, "Either A or not-A" as a logical truth, for it does not actually help one in knowing whether A or not-A is the case. This does not close, however, the philosophical problem about intuitionism. Special problems arise from intuitionists' rejection (in effect) of the nonepistemic aspects of logic, as illustrated by the fact that only a part of epistemic logic is needed in this translation of intuitionistic logic into epistemic logic.

Other new logics are obtained by modifying the rules of those games that are involved in the game-theoretical interpretation of first-order logic mentioned above. The logician may reject, for instance, the assumption that he possesses perfect information, an assumption that characterizes classical first-order logic. One may also try to restrict the strategy sets of the players—to recursive strategies, for example.

Many-valued logics

Among the oldest kinds of alternative logics are many-valued logics. In them, more truth values than the usual true and false are assumed. The idea seems very natural when considered in abstraction from the actual use of logic. But a philosophically satisfactory interpretation of many-valued logics is not equally straightforward. The interest in finite-valued logics and the applicability of them are sometimes exaggerated. The idea, however, of using the elements of an arbitrary Boolean algebra—a generalized calculus of classes—as abstract truth-values has provided a powerful tool for systematic logical theory.

LOGIC AND OTHER DISCIPLINES

Technical disciplines. The relations of logic to mathematics, to computer technology, and to the empirical sciences are here considered.

Mathematics. It is usually said that all of mathematics can, in principle, be formulated in a sufficiently theorem-rich system of axiomatic set theory. What the axioms of a set theory that could accomplish this might be, however, and whether they are at all natural is not obvious in every case. (The recent development in abstract algebra known as category theory offers the most conspicuous examples of these problems.) The axioms of set theory may be presumed to hold in virtúe of the meanings of the terms set, member of, and so on. Thus, in some loose sense all of pure mathematics falls within the scope of logic in the wider sense. This assertion is not very informative, however, as long as the logician has no ways of analyzing these meanings so as to be able to tell what assumptions (axioms of set theory) should be adopted. The definitions of basic mathematical concepts (such as "number") in logical terms proposed by Gottlob Frege (in 1884), by Bertrand Russell (in 1903), and by their successors do not help in this enterprise. It is not clear that more recent insights in logic help very much, either, in the search for strong set-theoretical assumptions. The relationship of mathematics to logic on this level therefore remains ambiguous.

Notwithstanding these deep problems, virtually all normal mathematical argumentation is carried out in logical terms—mostly in first-order terms, but with a generous sprinkling of second-order reasoning and various principles of set theory. Historically speaking, most specific early examples of nontrivial logical reasoning were taken from mathematics.

Often these examples were set in contrast to logical arguments understood in a narrow traditional sense—in a sense narrower still than the idea of logic as being exhausted by quantification theory. According to this traditional view, logic is equated with syllogistic; *i.e.,* with a part of that part of first-order logic that deals with properties and not with relations. Much of what earlier philosophers said of mathematical reasoning must, thus, be understood as applying to relational (first-order) reasoning. The present-day philosophy of logic is therefore as much an heir to traditional philosophy of mathematics as to traditional philosophy of logic.

Specific logical results are applicable in several parts of mathematics, especially in algebra, and various concepts and techniques used by logicians have often been borrowed from mathematics. (Thus one can even speak of "the mathematics of metamathematics.")

Computers. It has already been indicated that recursive function theory is, in effect, the study of certain idealized automata (computers). It is, in fact, a matter of indifference whether this theory belongs to logic or to computer science. The idealized assumption of a potentially infinite computer tape, however, is not a trivial one: Turing machines typically need plenty of tape in their calculations. Hence the step from Turing machines to finite automata (which are not assumed to have access to an infinite tape) is an important one.

Turing machines and the infinite tape

This limitation does not dissociate computer science from logic, however, for other parts of logic are also relevant to computer science and are constantly employed there. Propositional logic may be thought of as the "logic" of certain simple types of switching circuits. There are also close connections between automata theory and the logical and algebraic study of formal languages. An interesting topic on the borderline of logic and computer science is mechanical theorem proving, which derives some of its interest from being a clear-cut instance of the problems of artificial intelligence, especially of the problems of realizing various heuristic modes of thinking on computers. In theoretical discussions in this area, it is nevertheless not always understood how much textbook logic is basically trivial and where the distinctively nontrivial truths of logic (including first-order logic) lie.

Methodology of the empirical sciences. The quest for theoretical self-awareness in the empirical sciences has led to interest in methodological and foundational problems as well as to attempts to axiomatize different empirical theories. Moreover, general methodological problems, such as the nature of scientific explanations, have been discussed intensively among philosophers of science. In all of these endeavours, logic plays an important role.

By and large, there are here three different lines of thought. (1) Often, only the simplest parts of logic—*e.g.*, propositional logic—are appealed to (over and above the mere use of logical notation). Sometimes, claims regarding the usefulness of logic in the methodology of the empirical sciences are, in effect, restricted to such rudimentary applications. This restriction is misleading, however, for most of the interesting and promising connections between methodology and logic lie on a higher level, especially in the area of model theory. In econometrics, for instance, a special case of the logicians' problems of definability plays an important role under the title "identification problem." On a more general level, logicians have been able to clarify the concept of a model as it is used in the empirical sciences.

Use of first-order versus set-theoretical formulations

In addition to those employing simple logic, two other contrasting types of theorists can be distinguished: (2) philosophers of science, who rely mostly on first-order formulations, and (3) methodologists (*e.g.*, Patrick Suppes, a U.S. philosopher and behavioral scientist), who want to use the full power of set theory and of the mathematics based on it. Both approaches have advantages. Usually realistic axiomatizations and other reconstructions of actual scientific theories are possible only in terms of set theoretical and other strong mathematical conceptualizations (theories conceived of as "set-theoretical predicates"). In spite of the oversimplification that first-order formulations often entail, however, they can yield theoretical insights because first-order logic (including its model theory) is mastered by logicians much more thoroughly than is set theory.

Many empirical sciences, especially the social sciences, use mathematical tools borrowed from probability theory and statistics, together with such outgrowths of these as decision theory, game theory, utility theory, and operations research. A modest but not uninteresting beginning in the study of their foundations has been made in modern inductive logic.

Human disciplines. The relations of logic to linguistics, psychology, law, and education are here considered.

Linguistics. The revival of interest in semantics among theoretical linguists in the late 1960s awakened their interest in the interrelations of logic and linguistic theory as well. It was also discovered that certain grammatical problems are closely related to logicians' concepts and theories. A near-identity of linguistics and "natural logic" has been claimed by the U.S. linguist George Lakoff. Among the many conflicting and controversial developments in this area, special mention may perhaps be made of attempts by Jerrold J. Katz, a U.S. grammarian-philosopher, and others to give a linguistic characterization of such fundamental logical notions as analyticity; the sketch by Montague of a "universal grammar" based on his intensional logic; and the suggestion (by several logicians and linguists) that what linguists call "deep structure" is to be identified with logical form. Of a much less controversial nature is the extensive and fruitful use of recursive function theory and related areas of logic in formal grammars and in the formal models of language users.

Psychology. Although the "laws of thought" studied in logic are not the empirical generalizations of a psychologist, they can serve as a conceptual framework for psychological theorizing. Probably the best known recent example of such theorizing is the large-scale attempt made in the mid-20th century by Jean Piaget, a Swiss psychologist, to characterize the developmental stages of a child's thought by reference to the logical structures that he can master.

Elsewhere in psychology, logic is employed mostly as an ingredient of various models using mathematical ideas or ideas drawn from such areas as automata or information theory. Large-scale direct uses are rare, however, partly because of the problems mentioned above in the section on logic and information.

Law. Of the great variety of kinds of argumentation used in the law, some are persuasive rather than strictly logical, and others exemplify different procedures in applied logic rather than the formulas of pure logic. Examinations of "Lawiers Logike"—as the subject was called in 1588—have also uncovered a variety of arguments belonging to the various departments of logic mentioned above. Such inquiries do not seem to catch the most characteristic kinds of legal conceptualization, however—with one exception, viz., a theory developed by Wesley Newcomb Hohfeld, a pre-World War I U.S. legal scholar, of what he called the fundamental legal conceptions. Although originally presented in informal terms, this theory is closely related to recent deontic logic (in some cases in combination with suitable causal notions). Even some of the apparent difficulties are shared by the two approaches: the deontic logician's notion of permission, for example, which is often thought of as being unduly weak, is to all practical purposes a generalization of Hohfeld's concept of privilege.

Hohfeld's theory and deontic logic

Education. After having been one of the main ingredients of the general school curriculum for centuries, logic virtually disappeared from liberal education during the first half of the 20th century. It has made major inroads back into school curricula, however, as a part of the new mathematical curriculum that came into fairly general use in the 1960s, which normally includes the elements of propositional logic and of set theory. Logic is also easily adapted to being taught by computers and has been used in experiments with computer-based education.

(K.J.Hi.)

The nature of mathematics

A chief point of interest that has emerged from modern attempts to characterize philosophy is the importance of distinguishing dialectical or analytical inquiries about meaning from empirical inquiries about fact. A primary, traditional task of the philosopher has been to present things in such a light that human feelings may be reasonably grounded. The need for this is especially obvious in the case of the moral philosopher or the aesthetician, whose work treats explicitly of subjective concerns. But the need remains the same in all philosophical inquiries, including even discussions of the foundations of mathematics. For it is obvious to a careful observer that persons who put forward theses about the nature of mathematics are involved not just intellectually but also emotionally in their pursuits; while it must be supposed that in some cases this involvement stems from or inevitably leads to intellectual confusion, it must also be allowed that a certain emotional commitment may perhaps be a necessary condition for the making of discoveries. Thus it can scarcely be an accident that no great mathematician has ever accepted the conventionalist view according to which mathematical truths are man-made.

The inquiry into the nature, underlying assumptions, and scope of mathematics has emerged in the 20th century as a subdiscipline of mathematics itself, known as the study of foundations. For a full historical treatment of this field, see the *Macropædia* article MATHEMATICS, THE FOUNDATIONS OF. The material below, edited from an article originally written by Alfred North Whitehead for the 11th edition of the *Encyclopædia Britannica,* treats mathematics itself as an object of philosophical investigation.

A cursory definition of mathematics

It has been usual to define mathematics as "the science of discrete and continuous magnitude." Even Leibnitz, who initiated a more modern point of view, follows the tradition in thus confining the scope of mathematics properly so called, while apparently conceiving it as a department of a yet wider science of reasoning. A short consideration of some leading topics of the science will exemplify both the plausibility and inadequacy of the above definition. Arithmetic, algebra, and the infinitesimal calculus, are sciences directly concerned with integral numbers, rational (or fractional) numbers, and real numbers generally, which include incommensurable numbers. It would seem that

"the general theory of discrete and continuous quantity" is the exact description of the topics of these sciences. Furthermore, can we not complete the circle of the mathematical sciences by adding geometry? Now geometry deals with points, lines, planes, and cubic contents. Of these all except points are quantities. Also, as the Cartesian geometry shows, all the relations between points are expressible in terms of geometric quantities. Accordingly, at first sight it seems reasonable to define geometry in some such way as "the science of dimensional quantity." Thus every subdivision of mathematical science would appear to deal with quantity, and the definition of mathematics as "the science of quantity" would appear to be justified. We have now to see why the definition is inadequate.

CRITICAL QUESTIONS

Types relating to numbers. What are numbers? We can talk of five apples and 10 pears. But what are "five" and "10" apart from the apples and pears? Also in addition to the cardinal numbers there are the ordinal numbers: the fifth apple and the 10th pear claim thought. What is the relation of "the fifth" and "10th" to "five" and "10"? "The first rose of summer" and "the last rose of summer" are parallel phrases, yet one explicitly introduces an ordinal number and the other does not. Again, "half a foot" and "half a pound" are easily defined. But in what sense is there "a half," which is the same for "half a foot" as "half a pound"? Furthermore, incommensurable numbers are defined as the limits arrived at as the result of certain procedures with rational numbers. But how do we know that there is anything to reach? We must know that $\sqrt{2}$ exists before we can prove that any procedure will reach it.

Types relating to geometry. Also in geometry, what is a point? The straightness of a straight line and the planeness of a plane require consideration. Furthermore, "congruence" is a difficulty. For when a triangle "moves," the points do not move with it. So what is it that keeps unaltered in the moving triangle? Thus the whole method of measurement in geometry as described in the elementary textbooks and the older treatises is obscure to the last degree. Lastly, what are "dimensions"? All these topics require thorough discussion before we can rest content with the definition of mathematics as the general science of magnitude; and by the time they are discussed the definition has evaporated. An outline of the modern answers to questions such as the above will now be given. A critical defense of them would require a volume.

Nature of cardinal numbers. A one–one relation between the members of two classes α and β is any method of correlating all the members of α to all the members of β, so that any member of α has one and only one correlate in β, and any member of β has one and only one correlate in α. Two classes between which a one-one relation exists have the same cardinal number and are called cardinally similar; and the cardinal number of the class α is a certain class whose members are themselves classes—namely, it is the class composed of all those classes for which a one-one correlation with α exists. Thus the cardinal number of α is itself a class, and furthermore α is a member of it. For a one-one relation can be established between the members of α and α by the simple process of correlating each member of α with itself. Thus the cardinal number one is the class of unit classes, the cardinal number two is the class of doublets, and so on. Also a unit class is any class with the property that it possesses a member x such that, if y is any member of the class, then x and y are identical. A doublet is any class which possesses a member x such that the modified class formed by all the other members except x is a unit class. And so on for all the finite cardinals, which are thus defined successively. The cardinal number zero is the class of classes with no members; but there is only one such class, namely—the null class. Thus this cardinal number has only one member. The operations of addition and multiplication of two given cardinal numbers can be defined by taking two classes α and β, satisfying the conditions (1) that their cardinal numbers are respectively the given numbers, and (2) that they contain no member in common, and then by defining by reference to α and β two other suitable classes whose cardinal numbers are defined to be respectively the required sum and product of the cardinal numbers in question.

Cardinal numbers as classes

With these definitions it is now possible to *prove* the following six premises applying to finite cardinal numbers, from which Peano has shown that all arithmetic can be deduced:—

i. Cardinal numbers form a class. ii. Zero is a cardinal number. iii. If a is a cardinal number, $a+1$ is a cardinal number. iv. If s is any class and zero is a member of it, also if when x is a cardinal number and a member of s, also $x+1$ is a member of s, then the whole class of cardinal numbers is contained in s. v. If a and b are cardinal numbers, and $a+1=b+1$, then $a=b$. vi. If a is a cardinal number, then $a+1=0$.

It may be noticed that (iv.) is the familiar principle of mathematical induction. Peano in a historical note refers its first explicit employment, although without a general enunciation, to Maurolycus in his work, *Arithmeticorum libri duo* (Venice, 1575).

But now the difficulty of confining mathematics to being the science of number and quantity is immediately apparent. For there is no self-contained science of cardinal numbers. The proof of the six premises requires an elaborate investigation into the general properties of classes and relations that can be deduced by the strictest reasoning from our ultimate logical principles. Also it is purely arbitrary to erect the consequences of these six principles into a separate science. They are excellent principles of the highest value, but they are in no sense the necessary premises that must be proved before any other propositions of cardinal numbers can be established. On the contrary, the premises of arithmetic can be put in other forms, and, furthermore, an indefinite number of propositions of arithmetic can be proved directly from logical principles without mentioning them. Thus, while arithmetic may be defined as that branch of deductive reasoning concerning classes and relations that is concerned with the establishment of propositions concerning cardinal numbers, the introduction of cardinal numbers makes no great break in this general science. It is merely a subdivision in a general theory.

Nature of ordinal numbers. We must first understand what is meant by "order," that is, by "serial arrangement." An order of a set of things is to be sought in that relation holding between members of the set that constitutes that order. The set viewed as a class has many orders. Thus the telegraph posts along a certain road have a space-order very obvious to our senses; but they have also a time-order according to dates of erection, perhaps more important to the postal authorities who replace them after fixed intervals. A set of cardinal numbers has an order of magnitude, often called *the* order of the set because of its insistent obviousness to us; if they are the numbers drawn in a lottery, their time-order of occurrence in that drawing also ranges them in an order of some importance. Thus the order is defined by the "serial" relation. A relation (R) is serial when (1) it implies diversity, so that, if x has the relation R to y, x is diverse from y; (2) it is transitive, so that if x has the relation R to y, and y to z, then x has the relation R to z; (3) it has the property of connexity, so that if x and y are things to which any things bear the relation R, or which bear the relation R to any things, then *either* x is identical with y, *or* x has the relation R to y, *or* y has the relation R to x. These conditions are necessary and sufficient to secure that our ordinary ideas of "preceding" and "succeeding" hold in respect to the relation R. The "field" of the relation R is the class of things ranged in order by it. Two relations R and R' are said to be ordinally similar, if a one-one relation holds between the members of the two fields of R and R', such that if x and y are any two members of the field of R, such that x has the relation R to y, and if x' and y' are the correlates in the field of R' of x and y, then in all such cases x' has the relation R' to y', and conversely, interchanging the dashes on the letters; *i.e.*, R and R', x and x', etc. It is evident that the ordinal similarity of two relations implies the cardinal similarity of their fields, but not conversely. Also, two relations need not be serial in order to be ordinally similar; but if one is serial, so is the other.

Serial relations

The relationship-number of a relation is the class whose members are all those relations that are ordinarily similar to it. This class will include the original relation itself. The relation-number of a relation should be compared with the cardinal number of a class. When a relation is serial its relation-number is often called its serial type. The addition and multiplication of two relation-numbers is defined by taking two relations R and S, such that (1) their fields have no terms in common; (2) their relation-numbers are the two relation-numbers in question, and then by defining by reference to R and S two other suitable relations whose relation-numbers are defined to be respectively the sum and product of the relation-numbers in question. We need not consider the details of this process. Now if n be any finite cardinal number, it can be proved that the class of those serial relations, which have a field whose cardinal number is n, is a relation-number. This relation-number is the ordinal number corresponding to n; let it be symbolized by $\dot{n}$. Thus, corresponding to the cardinal numbers 2, 3, 4 . . . there are the ordinal numbers $\dot{2}$, $\dot{3}$, $\dot{4}$. . . . The definition of the ordinal number 1 requires some little ingenuity owing to the fact that no serial relation can have a field whose cardinal number is 1; but we must omit here the explanation of the process. The ordinal number 0 is the class whose sole member is the null relation—that is, the relation that never holds between any pair of entities. The definitions of the finite ordinals can be expressed without use of the corresponding cardinals, so there is no essential priority of cardinals to ordinals. Here also it can be seen that the science of the finite ordinals is merely a subdivision of the general theory of classes and relations.

Cantor's infinite numbers. Owing to the correspondence between the finite cardinals and the finite ordinals, the propositions of cardinal arithmetic and ordinal arithmetic correspond point by point. But the definition of the cardinal number of a class applies when the class is not finite, and it can be proved that there are different infinite cardinal numbers, and that there is a least infinite cardinal, now usually denoted by $\aleph_0$, where $\aleph$ is the Hebrew letter aleph. Similarly, a class of serial relations, called *well-ordered* serial relations, can be defined, such that their corresponding relation-numbers include the ordinary finite ordinals, but also include relation-numbers which have many properties like those of the finite ordinals, though the fields of the relations belonging to them are not finite. These relation-numbers are the infinite ordinal numbers. The arithmetic of the infinite cardinals does not correspond to that of the infinite ordinals. It will suffice to mention here that Peano's fourth premise of arithmetic does not hold for infinite cardinals or for infinite ordinals. Contrasting the above definitions of number, cardinals and ordinals, with the alternative theory that number is an ultimate idea incapable of definition, we find that our procedure exacts greater attention and less credulity.

Rational and real numbers

The data of analysis. Rational numbers and real numbers in general can now be defined according to the same general method. If m and n are finite cardinal numbers, the rational number m/n is the relation that any finite cardinal number x bears to any finite cardinal number y when $n \times x = m \times y$. Thus the rational number one, which we will denote by 1_r, is not the cardinal number 1; for 1_r is the relation 1/1 as defined above, and is thus a relation holding between certain pairs of cardinals. Similarly, the other rational integers must be distinguished from the corresponding cardinals. The arithmetic of rational numbers is now established by means of appropriate definitions, which indicate the entities meant by the operations of addition and multiplication. But in order to obtain general enunciations of theorems without exceptional cases, mathematicians employ entities of ever-ascending types of elaboration. These entities are not created but are employed by mathematicians, and their definitions should show the construction of the new entities in terms of the old. The real numbers, including irrational numbers, have now to be defined. Consider the serial arrangement of the rationals in their order of magnitude. A real number is a class (α, say) of rational numbers that satisfies the condition that it is the same as the class of those rationals each of which precedes at least one member of α. Thus, consider the class of rationals less than 2_r; any member of this class precedes some other members of the class—thus 1/2 precedes 4/3, 3/2 and so on; also the class of predecessors of predecessors of 2_r is itself the class of predecessors of 2_r. Accordingly this class is a real number; it will be called the real number 2_R. Note that the class of rationals less than or equal to 2_r is not a real number. For 2_r is not a predecessor of some member of the class. In the above example 2_R is an integral real number, which is distinct from a rational integer, and from a cardinal number. Similarly, any rational real number is distinct from the corresponding rational number. But now the irrational real numbers have all made their appearance. For example, the class of rationals whose squares are less than 2_r satisfies the definition of a real number; it is the real number $\sqrt{2}$. The arithmetic of real numbers follows from appropriate definitions of the operations of addition and multiplication. Except for the immediate purposes of an explanation, such as the above, it is unnecessary for mathematicians to have separate symbols, such as 2_r, 2_n and 2_R, or $2/3$ and $(2/3)_R$. Real numbers with signs (+ or −) are now defined. If a is a real number, $+a$ is defined to be the relation that any real number of the form $x + a$ bears to the real number x, and $-a$ is the relation that any real number x bears to the real number $x + a$. The addition and multiplication of these "signed" real numbers is suitably defined, and it is proved that the usual arithmetic of such numbers follows. Finally, we reach a complex number of the nth order. Such a number is a "one-many" relation which relates n signed real numbers (or n algebraic complex numbers when they are already defined by this procedure) to the n cardinal numbers $1, 2, \ldots n$ respectively. If such a complex number is written (as usual) in the form $x_1e_1 + x_2e_2 + \ldots + x_ne_n$, then this particular complex number relates x_1 to 1, x_2 to $2, \ldots x_n$ to n. Also the "unit" e_1 (or e_2) considered as a number of the system is merely a shortened form for the complex number $(+1)e_1 + 0e_2 \ldots + 0e_n$. This last number exemplifies the fact that one signed real number, such as 0, may be correlated to many of the n cardinals, such as $2 \ldots n$ in the example, but that each cardinal is only correlated with one signed number. Hence the relation has been called above "one-many." The sum of two complex numbers $x_1e_1 + x_2e_2 + \ldots + x_ne_n$ and $y_1e_1 + y_2e_2 + \ldots + y_ne_n$ is always defined to be the complex number $(x_1 + y_1)e_1 + (x_2 + y_2)e_2 + \ldots + (x_n + y_n)e_n$. But an indefinite number of definitions of the product of two complex numbers yield interesting results. Each definition gives rise to a corresponding algebra of higher complex numbers. We will confine ourselves here to algebraic complex numbers—that is, to complex numbers of the second order taken in connection with that definition of multiplication that leads to ordinary algebra. The product of two complex numbers of the second order—namely, $x_1e_1 + x_2e_2$ and $y_1e_1 + y_2e_2$, is in this case defined to mean the complex $(x_1y_1 + x_2y_2)\,e_1 + (x_1y_2 + x_2y_1)e_2$. Thus $e_1 \times e_1 = e_1$, $e_2 \times e_2 = -e_1$, $e_1 \times e_2 = e_2 \times e_1 = e_2$. With this definition it is usual to omit the first symbol e_1, and to write i or $\sqrt{-1}$ instead of e_2. Accordingly, the typical form for such a complex number is $x + yi$, and then with this notation the above-mentioned definition of multiplication is invariably adopted. The importance of this algebra arises from the fact that in terms of such complex numbers with this definition of multiplication the utmost generality of expression, to the exclusion of exceptional cases, can be obtained for theorems that occur in analogous forms, but complicated with exceptional cases, in the algebras of real numbers and of signed real numbers. This is exactly the same reason as that which has led mathematicians to work with signed real numbers in preference to real numbers, and with real numbers in preference to rational numbers.

DEFINITION OF MATHEMATICS

It has now become apparent that the traditional field of mathematics in the province of discrete and continuous number can only be separated from the general abstract theory of classes and relations by a wavering and inde-

terminate line. Of course a discussion as to the mere application of a word degenerates into the most fruitless logomachy. But on the assumption that "mathematics" is to denote a science well marked out by its subject matter and its methods, and that at least it is to include all topics habitually assigned to it, "mathematics" is employed in the general sense of the "science concerned with the logical deduction of consequences from the general premises of all reasoning."

Kinds of mathematical propositions

Geometry. The typical mathematical proposition is: "If $x, y, z \ldots$ satisfy such and such conditions, then such and such other conditions hold with respect to them." By taking fixed conditions for the hypothesis of such a proposition a definite department of mathematics is marked out. For example, geometry is such a department. The "axioms" of geometry are the fixed conditions that occur in the hypotheses of the geometrical propositions. It is sufficient to observe here that they are concerned with special types of classes of classes and of classes of relations, and that the connection of geometry with number and magnitude is in no way an essential part of the foundation of the science.

Classes and relations. We now must deduce the general properties of classes and relations from the ultimate logical premises. In the course of this process, some contradictions have become apparent. That first discovered is known as Burali-Forti's contradiction and consists in the proof that there both is and is not a greatest infinite ordinal number. But these contradictions do not depend upon any theory of number, for Russell's contradiction does not involve number in any form. This contradiction arises from considering the class possessing as members all classes that are not members of themselves. Call this class w; then to say that x is a w is equivalent to saying that x is not an x. Accordingly, to say that w is a w is equivalent to saying that w is not a w. An analogous contradiction can be found for relations. It follows that a careful scrutiny of the very idea of classes and relations is required. Note that classes are here required in extension, so that the class of human beings and the class of rational featherless bipeds are identical; similarly for relations, which are to be determined by the entities related. Now a class in respect to its components is many. In what sense then can it be one? This problem of "the one and the many" has been discussed continuously by the philosophers. All the contradictions can be avoided, and yet the use of classes and relations can be preserved as required by mathematics, and indeed by common sense, by a theory that denies to a class—or relation—existence or being in any sense in which the entities composing it—or related by it—exist. Thus, to say that a pen is an entity and the class of pens is an entity is merely a play upon the word "entity": the second sense of "entity" (if any) is indeed derived from the first but has a more complex signification. Consider an incomplete proposition, incomplete in the sense that some entity that ought to be involved in it is represented by an undetermined x, which may stand for any entity. Call it a propositional function; and, if φx be a propositional function, the undetermined variable x is the argument. Two propositional functions φx and ψx are "extensionally identical" if any determination of x in φx that converts φx into a true proposition also converts ψx into a true proposition, and conversely for ψ and φ. Now consider a propositional function $F\chi$ in which the variable argument χ is itself a propositional function. If $F\chi$ is true when, and only when, χ is determined to be either φ or some other propositional function extensionally equivalent to φ, then the proposition $F\varphi$ is of the form which is ordinarily recognized as being about the class determined by φx taken in extension—that is, the class of entities for which φx is a true proposition when x is determined to be any one of them.

Classes of classes and of relations

A similar theory holds for relations that arise from the consideration of propositional functions with two or more variable arguments. It is then possible to define by a parallel elaboration what is meant by classes of classes, classes of relations, relations between classes, and so on. Accordingly, the number of a class of relations can be defined, or of a class of classes, and so on. This theory is in effect a theory of the *use* of classes and relations and does not decide the philosophic question as to the sense (if any) in which a class in extension is one entity. It does indeed deny that it is an entity in the sense in which one of its members is an entity. Accordingly, it is a fallacy for any determination of x *to* consider "x is an x" or "x is not an x" as having the meaning of propositions. Note that for any determination of x, "x is an x" and "x is not an x" are neither of them fallacies but are both meaningless, according to this theory. Thus Russell's contradiction vanishes, and the other contradictions vanish also.

APPLIED MATHEMATICS

Selection of topics. The selection of the topics of mathematical inquiry among the infinite variety open to it has been guided by the useful applications, and indeed the abstract theory has only recently been disentangled from the empirical elements connected with these applications. For example, the application of the theory of cardinal numbers to classes of physical entities involves in practice some process of counting. It is only recently that the *succession* of processes that is involved in any act of counting has been seen to be irrelevant to the idea of number. Indeed, it is only by experience that we can know that any definite process of counting will give the true cardinal number of some class of entities. It is perfectly possible to imagine a universe in which any act of counting by a being in it annihilates some members of the class counted during the time and only during the time of its continuance. A legend of the Council of Nicaea illustrates this point: "When the Bishops took their places on their thrones, they were 318; when they rose up to be called over, it appeared that they were 319; so that they never could make the number come right, and whenever they approached the last of the series, he immediately turned into the likeness of his next neighbour." Such a story cannot be disproved by deductive reasoning from the premises of abstract logic. We can only assert that a universe in which such things are liable to happen on a large scale is unfitted for the practical application of the theory of cardinal numbers.

Relation of theory and physical reality

The application of the theory of real numbers to physical quantities involves analogous considerations. In the first place, some physical process of addition is presupposed, involving some inductively inferred law of permanence during that process. Thus in the theory of masses we must know that two pounds of lead when put together will counterbalance in the scales two pounds of sugar, or a pound of lead and a pound of sugar. Furthermore, the sort of continuity of the series (in order of magnitude) of rational numbers is known to be different from that of the series of real numbers. Indeed, mathematicians now reserve "continuity" as the term for the latter kind of continuity; the mere property of having an infinite number of terms between any two terms is called "compactness." The compactness of the series of rational numbers is consistent with quasi-gaps in it—that is, with the possible absence of limits to classes in it. Thus the class of rational numbers whose squares are less than 2 has no upper limit among the rational numbers. But among the real numbers all classes have limits. Now, owing to the necessary inexactness of measurement, it is impossible to discriminate directly whether any kind of continuous physical quantity possesses the compactness of the series of rationals or the continuity of the series of real numbers. In calculations the latter hypothesis is made because of its mathematical simplicity. But the assumption has certainly no a priori grounds in its favour, and it is not very easy to see how to base it upon experience. For example, the continuity of space apparently rests upon sheer assumption unsupported by any a priori or experimental grounds. Thus the current application of mathematics to the analysis of phenomena can be justified by no a priori necessity.

Existence of applied mathematics. In one sense there is no science of applied mathematics. When once the fixed conditions that any hypothetical group of entities are to satisfy have been precisely formulated, the deduction of the further propositions, which also will hold respecting them, can proceed in complete independence of the question as to whether or not any such group of entities

can be found in the world of phenomena. Thus rational mechanics, based on the Newtonian Laws and viewed as mathematics, is independent of its supposed application, and hydrodynamics remains a coherent and respected science though it is extremely improbable that any perfect fluid exists in the physical world. But this unbendingly logical point of view cannot be the last word upon the matter. For no one can doubt the essential difference between characteristic treatises upon "pure" and "applied" mathematics. The difference is a difference in method. In pure mathematics the hypotheses that a set of entities are to satisfy are given, and a group of interesting deductions are sought. In "applied mathematics" the "deductions" are given in the shape of the experimental evidence of natural science, and the hypotheses from which the "deductions" can be deduced are sought. Accordingly, every treatise on applied mathematics, properly so-called, is directed to the criticism of the "laws" from which the reasoning starts, or to a suggestion of results that experiment may hope to find. Thus if it calculates the result of some experiment, it is not the experimentalist's well-attested results that are on their trial but the basis of the calculation. (A.N.W.)

Western philosophy of law

Philosophy of law is concerned with the formulation of concepts and theories to aid in understanding the nature of law, the sources of its authority, and its role in society. In English-speaking countries the term "jurisprudence" is often used synonymously and is invariably used in reference to particular subdivisions of the field.

To the extent that it implies some necessary link or coincidence with general philosophy, the phrase "philosophy of law" may be somewhat misleading, for philosophy of law is mostly untouched by the conflicts of different philosophical schools, and its practitioners may without incongruity draw on diverse philosophical outlooks—on Logical Positivism for some analytical problem of the structure of legal orders and, simultaneously, on Existentialism for a problem of sociological jurisprudence or justice, for example—without commitment in any such instance to an entire philosophical outlook.

Nor can one treat philosophy of law as a specialized branch of philosophy such as ethical or political philosophy, epistemology, or logic, for in philosophy of law all these branches may make contributions. Ideas that may illumine jurisprudential problems must indeed be sought not only in philosophy but in all systematic bodies of thought. Only if "philosophy" is interpreted in its least technical and broadest sense does "philosophy of law" cease to be a misnomer.

PROBLEMS OF THE PHILOSOPHY OF LAW

Three major subdivisions of philosophy of law

Various approaches. For practical reasons, such as to avoid overlappings, it is convenient to organize jurisprudence into three principal branches only: analytical jurisprudence, sociological jurisprudence, and the theory of justice.

Analytical jurisprudence. The analytical questions in jurisprudence are concerned with articulating the axioms, defining the terms, and prescribing the methods that best enable one to view the legal order (or part of it) as a self-consistent system and that maximize awareness of its logical structure. Perhaps the most rigorous solutions are those which, like that of Hans Kelsen, a contemporary Austrian–American legal philosopher, attempt to identify structural or relational features as being necessarily entailed in the meaning of legal norms or in lawyers' intellectual operations with them (see below *Pure theory of law*). Alternatively, the basis for logical structuring may be found in some imputed attribute of law not itself inherently structural. The 19th-century English legal philosopher John Austin, for example, thought it an essential preliminary to his quest for a logical system in law to clarify what was involved in his assumption that law always consists of "commands." This clarification is important, but the claim that such a clarified version of a common assumption necessarily amounts to an analytical model of law seems unwarranted.

On more modest levels, the analyst may seek to infuse clarity and orderliness into some particular branch of a legal system or even into the applications of some particular rule. Such work shades over, on the one hand, into ordinary legal analysis and, on the other, into jurisprudential efforts to clarify the meaning of particular legal terms. Analysis of a particular word or even a particular branch of law will usually, by necessity, be particularist in the sense that it works upon legal materials found only in one particular legal system. In between particularism and universalism lie what are sometimes called comparative approaches, in which analytical jurisprudence is applied to materials drawn from more than one (but not from all) legal systems.

Sociological jurisprudence. The sociological questions in jurisprudence are concerned with the actual effects of the law upon the complex of attitudes, behaviour, organization, environment, skills, and powers involved in the maintenance of a particular society. Conversely, sociological jurisprudence is also concerned with the effects of social phenomena on both the substantive and procedural aspects of law, as well as on the legislative, judicial, and other means of forming, operating, changing, and disrupting the legal order. The fact that men in a given time and place hold particular ideas and values, including ideals of justice, is itself a fact the relation of which to law must be studied; but the focus is sharply different from that in the study of theories of justice. Its focus is descriptive, not normative; it is concerned with what is or with what goes on, not with what ought to be or ought to go on.

The theory of justice. The theory of justice is concerned with the evaluation and criticism of law in terms of the ideals or goals postulated for it. This involves the identification and articulation of the values that the legal order seeks to realize. This aspect of jurisprudence is inextricably interwoven with ethical and political philosophy, and theories of justice thus tend to parallel the full range of ethical and political philosophies.

Law, morality, and natural law. A consideration of fundamental importance in the philosophy of law is that of the distinction between law and morality. The importance of the distinction is illustrated by the main questions to which it gives rise: (1) How far and in what sense should the law of a community seek to give effect to its morality? (2) Is there a moral duty to obey the law even when it does not embody morality, and, if so, are there any limits to this duty? (3) When a legal rule directs conduct that morality forbids, which should the citizen obey? (4) Is there ever (and, if so, when is there) a duty to overthrow an entire legal system because of its conflict with morality?

In all these questions the word "law" refers to the specialized form of social control familiar in modern, secular, politically organized societies. The word "morality" in the four questions may, however, refer to any of the following: (1) the community's relevant factual behaviour patterns (its mores); (2) its socially approved behaviour patterns, as sanctified by some widely held rational or religious ideal, whether observed in practice or not (social morality); or (3) the moral ideals accepted by each individual as binding on himself and on others, whether or not those others agree (individual morality). All these, like law, are means of controlling human conduct by setting normative standards; and all three have a constantly changing interaction with each other, as well as with law.

Definitions of morality

The fact that legal and moral norms vary from place to place and from one historical period to another lies in part behind a persistent theme in the philosophy of law: the search for unchanging norms that are universally valid. Clearly, the most certain way of establishing such norms would be to base them on widely observed facts, such as man's social propensities or the ubiquitous importance of kinship in social organization, which supposedly reveal something fundamental about the nature of man and his adjustment to the world. The attempt to base norms on some such category of facts has for two millennia been associated with the concept of natural law. This concept has many versions, the principal of which are outlined in the historical survey below, but the significance of the topic merits some separate preliminary discussions.

Aristotle's conception of natural law

It has always been possible to trace a mainstream of natural-law thought, flowing from Aristotle's premise that the "nature" of any creature, from which obligations must be derived, is what it will be in its fullest and most perfect development. For man, this means what he is when the powers and qualities distinguishing him from other creatures, namely, his reason and his impulse to social living, are fully developed. Natural law embodies those obligations that will appear if mankind's reason and sociality are fully unfolded.

A major difficulty presented by this attempt to develop normative standards appears to be that it is very difficult to demonstrate, let alone create a sense of obligation toward, values that are only immanent. All theories of natural law, moreover, have found it necessary to rely on what are essentially intuitions or preconceptions as to what man's true nature is. All such theories acknowledge, for instance, that the full development or fulfillment of an entity is not the same as its mere continued existence, that there may be a "warping" or "impeding" of the natural tendencies, so that what exists may then "be said to be unsound or incorrect." Thus, mere factuality is not a sufficient source of obligation. Similarly, St. Thomas Aquinas himself, in identifying the "inclinations" from which men may learn natural law, found it necessary to order these in grades of inclination, so that those inclinations most closely related to reason and sociality take priority over those concerned (for example) with procreation and self-preservation. The criteria by which such a hierarchy is ordered must be drawn from sources other than the factual inclinations themselves. The "lower" grades (such as self-preservation) may well be based on something like instinct; but the question arises at the higher grades whether there is any comparable instinct by which men seek to find moral precepts binding all of them in common. Aquinas here appealed to synderesis, a kind of sympathetic understanding found in men, a disposition (habit) of the practical intellect inclining them to the good and murmuring against evil.

To derive from this synderesis a universal natural law, however, it would be necessary to demonstrate some "universal conscience" of all mankind. But natural lawyers faced with the fact that men's consciences do not coincide explain that conscience may err and reason be corrupt. Invocation of synderesis is in fact helpful not as an account of how one may arrive at factually based normative standards but as an illustration of the psychological tendency of men to assert values.

HISTORICAL SURVEY OF LEGAL THEORIES

The ancient world. *Greek thought.* The major contribution of Greece was a body of philosophical and cosmological ideals about justice, more apt for orators' appeals to popular assemblies than for preceptual application to day-to-day life situations.

Early Greek cosmologies, embedded in some of the earliest myths, had seen the individual as held within a kind of transcending harmony of the universe, emanating from the divine law (*logos*) and expressed in relation to human life in the law (*nomos*) of the polis, the city-state. The later Sophists, however, who examined critically all assumptions relating to life in the city-state, pointed to the wide disparities in human law and morals and rejected the claim that this human law (*nomos*) necessarily reflected any universal law (*logos*). Taking man as "the measure of all things," they rejected any claims of his law (*nomos*) to absolute value and saw law and justice and values generally as created by men's reasons, in their multitudes and generations, in all their individuated, relativistic, and historically changing dimensions.

Plato's idea of law

In the restless intellectual and political climate of 5th-century Athens, Plato was concerned to redefine the nature of justice by relating it to something far more permanent and absolute than the *nomos* of the city-state. He assigned "reality" to the unchanging archetypal forms—*i.e.,* the ideas—of things rather than to the ephemeral phenomena as superficially and confusedly perceived by individual men unenlightened by philosophy. In the utopia described in *The Republic,* Plato defines justice in an architectonic sense: justice prevails when the state is ordered in accordance with the ideal forms ascertained by its philosopher-kings and is thus unrelated to the *nomos* of the city-state. There is no need for human law, since transcendental knowledge rules. In his later thought, however, as revealed in *Politicus* (the "Statesman") and the *Laws,* when he is concerned to describe a more practicable but nevertheless "second best" state, Plato assigns to law a role almost as important as that of knowledge in *The Republic.* A famous classification of states given in *Politicus* is indeed based on the criterion of whether or not they are ruled by law. The law as Plato here conceived it, however, was not mere convention or the imperfect individual judgments of men but a reflection of the common human reason in its full development. To this extent the rule of law might approximate the ideal rule of knowledge envisaged in *The Republic,* for in the inherited law of men is crystallized that much wisdom of which they are capable.

And yet it was difficult for Plato to find justification for such an argument in his basic philosophical position, with its emphasis on the contrast between the mere opinion of ordinary men and the transcendental knowledge of the philosopher. Aristotle, who in common with Plato held a view of nature or reality that transcended the variability of things as perceived by the senses, was, however, able more successfully to defend the validity of a law resulting from the practice of ordinary men. For Aristotle's transcendental reality is more firmly related to things as they are: it comprises that which they will become as their potentialities unfold in nature toward the end that is theirs in nature. Man, in his nature, is moral, rational, and social, and his law may be judged by the extent to which it facilitates the development of these innate qualities.

The Stoic conception of natural law

The Greek conception of natural law underwent further refinement by the Stoic school of philosophy, which became active toward the end of the 4th century BC. The Stoics posited the existence of a natural law, the *jus naturale,* which was an emanation of the *lex aeterna,* the law of reason of the cosmos. The existence of an innate reason in men linked everyone with the cosmic order and subjected all to a universally valid moral law. This latter concept thoroughly infused Roman thinking, largely as a result of the influence of Stoic philosophy on Rome.

Roman thought. Greek law scarcely survived as a system, because it never developed a class of legal specialists or abandoned its lay administrators or its popular tribunals of grotesque size. Roman law, on the other hand, developed through the efforts of expert jurisconsults (learned lawyers) and praetors (magistrates) into a permanent heritage of Western society. By its adoption into works such as Cicero's *De republica* as well as in the work of the great jurisconsults, Stoic speculation concerning reason and nature was brought onto the level of precepts for concrete problem solving. The crude, tribal *jus civile* ("civil law") of the Romans was thus transformed into a natural-law-based *jus gentium* (law applying to all people), a set of principles common to all nations and appropriate, therefore, for application to foreigners as well as Romans.

Hebrew thought. In the Talmud there is an assertion that "Whatever decision of a mature scholar in the presence of his teacher will yet derive from the Law (Torah) that was already spoken to Moses on Mt. Sinai." In theory, this presupposed that the Oral Law must respect every jot and tittle of the revealed written law. Yet the richness, ambivalences, and silences of what was written, in relation to a changing world, still left the widest freedom to the scholarly reason of the rabbinical exegetes into whose care both the written law and the Oral Law finally came.

The operations of the rabbinical schools and courts over many creative centuries, especially during and following the first Babylonian Exile, resembled those of the great Roman jurisconsults and the great judges of the common-law tradition. One Talmudic story tells of a doctrinal rift between the majority of a rabbinical court led by a great rabbi and a dissenting but no less great rabbi, in which the dissenter successfully summoned the authentic voice of God onto his side of the argument. To this intervention the majority of the court responded: "The law is not in heaven, the law has been handed down to us on earth from Mt. Sinai, and we no longer take notice of

heavenly voices. . . ." And the story relates that, at that point, God said with a smile to Elijah the prophet, with whom he was walking: "My children have defeated me, my children have defeated me." But this was an indulgent ratification, not an implacably cruel wrath such as the Greek god brought down on the head of Prometheus. Thus, even against divine intervention, the learned stood their ground, relying for the interpretation of the law on their own wisdom and reason.

The Middle Ages. *Augustine.* St. Augustine of Hippo, in attempting to refute the pagan assertion that Christianity was responsible for the decline of Roman power, reintroduced Stoic philosophy alongside Judeo-Christian thought into the stream of modern jurisprudential speculation. He placed God's reason beside God's will as the highest source of the unchangeable, eternal, divine law binding directly on man and all other creatures. The divine law was thus accessible to both man's reason and his faith and was not, as St. Paul had largely concluded, the product of his will alone and hence not rational in terms of human as opposed to divine reason.

At a second level, Augustine placed the no less unchangeable natural law, being the divine law as man is given the reason, heart, and soul to understand it. The third level, of temporal, or positive, law (for him, the Roman law of the Christian Roman Empire), was warranted by the eternal divine law, even though it changed from time to time and from place to place, so long as it respected the limits laid down by the divine and natural law. This rationale of secular power, some have thought, preserved the idea of government under law through the disintegration of the ancient world, for recultivation in the revival of learning of the 12th and 13th centuries.

Scholasticism. Aquinas, like Augustine long before, succeeded in quieting momentarily the competing claims of the will against the reason of God, the struggle between "voluntarism" and "rationalism," as the underlying basis of the eternal and natural law. Aquinas, like Augustine, gave a plausible place to both natural law and temporal (or positive) law under the eternal law. Human, or positive, law is a creation of human reason for the common good, within limits that natural law prescribes, so that even this proceeds from right reason and therefore from the eternal law. Such positive law as violated the natural and thus the eternal law "was not law" or merely was not binding "in conscience."

Voluntarism of Duns Scotus

The tendency to make reason prevail over will (as in Plato's call for philosophers to be kings or the Arab Averroës' call for philosophers to interpret what is revealed) was challenged by a voluntarist countermovement at Paris and Oxford in the quarter of a century after Aquinas' death in 1274. A Franciscan, John Duns Scotus, insisted on the uniqueness of all beings as finally traceable to the uniqueness of God's will. All precepts, even of the divine law, depend on the single precept "Love God," and, since not reason but will gives access to this, there is no natural law accessible to man's reason. All that can be required of human, or positive, law is that it must be "consonant" with the precept "Love God," or with any other precept willed by God.

The Renaissance period to the 18th century. *Machiavelli.* Machiavelli presented himself (on one interpretation, at least) as seeking to escape from both transcendent will and transcendent reason into the empirical, into life as it is, observed through the eyes of a worldly man whose mind is uncluttered with philosophical and theological preconceptions. He can be understood, in his own words, to be seeking "what a principality is, the variety of such States, how they are won, how they are held, how they are lost." This conception was the more remarkable in 1513, since such an approach had then barely been promulgated for study of the physical world. It had still, indeed, to await its major manifesto in that sphere until Francis Bacon's *Advancement of Learning* at the end of the century.

The normative element in Machiavelli's thought

Even on the more favourable view of Machiavelli's aim—*i.e.*, as describing, rather than prescribing, political behaviour—it remains true that he saw this description as ancillary to the art of maintaining the state and its ruler, so that this maintenance is a kind of end in itself. The omnipotence—unrestrained by law or morality—that he both ascribes and prescribes to the prince is thus a product not so much of his scientific detachment as of his tendency to view political power as a value, as an end in itself.

Natural law and social-contract theory. The supremacy of the human lawgiver, as posited by Machiavelli and in their diverse ways also by the French and English political theorists Jean Bodin and Thomas Hobbes and others, interwove in the following centuries with continued insistence by Grotius and others on the dominance of the divine reason and man's participation in it, by which he has access to the natural law.

The views of Hugo Grotius

The Dutch political and legal philosopher Hugo Grotius, amid the political expediencies and anarchy of the Thirty Years' War (1618–48), sought to introduce a degree of normative restraint among the monarchical rulers of the newly emerged sovereign states of Europe and to establish a basis in natural law for a rejection of *raison d'état* as a just cause for war, as well as for legal limits on the means and modes of violence in war. Even if the wills of sovereign states form the basis of the international order, Grotius argued, "the totality of the relations between States" is still "governed by law." That law he found in an updated version of the Stoic natural law, as naturalized into Roman law and Christian theology.

With Grotius, as with the Stoics, the normative or moral power of the natural law derives from the fact that man's innate nature (itself part of the nature of the cosmos) and his propensities are viewed as ideal or inherently good. In Grotius' own time, however, there arose a skepticism toward such unfounded optimism, a skepticism that underlies the thought of Hobbes.

With Hobbes (1588–1679), as with the Greek Sophists, the nature of man is not the ideal nature of Grotius and the Stoics. It is rather man's supposed actual nature, before sociality and authority have tempered it. Man, in a state of nature, is motivated by desires and aversions and most of all by the desire to preserve his biological existence. This need for security is best met by all men vesting their rights of self-help in a sovereign, whether that sovereign be a single man or an assembly of men, and subjecting themselves to the laws of that sovereign, or "great Leviathan."

The reason why men must obey the law of the sovereign state, which is the only institution capable of protecting men against each other, is thus based firmly in Hobbes's conception of man's nature, albeit a very different conception from the idealist premises of earlier theories of natural law. Natural-law theorizing after Hobbes is thus divided into these two major streams.

Judicial supremacy. By the beginning of the 17th century the idea of applying natural law as a test of the validity of the positive law (the law of the particular human jurisdiction) had passed from the province of speculative writers to courts of law. The English jurist Lord Coke, in Bonham's case in 1610, was already referring to the tradition that "when an act of Parliament is against common right or reason or repugant or impossible to be performed, the common law will control it, and adjudge such act to be void." About a century before that, an English treatise known as "St. Germain, Doctor and Student" had already presented a three-tier hierarchy of the law of God, natural law (the law of reason), and human (positive) law, obviously deriving from Augustine and Aquinas.

Natural law and American constitutionalism

In the United States in the next century, constitutional theory became highly infused with ideas of natural rights. The Declaration of Independence, with its assertion of the self-evident rights of life, liberty, and the pursuit of happiness, marked the beginning of a continuing natural-law influence on American constitutional development. The power of the judiciary to "review" legislation for consistency with a written constitution was taken in the United States to import the power to declare it void, constitutional law being analogized to natural law. Indeed, American judicial statements of 1814, 1822, and 1831 asserted the power of the judiciary to strike down statutes for violation not only of explicit constitutional restraints but also of "eternal principles of justice which no government has a right to disregard." The analogy of constitutional and natural law did not necessarily require that the power to

strike down legislation should be a judicial power; this was not so in ancient Rome, nor is it always so in modern civil-law countries. It is arguable that such a judicial repository of the power of final review is unavoidable, since the legislature cannot be expected to annul its own acts; and the executive, even if it were not a party to such acts, is scarcely equipped for the tasks of objective interpretation involved. Yet there are real difficulties of policy and principle raised by giving the judiciary the final word. A distinction must first of all be made among diverse constitutional restraints. Safeguards for such rights as free speech and assembly and access to courts, which help to assure the responsibility of rulers and to prevent the fall of democracy into tyranny or demagoguery, may well be placed in the final custody of judges. But, beyond this point, others have argued, judicial supremacy, in enforcing restraints laid down by the Founding Fathers of an earlier generation, may clearly constitute an obstacle to the implementation by the courts of a society's present convictions.

Decline of natural law. If man is the measure of all things, as the Sophists taught, then a given society of men is the measure of its culture, including its moral and legal standards. In the modern period the French jurist and political philosopher Montesquieu's *De l'esprit des lois* (1748) and *Lettres persanes* (1721) offered the thesis that a people's law and justice are determined by the particular factors and environment that operate upon them. They thus could not, as the natural-law theory of the time held, be unchanging from age to age and from people to people. The French sociologist Auguste Comte's *Système de philosophie positive* (1851–54), which set out to explain positive laws, like other social facts, by reference to verified hypotheses concerning cause and effect and interaction, was similarly antithetical to natural-law theory as it had so far developed. To Comte, metaphysical concepts about such abstractions as ideal essences belonged to a past stage in man's intellectual development. And Darwin's *On the Origin of Species by Means of Natural Selection* (1859), the English philosopher Herbert Spencer's positivism, and other related thinking of the period provided a biological model of self-development of organisms and institutions through a struggle in which survival was a function of challenge and response in the given environment. Change and adaptation, rather than constancy and inviolability, were thus at the heart of their system.

Montesquieu and early social science

Under the leadership of anthropologists, analyses of man's internal process of response to the exigencies of existence within a particular culture—to conscious and subconscious psychic drives and motivations—deeply affected the jurisprudential study of law and society and helped to bring natural-law thinking to a 19th-century nadir. In the anthropologist Bronislaw Malinowski's most mature statement on the matter, he distinguished four major meanings of the word law as important in understanding the growth of civilization. They included "laws of nature" in the scientific sense of rules governing men's conscious adaptations to the environment; rules of "efficiency" and "convenience" according to which the group lives; rules for conflict adjustment; and rules about enforcement of the last two. No conception of natural law, which had engaged earlier thinkers for two millennia and more, was included.

"Idealism" and justice. Another line of thought, which was also divorced from natural-law concepts, was contained in the Idealist philosophy of Immanuel Kant. Fundamental to Kant's ethical and jurisprudential reasoning is the premise that all moral concepts have their basis wholly in a priori thought, that they can be arrived at by reason alone, without reference to experience or recourse to intuition of rules alleged immanent in experience. Man, furthermore, is a free agent whose actions are determined by aims that he is at liberty to select. From such premises Kant deduced the nature of an ideal law, in which is implicit a theory or criterion of justice. This ideal law comprises the conditions under which all members of society can enjoy the maximum freedom from subjection to the arbitrary will of others.

Kant's ideal law

But Kant's supposedly a priori concepts are in fact as transcendental as anything natural lawyers have offered. It is thus not surprising that later thinkers, such as Johann Fichte, Kant's Idealist successor, had little difficulty in putting the new Kantian wine into natural-law bottles.

The 20th century saw a fresh attempt at the Kantian approach in the work of the German legal philosopher Rudolf Stammler. Adopting the Kantian position that knowledge is independent of sensory experience, Stammler set out to discover pre-experiential categories, or "pure forms," of thinking about law. Stammler arrived at a social ideal of a "community of free-willing men," an ideal that he claimed to have universal validity because of its supposed a priori basis. Having thus arrived at a "pure" ideal of society, untainted by empirical content deriving from sense perception, he felt able to formulate equally pure principles for just law that would regulate his ideal society. Stammler's pure idea of society comprised the harmony of individual and common purposes: his pure idea of just law thus comprised those principles conducive to such harmony—the mutual respect of individuals for each other's purposes and the participation of all in the achievement of the common purposes.

The different stream of Idealism flowing from Hegel's philosophy of history was fed into jurisprudence by Josef Kohler, Stammler's close predecessor in that subject in the Berlin University. His work is still another effort to relate social facts and the norms of justice by exposing the immanence of values in facts—in "civilization" in Kohler's case.

In perspective, these idealisms, despite their formal or philosophical antagonism to "rationalism" and natural-law thinking, seem to have reinforced in the age of the Industrial Revolution the individualist and libertarian trends that natural law had built up successively against medieval church and empire, the shackles of medieval social, political, and economic organization, and 18th-century despotism.

The 19th and 20th centuries. *Analytical Positivism.* The early 19th century witnessed a reaction against both Kantian Idealism and iusnaturalism (natural-law theorizing). The scientific temper of the age, reflected in the practical achievements of the early decades of the Industrial Revolution, was not conducive to deductive reasoning from a priori hypotheses, which appeared an impractical method of solving the problems of complex societies. Such problems might better be approached via a thorough analysis of existing law and institutions. This new climate of opinion came to be known as Positivism.

Among the chief meanings of Positivism in the legal-analytical sphere are the separation of law as it is and law as it ought to be, stress on the analysis of legal concepts, reliance on logical reasoning in the search for applicable law, and denial that moral judgments can be based on observation and rational proof. Anglo-Saxon analytical Positivism has directed itself mainly to the logical dissection, appraisal, and clarification of the precept element of law, ignoring the elements consisting of lawyers' traditional techniques and received ideals. By the nature of its tasks, analytical jurisprudence does not concern itself with either the facts surrounding or the consequences flowing from legal precepts or with their ethical evaluation, though particular analysts may also be interested in those matters.

Meaning of Positivism in jurisprudence

Analytical Positivism in England began with the work of the philosopher and legal reformer Jeremy Bentham. His work influenced John Austin, the most outstanding figure in English jurisprudence, who set out to analyze the notions pervading English law. In order to delimit his subject, he defined positive law as the commands of a sovereign addressed to political inferiors and backed by threats of evil in the event of disobedience. Positive law might well be derived from moral precepts and other sources, but such precepts become law only when commanded by a sovereign.

The analytical-Positivist attitude has continued to influence thinkers, although the particular approach of Austin is now of historical interest only. Logical analysis is clearly a tool that may be employed in many spheres of jurisprudence, and its importance thus transcends the limits of any one school. Analysis means little in itself; its value

depends largely on the validity of the premises from which the argument is made and on the relevance of the subject matter that is chosen for analysis. The paramountcy for the analytical Positivist of questions of logical order and consistency represents a permissible deviation, yet still a deviation, from the wider concerns of ordinary lawyers and students of law and society generally. This is not to say that there is a necessary conflict between the requirements of justice among changing social facts and those of logical consistency of precepts. It means only that there is surely no necessary (nor indeed usual) coincidence between them.

Historical Positivism. In discarding speculative cosmology, the a priori, and the self-evident, 19th-century historical jurisprudence opened the way for the search for the realities of law through empirical observation—for a sociodescriptive rather than a logico-analytical-Positivist jurisprudence.

The leading figure in the historical school was the German jurist Friedrich Karl von Savigny, who confronted the natural-law aspiration for a universal human code with the singularity of the law of particular peoples resulting from their unique sociocultural experiences.

For Savigny, law rests on the *Volksgeist,* or innate popular consciousness; law *par excellence* is customary law. He recognized, of course, that the details of a developed legal system do not spring from simple group intuition. With maturity, both life and law become more specialized and artificial, creating a dualism in more mature law. Part of such a system still rests directly on the popular consciousness and way of life ("the political element"); but this becomes elaborated by jurists, be they Roman jurisconsults or common-law judges, who in this respect represent the community ("the technical element").

Savigny's emphasis on the need of legal change to respect the continuity of the *Volksgeist* offers a pre-Darwinian concept of juristic evolution. The *Volksgeist* corresponds to modern notions of social rather than biological inheritance. Savigny's sense of the impotence of legislatures in the face of the restraints imposed by the *Volksgeist* foreshadows modern recognition of the social and psychological limits of effective legal action.

The English legal historian Sir Henry Maine's dual academic concern with both English law and Roman law challenged him to explain their independent yet often parallel growths and may well have redeemed him from Savigny-like overemphasis of national uniqueness. His concern led him to a comparative historical jurisprudence seeking hypothetical "laws" of development controlling all legal systems. He saw changes in substantive law and in the machinery and modes of legal enforcement and growth as moving in pace with certain recognizable stages in social growth, from primitive, kin-organized society to the mature, complex commercial and industrial societies of Europe.

Maine's experience in India after the publication of his *Ancient Law* in 1861 broadened his interests so that he embraced less well known and less developed systems, such as the Brehon, Hindu, Welsh, Germanic, Anglo-Saxon, and Hebrew. His breadth of interest matched the concurrent growth in anthropological study of primitive peoples.

Maine's work shows the strong combined influence of the analogy of biological to social evolution and of the Hegelian philosophy of history. The consequent, somewhat mechanistic tenor of his interpretations resulted in his being accused by many anthropologists and legal historians of making false assumptions concerning the pattern and sequence of social development.

Economic interpretations. Certain residues of the Marxist economic interpretation of history have won a central place in sociological jurisprudence (see below *Sociological jurisprudence*), as indeed in most branches of social science. One such persistent trend of thought is the close interrelatedness of legal, ethical, economic, and psychological inquiries; another is the pre-eminence among these of economic factors. According to Marxist doctrine, the political and judicial systems—the state and the law—represent the superstructure of society, their nature being determined by the economic base—the mode of production and exchange. The state and its repressive law are but instruments of class domination, becoming redundant under Communism, which has no need of coercion. During the transition to full Communism, they would "wither away." There were, of course, softenings of this bold doctrine in its original authors, with admissions that the ethical or legal superstructure should not be seen as a merely passive effect; and Lenin himself pressed to extremes both the passion of the original thesis and its qualifications. Lenin, indeed, saw state power as an essential weapon of the proletarian dictatorship until the movement to a full Communist society should be completed.

The Marxist conception of law

The first half-century of the Soviet Union, with its steady consolidation of state power and its attendant law, has imposed the severest strains on the withering-away prediction. The general tenor of explanation is that the "law" the disappearance of which is prophesied refers only to the kind of coercive order manifest in such instrumentalities as the courts, police, and jails of capitalist countries.

Within these sweeping theses of Marxist thinking, more modest subtheses have played a valuable part. The Socialist jurist Karl Renner, for example, in his *Rechts-institute des Privatrechts und ihre soziale Funktion* (1929), was concerned to show that the legal conception of ownership, formulated in early economies, had profound new effects when continued as an institution of the 19th-century economy. It then, through the law of property and contract, alienated into private hands great segments of what should be in the public domain.

Even more notable are the German sociologist Max Weber's studies of the correlations of socioeconomic and ethicojuristic change, freed of the straitjacket of economic determinism. In these, the impact of unique factors or combinations of factors in particular civilizations is taken into account, including the existence of accepted systems of values, immediate and ultimate, which may (and in Weber's view did) have a decisive effect on the emergence of the Western capitalist system.

Sociological jurisprudence. The historical jurisprudence of the earlier part of the 19th century became subject to the influence of the developing social sciences, which attempted to explain law in its social context. The result was the emergence of a sociological school of jurisprudence.

The early decades of sociological jurisprudence combined 19th-century faith in progress, social evolution, rationalism, humanitarianism, and political pluralism with a sanguine belief that the Newtonian model of natural science would also hold for the social sciences. It was affected by questions of whether the social sciences are truly sciences, what their mutual boundaries are, and whether they can be integrated or somehow transcended by some subject such as sociology or anthropology.

An outstanding figure of the early sociological school was a German, Rudolf von Jhering, who in the 1860s contributed to the intellectual stream a theory of justice predicated on a view of law as a social phenomenon. He saw law as an outcome of the struggle of men to fulfill their purposes and of the force that they marshal behind this. Another historical jurist, the German Otto von Gierke, stirred a related interest with his emphasis on the importance of the inner life and activities of groups and associations as sources of binding social norms. This opened up jurisprudence to some psychological issues. Gierke's work also contributed to the later American Neorealism through its influence on Oliver Wendell Holmes, Jr., and to the theory of the "living law" of the Austrian jurist Eugen Ehrlich, in the first decade of the 20th century. Ehrlich insisted on the profuse norm-creating activities of the countless associations in which men are involved.

Early 20th-century views

At the beginning of the 20th century a great variety of psychological hypotheses were brought to bear on law. A theory of dynamic psychic drives, for example, was propounded by an American sociologist, Lester F. Ward, who argued that such drives could be utilized in social planning. Freud's exploration of psychic activity on a subconscious level, as well as studies of the nonrational and the irrational in the social process by the Italian and German sociologists Vilfredo Pareto and Max Weber, were also profoundly influential.

Revival of natural-law theories. Iusnaturalism, in the sense of the assertion of an order of norms for human conduct transcending human will, to which the validity of positive law is subjected, has certainly experienced a 20th-century revival. The massive human delinquencies of the century, such as those of the Nazis, have been important in stimulating these modern natural-law yearnings. The revival, indeed, has rarely overthrown dominant Positivist positions, but it has certainly reopened some questions that Positivists have not adequately faced.

Contributions to this re-emergence have come from varied directions, rather than from a single intellectual movement. They have often avoided explicit reference to natural law and have even expressed hostility or ridicule toward it. The German Stammler and the French jurist François Gény were certainly among its pioneers. Gény's *Méthode d'interprétation* (1899) displayed the inescapably creative (or lawmaking) role of the judiciary even under a comprehensive code such as the Code Napoléon. It led him to the questions of what are "the sources of law" and where does the legislator's prescription fall short. Answers to such questions must be based on the facts of each particular situation to be adjusted—the legislator cannot impose his view on the court. This line of thinking foreshadowed a variety of doctrines about "the nature of things" or "the nature of facts," all of which shared the idea that the decisive nature of a situation has its base in the facts for which men seek governing law. The properties and circumstances of these facts themselves afford immediate guidelines for just regulation. The fact situation, if only its essence will be perceived, has the superior applicable norms immanent within it.

In his *Lehre von dem richtigen Recht* (1902), Stammler sought, as described above, the a priori social principles of just law concerning respect for and participation by all members. His call for "natural law with a changing content" based on these a priori principles quickly became a 20th-century slogan.

Even as this express reinvocation of natural law was proceeding, the French public lawyer Léon Duguit was expressly denouncing it. Duguit's concern was to place law and lawyers within what he saw to be the correct frame. This he found in Émile Durkheim's Positivist sociology. This led him, with some paradox for a contemner of natural law, to insist that law is but "*le produit spontané des faits*" ("the immediate result of the facts"). The observed "facts" of social solidarity arising from economic specialization of functions generated, Duguit argued, the society's norms. Breach of these norms causes social disorder and a spontaneous movement toward readjustment. Even a supreme legislator was bound (Duguit affirmed) by this objective "rule of law," so that his acts violating it are void, even apart from any other constitutional restraint. All this bears the clear iusnaturalist mark of the assumed immanence in observed facts of a transcending and overriding order. It pays cryptic homage to a natural law, fealty to which Duguit denied.

Law in Positivist sociology

The German legal philosopher Gustav Radbruch's turn toward natural law at the end of a life of great contributions to democratic legal relativism and Positivism was very different. Positivism, Radbruch argued, had encouraged German lawyers to stand by at Nazi barbarism, declaring "*Gesetz ist Gesetz*" ("Law is Law"). Nor was Radbruch's turn to natural law in any way cryptic. He came to declare quite openly that:

> where justice is not even striven for, where equality which is the core of justice is constantly denied in the enactment of positive law, there the law is not only "unjust law" but lacks the nature of law altogether.

The linkage with the revived natural law of the legal institutionalism of the French legal philosopher Maurice Hauriou and the writer and historian Georges Renard is different again. As with Duguit, the linkage is not proclaimed, but no overt hostility disguised their obvious sympathy for Thomist positions. Theirs is a Catholic version of institutionalism (which regards social institutions such as the family or the corporation as expressing the social reality underlying the law). The natural-law assumptions are apparent in the insistence on "the principles of organization," the "communion" of members in realizing "durable ideals," and the placing of men's powers of organization into the service of such ideals, as essential elements of any institution. For them, as for Duguit, the principles of justice were principles of social organization, immanent and self-evident.

It has been tempting for many to seek kinships between natural law and Existentialism, as was attempted by the German legal philosopher Werner Maihofer. Such efforts seem, however, destined to denature either Existentialism or natural law itself. Even in all their varieties, Existentialist positions approach no nearer to natural law than to assert that the traumas, anxieties, and demands of mere "existence" confront men with fateful value choices. Yet this is far short of asserting that any transcending principles of harmony may be discoverable.

Abstract symbols such as "social solidarity," "the principles of social organization," or "immanence in the facts of social life" are by virtue of their ambiguity susceptible to misappropriation by absolutist governments. The same may be said of Savigny's *Volksgeist* notion, as witness its affinity to the racialism of Nazi law. Thus, while the modern revival of natural law has been in part a revulsion from totalitarianism, it can also be exploited to rationalize totalitarianism.

There is another paradox also: The growth of the social sciences has invited restatement of natural-law traditions in terms of social ideals. Yet the very complexity of the social and economic orders and of their attendant sciences has placed forbidding barriers before the aspiration to base justice or other values on "objective" knowledge. Some have been tempted to hope that natural law may somehow overleap such barriers.

Pure theory of law. In part, at least, the influence of the distinguished legal philosopher Hans Kelsen's "pure theory of law" reflects early 20th-century skepticism about natural law and sociology, to both of which Kelsen opposed his claimed purity of method; *i.e.*, a method free from contamination by values of any sort.

He asserted, first, that legal theory was properly a science in the sense of an uncommitted, value-free, methodical concern with a determined object of knowledge. Second, he argued, legal theory must be isolated from psychological, sociological, and ethical matters. Third, purity of method permits the analyst to see that every legal system is in essence a hierarchy of norms in which every proposition is dependent for its validity on another proposition. The justification for describing any particular rule as law thus depends on whether there is some other proposition standing behind it, imparting to it the quality of law. This regression is continued until the *Grundnorm*, or "basic norm," is arrived at. The basic norm derives its validity from the fact that it has been accepted by some sufficient minimum number of people in the community.

Kelsen's hierarchy of norms

Kelsen's assertion that norms can spring only from other norms seems but another way of stating his rejection of the relevance of facts to values and, therefore, of iusnaturalism and sociology to his pure science of law. Yet finally it has seemed to many dubious whether the Kelsenite theory itself escaped the *liaison fatal* between facts and norms; for, if all legal norms must finally hang on the basic norm, then whatever it is that the basic norm hangs on must be nonlaw. And whether the basic norm hangs on "habitual obedience to determinate persons," as the English legal philosopher John Austin in effect proposed a century before, or on "efficaciousness," as Kelsen proposed, what it hangs on is fact rather than norm. Critics have complained that, at most points in the creation of norms in Kelsen's system, what is decisive is the intervention of acts of will of persons endowed by higher norms with norm-making authority. The determination whether such acts of will have occurred is a factual inquiry, to the decisiveness of which Kelsen's pure theory gave little weight.

Modern schools of realism. The American jurist Oliver Wendell Holmes's description of law in 1897 as "what the courts will do in fact" and of the "real ground" of decisions as resting often in some "inarticulate major premise" rather than in expressed reasons gave 20th-century legal realism its central theme.

Certain features are common to the "realist" jurists. They include (besides the above-mentioned concern with "the law in action") stress on the social purposiveness of law, on the endless flux in both society and law, on the need to divorce the "is" and the "ought" for purposes of study and to question all orthodox assumptions made by lawyers, and in particular on the need to substitute more realistic working categories for current lawyers' generalities. Among the orthodoxies thus challenged, these writers tended to include the works of early sociological jurisprudence. Yet it is clear, from the present perspective, that the concerns common to the realists and the more orthodox sociological jurists were far more important than the ephemeral if bitter conflicts that at first flared up between them. The American realists in their important surviving contributions have for the most part reinforced, clarified, and elaborated a number of main insights, notably about rule uncertainty and fact uncertainty, which they shared with sociological jurisprudence.

Scandinavian realism

Scandinavian realists, while temperamentally akin to their American colleagues, were rather different in intellectual concerns. Methodologically, they invoked a somewhat gross empiricism, leading them to deny that the law could be the subject of scientific inquiry at all, since its concepts and principles are not founded on spatial and temporal data of experience. Taking lawyers' talk of the will of the sovereign very literally, they were concerned to show that there is no such will of common content and that even legislators who enact a code are merely rubber stamping what others drafted.

The Swedish jurist Axel Hagerström insisted that the idea of rules of law as commands is an idea not corresponding with facts. His disciple Karl Olivecrona added that this false idea results from the syntactical imperative form used in modern legislation. Such rules, he urged, were commands only in a depersonalized sense. He preferred to describe them as "independent imperatives." Such "imperative statements about imaginary actions, rights, duties" may not be directed to any particular persons. Yet, even if some legal rules are directed "so to say, into the air," others are certainly directed to particular persons. If any form of imperative notion is to be preserved, it should be one that accommodates both situations.

Some of the problems that these writings address are rather tied to the special experience of their authors' own legal cultures. Others reach out independently toward truths already reached earlier in Anglo-American jurisprudential scholarship, especially as to the merely noetic and conceptual (rather than physical or psychological) nature of rights, duties, and liabilities.

THE STATUS OF CONTEMPORARY PHILOSOPHY OF LAW

By the middle of the 20th century, serious scholars no longer argued for or against the exclusive imperium of either the analytical–logical, the justice–ethical, or the sociological approach. Whether jurisprudence is a single field in some scientific sense or whether its unity lies in the need to serve the intellectual needs of those concerned with making, applying, improving, or generally understanding law, all the above areas are included within it.

The "revolt against formalism"

A characteristic feature of contemporary jurisprudence is what has come to be known as "the revolt against formalism"; that is, against preoccupation with the technical and logical aspects of law. It can be traced back to Savigny's early 19th-century reaction against natural law, to Jhering's attacks on the German Pandectists (commentators on Roman law), and to Maine and the work of the anthropologists and early sociological jurists. Its early pressure was toward broader and deeper history, toward recognition of the organic nature of the processes of cultural growth, and toward problems of social action and the value choices therein entailed.

In the United States the legal philosopher Morton White identified five later contributing strains of thought, including the pragmatism of John Dewey; the economist Thorstein Veblen's institutionalism, rejecting both the abstractions of classical political economy and the fatalism of the Marxist interpretation of history; the revolt within jurisprudence of the American legal realists already described; and the approach to history as no mere chronicle of kings and battles but rather as a product of underlying economic forces and a guide to present and future civilizations.

The sometimes-overhasty iconoclasms of this revolt have proved less important than its positive affirmations. It has affirmed, for example, that the evaluating activities of justice must somehow move alongside the describing activities of sociological jurisprudence, that the choices of ethics, social policy, and justice still remain to be made when all the empirics of social science are done. The central question includes not merely what are the facts but also what should be done about the facts. These affirmations reject any regression to simple amoralism, stirring new temptations to return to natural law or other intuitive absolutes.

There are important advantages in the drive, characteristic of much contemporary social science, for overall cognition of the social and legal orders and the identification of key points for social action within them. But there are also dangers, for, especially with subject matters such as the law, systematic theory and overall cognition can rarely be of aid save in the rather long run, for which present decision makers cannot usually wait.

Growth of the sociological school. The most eminent pioneers and champions of modern sociological jurisprudence were Roscoe Pound in the United States and Hermann Kantorowicz in Europe. For both, the task of sociological jurisprudence, though orientated mainly to practical administrative or legislative problems, included that of framing hypotheses (as to the limits of effective legal action, for example) on which to base general laws of the operation of law in society.

Methodology of sociological jurisprudence

As with the social sciences, the principal methods available to sociological jurisprudence are those of survey, statistical analysis, comparative observation, and experimentation. The controls and corrections available usually fall far short of those of the natural-science models. Much work in sociological jurisprudence has merely brought to bear upon the law relevant findings from other social sciences. But it may also generate its own findings, as it has done in relation to traffic laws, control of moneylending, credit unions, bankruptcy laws, the effect of antitrust practices or of poverty on legal rights, the theory of appellate judicial decision making, and a host of other matters. Examinations of the prehistory and aftercareers of convicted criminals and of persons on probation or parole, probings of family and environmental influences bearing on potential deviance, and attempts to identify decisive factors predictive of future deviance are among the staples of sociological jurisprudence.

Sociological jurisprudence is confronted by the questions whether (and, if so, how and how far) it is possible through empirical methods to approach central issues of social action that involve value judgments. The fact that lawyers are necessarily involved with ideas of obligation, values, and norms sharpens this confrontation. A second group of problems arises from the high level of individuality of men, groups, and societies, from the unending variety of their emotions, roles, and expectations, and from the feedback effects on human behaviour that the empirical observation and testing of that behaviour brings about.

These problems give central importance to efforts to develop frames of social knowledge that give due place to both facts and values. Such inquiries show the great complexities of values held and their intricate and dynamic relation to the physical and cultural environments.

The study of law in society thus shares with anthropology and other social sciences a central interest in roles and functions as basic meaningful categories and in certain mechanisms and channels whereby conduct is thought to become socially meaningful. These notions are thought to permit the analysis of complex social situations into more refined terms, such as constituent goals, tasks, expectations, and allocated rights, powers, and duties.

As to the mechanisms or channels through which conduct becomes socially meaningful, earlier thought tended to explain social norms as built up from individual instances through group usages and mores that then crystallize in in-

stitutions such as law. Insofar as this suggests a cumulative movement or process, current thought would regard it as oversimplified. The growth of socio-ethical convictions is rather to be seen in terms of symbolic interaction between individuals. A particular society may be seen, in this light, as a collection of individuals with a culture that has been learned by symbolic communication from other individuals back through time, enabling members to gauge their behaviour to each other and to the society as a whole.

The future of sociological jurisprudence. In his famous program of 1911–12, Pound formulated a series of rather practical objectives for the movement, including making studies of the law in action, of the means of more effective legislation and law enforcement (by creation of ministries of justice, for example), of legal and judicial reasoning, of legal history in its social context, and of the role of the legal profession. An early quip against the sociological school was that it was like a great orchestra constantly tuning its instruments but never actually playing. Yet many practical tasks have been performed, and the school continued to show a gathering momentum and a widening range of concerns.

The maladjustments and inadequacies of the law gave to early sociological jurisprudence an intensely activist drive, directed to ad hoc remedies, and a great deal of the relevant work is still of this nature. Especially since 1945, however, juristic work on the relations of law and society has come into more fruitful contact with other social sciences, leading in turn to greater stress on cognition of the social and economic orders in their complex unity. Whatever the difficulties of designs for an overall analysis of the social system, some adjustment toward them is inevitable for sociological jurisprudence. This is in part, no doubt, a result of the waning of interest in many of the kinds of ad hoc problems with which it was initially concerned. But the interest in sociological theory also results from growing awareness that some problems require to be approached on a wider basis. This has created new stirrings of the turn-of-the-century ambition that the study of law in society become a specific branch of social science, concerned with framing and testing general laws governing law as a social phenomenon. (Ju.S.)

Philosophy of education

Different conceptions of the philosophy of education

Philosophy of education is a field of inquiry, speculation, and application in which philosophical methods are applied to the study of a problem, topic, or issue in education. Characteristic of these methods is the attempt to think as accurately, clearly, coherently, and systematically as possible. Analytical philosophers would say that philosophy of education should end with the attempt to clarify and justify educational statements and arguments. In practice, however, the field includes much more than that. It considers as relevant material much that has been written on education by influential philosophers of the past. There is, therefore, much overlap with the field of history of education, and ultimately no clear demarcation can be drawn between the two fields, which nourish and illuminate each other in their interconnectedness (see EDUCATION, HISTORY OF). Many philosophers of education also go beyond analysis in being concerned with establishing a commitment to value judgments and substantive positions. They take pains to attempt to clarify and justify those judgments and positions on the grounds that clear and substantiated judgments have greater probability of being sound.

The philosophy of education has a special concern with the *applications* of knowledge and theories. Thus, many philosophers of education are especially interested in the relationship between theory and practice. Moreover, they are often concerned with the ways in which philosophy relates to other fields of study in the attempt to shed light on educational problems and issues. This gives them a wide-angled approach to education, which some philosophers have called "educational theory" to distinguish it from a more narrowly analytical form of philosophy of education.

With regard to the term "education," there is a similar spectrum of views ranging from the narrow view of education as that which goes on in schools and universities to a definition of education as all those experiences that affect the growth and development of a person throughout life. The former view has tended to give way, with increased recognition of the crucially important part that informal experiences and relationships play in determining what and how an individual learns.

COMMON ASSUMPTIONS ABOUT HUMAN LEARNING

Symbolization. First, it is to be noted that human learning is largely dominated by symbolic processes. Much of the learning that man acquires during his lifetime is gained through his growing ability to understand and manipulate symbols—verbal, mathematical, artistic, musical, and so on. This symbolic emphasis gives human learning much of its power. It brings power because symbols enable man to deal abstractly and at a high level of generality with words, data, and ideas. Thus, he can learn to think rapidly, to link varied items in fruitful ways, and to create shortcuts to new knowledge.

Dangers of abstraction

The symbolic emphasis also brings dangers. It can trap man into circling around at a high level of generality without ever feeling the need to tie his abstractions to concrete applications. Schooling thus can become an apparently self-justifying cycle, which in fact fails to pay off in terms of improving men's lives. Moreover, since learners assimilate and master these symbolic processes at different rates, it is tempting for educators to identify all education with symbol manipulation, to classify those who take longer to master the process as inferior, and to reserve the privileges of higher schooling for those who master this particular process rapidly. This is to mistake a part for the whole and leads to enormous losses of human power and development.

Human motivation. Motivations to learning are more variable in man than in subhuman species. In other animals learning patterns develop largely through instinctual motivations. This means that lower animals need less training than humans before they are independent and self-reliant. Human beings are the most dependent of creatures when young, and they take longest to educate for independent activity. It is precisely this long period of dependence and education that constitutes man's superiority over lower animals, for it is in this period that the growing human being absorbs much of his culture and develops the skills and knowledge that enable him to build on the work of his predecessors. His lower degree of reliance on innate mechanisms also means that man is more flexible in his responses and is capable of adjusting those responses in the light of previous experiences. Thus, he is not condemned to repeat previous patterns of living and thinking but can create change, both in himself and in his culture.

Educators have often acted as if man were a lower animal, for schooling has often resembled training, with emphasis upon imitation, obedience, repetition, drill, and control, more than it has resembled education, with emphasis upon initiative, creativity, choice, decision making, and freedom. As a result, schooling has been more successful in improving man's skill in mechanical and repetitive activities at the lower end of his capacity scale than in developing higher qualities of intellect, feeling, and will that constitute his peculiar humanness.

Human capacity to learn. So far there have been discovered no limits to man's capacity to learn. From earliest times, however, men in positions of power or influence have suggested that the learning capacity of certain individuals or groups is severely limited and that they should not be expected to profit greatly, if at all, from education. These "ineducable" individuals have usually been members of minority or disadvantaged groups. But, repeatedly, when their cultural disadvantages have been removed, these groups have shown that their previous failure to learn has been due not to incapacity but to lack of fully realized opportunity.

Categorizing learners

These findings have led educators to be much more modest and less hasty in their labelling and classifying procedures. It has been realized that labels affixed to children tend to become self-fulfilling prophecies, that those

who are expected to learn usually do so, and those who are expected to fail to learn also usually do so. Hence, when educators resort to classifying children at all, they increasingly tend to use their labels as temporary rather than permanent, as saying something only about a quality of the child rather than about his person, and as something to be abandoned as soon as the child's performance proves the label wrong.

Similarly, no one has been able to confirm any certain limits to the speed with which man can learn. Schools and universities have usually been organized as if to suggest that all students learn at about the same rather plodding and regular speed. But, whenever the actual rates at which different people learn have been tested, nothing has been found to justify such an organization. Not only do individuals learn at vastly different speeds and in different ways, but man seems capable of astonishing feats of rapid learning when the attendant circumstances are favourable. It seems that, in customary educational settings, one habitually uses only a tiny fraction of one's learning capacities.

Complexity of human learning. Human learning is complex rather than simple. Learners are apt to learn more than one thing at a time. Sometimes this process is conscious, as when one simultaneously or rapidly assimilates many specific items of a whole. More often, the process is entirely or partly unconscious, as when the student learns some "content" consciously but at the same time absorbs unwittingly a great deal more from interrelationships, tones of voice, and so on.

Correlative learning

Educators are therefore becoming increasingly concerned with these concomitant learnings. They are aware that the long-term significance of the arithmetical skill that the student consciously learns may be nugatory compared with the importance of what he learns about himself as a learner, about his capacities and limits, about his relationship with his teacher, about power and authority, about his relationships with his fellow students, about equality, collaboration, competition, and friendship. As educators become more knowledgeable about the importance of learning climates, they are impelled to abandon simplified techniques of teaching in favour of a more complex approach that views learning in the context of a matrix of relationships and forces that act upon the student, the teacher, the school, and the community.

Intellectual–emotional–physical–volitional integrity of learning. Human learning concerns the whole person. The intellect is not the only agent of learning. This activity is shared by the body, the emotions, and the will. Moreover, the process cannot be limited to any one of these domains without affecting the others. Educators are most conscious of intellectual learnings, which tend to play the largest part in their plans and intentions. But there is increasing evidence that makes clear the folly of attempting to confine education to the training of intellects. If the teacher does so, he is destined to fare badly, for the child who is emotionally frozen or whose stomach is empty or who is determined to thwart the teacher will not perform intellectually as the teacher intends.

Educators are also becoming more aware of the other side of this coin—that is, that the learner's powers are vastly enhanced when not only is his intellect stimulated but also his feelings are respected, his body is nurtured, and his will to learn is strengthened. Effective education, therefore, is found when the learner is regarded as a person to be respected, nurtured, strengthened, and stimulated, rather than as an intellect to be trained.

DIFFERING CONCEPTIONS OF EDUCATION

Model theories of the educated person. In different historical eras and in different cultures a variety of conceptions has been put forward concerning the nature of mind, of truth, of knowledge, of moral goodness, of aesthetic beauty, and of the purposes of life. These views carry implications for the purpose and nature of the educational process, the ends toward which one should educate, the means by which one can best achieve those ends, the degree of consistency between the ends and the means, and the model, if any, of the ideally educated person that one has in mind when engaging in the educational process. These relate to some of the perennial questions that concern the philosopher of education. The following represent some of the most influential models that have been presented of the educated person.

Platonic view. The first great attempt to create a philosophically coherent model of the educated person was that of Plato. Reacting against the turbulence and chaos of his own times, Plato envisaged in *The Republic* and the *Laws* a permanent, stable, hierarchical society in which those most adept at education would rule, those moderately adept would become warriors and carry out the orders of the rulers or guardians, and those least adept at education would fill the lowliest worker functions in society. The education that would thus order and segregate the different groups in society would be based upon the ability to understand the perennial, ideal Forms that, Plato maintained, underlie the transient, concrete manifestations of those Forms in everyday life. Since, supposedly, not all people could surmount the difficulties of such an intellectually demanding program of study, it would be justifiable to select, classify, segregate, and reject students at the various stages of the educational process.

Intellectually segregated education

The Platonic scheme has been enormously influential, especially in the Western world, during the last 25 centuries. Most educational programs in the West have consistently reflected this hierarchical pattern. Schooling has usually been a process of selection and rejection, with great effort and ingenuity expended on testing, measuring, classifying, and segregating in accordance with the best available knowledge. Similarly, Plato's concept of mind has exerted lasting influence on Western education. Patterns of schooling have reflected the greater prestige accorded to the study of ideas and abstractions and the lower prestige given to practical studies and manual work. There has persisted a dichotomy between the so-called liberal arts, which have been considered suitable educational fare for potential leaders of society, and so-called vocational studies, which have been considered more suitable for potential followers. This pattern has lasted to the present day, not only in Europe and North America, but also in those parts of Asia, Africa, and Latin America greatly influenced by European culture. Thus, one finds in many former colonial societies that students often strongly prefer to follow "liberal" studies, even though the choice may condemn them to genteel unemployment in a society whose economy cries out for people with technological, practical, and manual skills.

Thomist view. In the 13th century, St. Thomas Aquinas made a monumental attempt to reconcile the two great streams of the Western tradition. In his teaching at the University of Paris and in his writings—particularly the *Summa theologiae* and the *Summa contra gentiles*—Aquinas tried to synthesize reason and faith, philosophy and theology, university and monastery, activity and contemplation. In his writings, however, faith and theology ultimately took precedence over reason and philosophy because the former were presumed to give access to truths that were not available through rational inquiry. Hence, Aquinas started with assumptions based on divine revelation and went on to a philosophical explication of man and nature. The model of the educated man that emerged from this process was the Scholastic, a man whose rational intelligence had been vigorously disciplined for the pursuit of moral excellence and whose highest happiness was found in contemplation of the Christian God.

This Scholastic model has greatly affected the development of Western education, especially in fostering the notion of intellectual discipline. Aquinas' theological-philosophical doctrine has been a powerful intellectual force throughout the West and has constituted the official basis of Catholic theology since 1879. Although Aquinas made an important place in his hierarchy of values for the practical uses of reason, later Thomists have often been more exclusively intellectual in their educational emphasis. For Aquinas, the primary agent of education was the learner, and his model was, thus, a person capable of self-education. Intellectually autonomous, he should be able to conduct his own process of research and discovery. The Roman Catholic Church, however, has usually put

the learner firmly under the authoritative superordination of the teacher.

Lockean view. John Locke, the 17th-century English philosopher, has been credited with formulating the classical liberal defense of individual freedom against the authorities of state and church. Opposed both to what he deemed the stagnation of unreflective tradition and the perils of enthusiastic radicalism, Locke saw science, reason, and experience as the best safeguards against these dangers. Responding to the rise of the new bourgeoisie and the new science, he became the principal spokesman for the increasingly powerful middle class, who were predominantly skeptical and practical in their intellectual temper. The model of the gentleman had traditionally been the English ideal of the educated person. Locke's achievement was to take this ideal and modify it to a form acceptable to the new bourgeoisie. Originally an aristocratic model, the gentleman ideal became infused, under Locke's influence, with democratic, Puritan, and practical characteristics.

Education of the practical gentleman

Locke's notion of the mind at birth as a *tabula rasa,* a "blank tablet" devoid of innate ideas, gave enormous importance to the role of experience and sense perception in the educational process. In *Some Thoughts Concerning Education* (1693) and the *Conduct of the Understanding* (1706), Locke outlined the heavily experiential education that would be appropriate for a gentleman. His four cardinal aims of education, in order of importance, were virtue, wisdom, breeding, and learning. Some critics have insisted that the order has remained important to the present day in English education, that learning must always be lightly worn and never ostentatious. Not only was Locke's thinking influential in shaping the subsequent development of English educational thought, but his Puritan individualism also had a considerable effect on the growth of American educational ideals. Locke's failure to recognize the possibilities of the uses of institutional power and legislation for interventions that would enhance rather than restrict freedom was an omission that has often bedevilled American educational thinking.

Naturalistic view. Most movements in philosophy and education have seemed to engender their own opposition. Hence, the rationalism and scientific objectivity of the Enlightenment found a reaction in the subjectivity and emotional spontaneity of romantic naturalism, the principal spokesman of which was Jean-Jacques Rousseau. In particular, Rousseau reacted against the excessive formalism and rationalism of 18th-century France. Out of this reaction there came a model of the educated person as the natural man, a figure presented in contrast to what Rousseau saw as the pathetic products of contemporary civilization. Against civilized values like rationalism, conscious reflection, control, complexity, and objectivity, Rousseau offered his own values of romanticism, intuitive spontaneity, freedom, simplicity, and subjectivity.

Subjectivity and spontaneity in learning

In *Émile* (1762), one of the most influential books on education ever written, Rousseau argued that one should protect the child from the corruptions of civilization and carefully nurture his natural, spontaneous impulses, which were always healthy. It was important, he maintained, to avoid premature intellectualization of emotion so that the child's intellectual powers could develop without distortion. Feeling should precede thinking, and the child should be controlled only by things, not by adults' wills. In these ideas lay some of the germs of progressive education, which spread throughout the world during the 19th and 20th centuries. Constantly in tension with the classical demands of reason, discipline, authority, and scholarship, the romantic naturalism of the progressive-education movement has continued to remind educators that their ultimate concern should be the growth of the unique, ultimately unfathomable child.

Marxist view. Probably the most influential writer of recent times was Karl Marx, whose writings, mostly in collaboration with Friedrich Engels, altered the course of history and continue to affect the lives of millions of people in all parts of the world. A central concern of Marx and Engels was to cure the alienation and dehumanization of man caused by what they saw as the exploitative forces of capitalism. In Marx's writings, the material dimension of history appeared as primary. Economic production was deemed the basis of life, and the prevailing ideas (religious, educational, and political) of a society were seen as being determined by its economic structure. The dominant ideas of an epoch or society were considered to be the ideas of its ruling class—that is, the class that controlled the means of material production. In order to find the hidden interest behind an idea, Marx argued, one had to examine its social function rather than its intellectual content. Marx saw the need for a proletarian revolution in order to bring about Communism. Under Communism, he argued, the opposition between the individual and the group would disappear; each man's interests would be seen to be identical with the interests of all, and alienation would be banished.

Marxist emphasis on social education

In Marx's view, what was needed for man's growth toward maturity was genuine community; that is, the voluntary drawing together of autonomous and socially responsible persons. The model of the educated person that Marx put forward was not the irresponsible individualist nor the coerced collectivist but the accountable communal man, who attained his freedom not by fleeing from social relationship but *through* social relationships. Individual freedom required social authority.

Marx's and Engels' writings, particularly *Das Kapital* (1867–94), *The German Ideology* (London ed., 1938) and *The Communist Manifesto* (1848), have reshaped the world. Millions have been inspired by their vision of unalienated work and education. Unfortunately, like many influential ideas, they have been abused and distorted. Several states have tried to bring about a Communist utopia through collectivist coercion and social manipulation in a distortion of Marx's arguments. There is a need for state intervention to remove gross inequalities and to expand opportunities, in Marx's view, but ultimately human regeneration is a task for each individual.

Pragmatist view. Out of the America that was created by immigration, urbanization, and industrialization in the second half of the 19th century came the philosophy of Pragmatism. Associated with such thinkers as C.S. Peirce, William James, and John Dewey, Pragmatism, as the dominant American philosophy, exerted a strong influence on the shape of education in the United States, and affected educational ideas and practices in Europe and Japan. In the hands of Dewey, Pragmatism evolved into a philosophy that saw man as formed through interaction with his natural and social environment. The educated person was always viewed by Dewey in a social context. Neither the individual nor society had any meaning without the other. Dewey created a model of the educated person as the reflective man, one who was critical of the authority of custom and tradition as the determinant of belief and action and who preferred the method of science, of "organized intelligence" as the best way to solve his problems.

Progressive emphasis on blending individual and social needs

In *My Pedagogic Creed* (1897), *How We Think* (1910), *Democracy and Education* (1916), *Human Nature and Conduct* (1922), and a stream of other philosophical and pedagogic writings, Dewey formulated a viewpoint that constituted a rigorous intellectual core of the progressive-education movement, although he criticized many of the manifestations of progressive pedagogy in practice. For example, he viewed the child's interests as vitally important and to be neither repressed nor humoured. To repress them was to commit the fault of much traditional education by ignoring the child's unique bent. But to humour them was to commit the fault of some progressive education by failing to discover the underlying power below the passing whim. Subject matter, Dewey argued, should consist of activities that enabled the child to reflect upon his social experiences. When subject matter preceded or was unrelated to the child's experiences, it was largely meaningless. It gained meaning through being made the medium for continued reflection upon, and reconstruction of, experience.

Behaviourist view. In the 20th century, the use of science has been extended to the study of virtually all aspects of human affairs. The possibilities of scientific control of men and events have brought profound changes in

philosophy and education. By studying only the behavioral aspects of man, science has been able to predict and control in ways that have powerful and sometimes frightening implications. The behaviourist view has been most notably represented by the American psychologist B.F. Skinner. In his writings, including the utopian novel, *Walden Two* (1948), *Science and Human Behavior* (1953), and *Beyond Freedom and Dignity* (1971), he has firmly rejected the conventional model of man as a free agent who acts in accordance with the decisions of an inner self that is neither fully explicable nor fully controllable by scientific means. Instead, Skinner envisages the use of scientific knowledge about the control of human behaviour to create a planned man, one who will be conditioned to behave in the way best calculated to achieve society's goals. Behavioral engineering will have removed all of his antisocial tendencies, and he will want only what is good for himself and his society.

Programmed learning and behavioral development

Behaviourism in general and Skinner in particular have exerted considerable influence on Western, and especially American, education through developments that range from the invention of programmed instruction (or the so-called "teaching machine") to the widespread emphasis on behavioral objectives in educational programs. The philosophy of Behaviourism compels one to examine carefully the issue of control in education. Skinner wants to use scientific control to bring about a society in which it will be easy to be good and to bring about an educational process through which it will be easy to be excellent. There is no alternative to control, in his view. It is simply a matter of *who* is to control. One does not grant the child "freedom" merely by leaving him alone. To refuse to use scientific control to shape human behaviour is, for the Behaviourist, a failure in responsibility.

Existentialist view. The application of science to ever more aspects of the study of man has been particularly marked in the West; but, in reaction to this trend, there has developed some criticism of the scientific way of viewing man as an object to be categorized, studied objectively, or subsumed under a generalization. Prominent in this reaction have been Existentialist philosophers, among whom the Jewish philosopher, Martin Buber, played an outstanding role in deepening contemporary understanding of man and education. In such books as *I and Thou* (1923) and *Between Man and Man* (1947), Buber's model of the educated person appeared as one whose life was shaped by existential decision making. Such a person did not determine choices in advance of existential situations. He used principles and traditions only as checks or reminders, not as infallible guides. His values were created in the concrete here and now and were manifested as he related to other men. Thus, each man was seen as a unique person rather than as a member of a category.

Relationship between learning and life

A vital concept for Buber was responsibility, viewed in terms of one's response to another. Thus, the dialogue became a central focus in his educational philosophy. The educated person was one who could listen as well as talk. And since genuine dialogue depended upon authenticity, upon *being* rather than *seeming*, one needed the courage to be oneself in relationships. Buber also urged the recognition of a continuity between learning and life, rather than the encouragement of knowledge for its own sake. He insisted that learning be related to consequent action. Thus, for Buber, the educated person was not one who merely had had his cognitive faculties trained but one whose inmost spirit had been infused by what he had learned.

Areas of disagreement among educational theorists. *Conceptions of what constitutes knowledge.* Although there is wide agreement that the assimilation of knowledge by the learner is a principal goal of education, this agreement is more apparent than real, because there is much disagreement about what constitutes knowledge. Clearly, if one person identifies knowledge with information, his pedagogic approach is likely to differ from that of one who identifies knowledge with a process of thinking. Some philosophers have attempted to clarify a distinction between "knowing that" (something is true or false) and "knowing how" (to do or learn to discover something). The former emphasis is common in educational programs that focus on the product—that is, on the information that the learner has amassed and can reproduce at the end of his program. The latter emphasis is common in programs that focus on the process—that is, on the skills and attitudes that the learner has adopted that enable him to learn more.

Conceptions of how knowledge should be communicated. Throughout history, the commonest way for knowledge to be communicated has been through what can be loosely called apprenticeship. The one who does not know (the apprentice) watches the one who does (the master); he imitates the master, probably fails, watches again, tries again, and so on until he knows what the master knows. This is primarily the way young people learn role identity, for example—what it means to be a husband or wife, a father or mother, a male or female. It is not an effective method for stimulating the growth of new knowledge. A second method, one that has been time-honoured in institutions of formal schooling, is telling. The one who knows (the teacher) tells the one who does not (the student). The student listens, tries to remember, and is usually required to reproduce his memorizations at some stage so that the teacher can judge whether or not the knowledge has been assimilated. This method is a much less effective way of communicating knowledge than would be guessed from its prevalence as an educational practice. Other methods try to remove the difficulties of human relationships from the act of communication by removing one party from the scene. For example, the apprentice can watch the master on film or videotape. Or the teacher can write his lectures in a book, which the student can then read. Another method is that of the dialogue. Unlike the previous methods, which imply superiority–inferiority relationships, this method implies equality between the members of the dialogue. All may at different moments be teachers or learners. The method assumes that all have some knowledge to give and that all need to learn.

Apprenticeship, telling, and dialogue

Conceptions of the purpose of education. One of the commonest areas of disagreement among educational theorists concerns what education should be for. Two broad divisions of opinion can be identified. They are, first, that education should serve the needs of the individual and second, that it should serve the needs of society. In industrialized nations it is common for both of these goals to be held but for different classes of the population. For the elite, the needs of the individual tend to prevail, and thus upper-class schooling often tolerates diversity and encourages idiosyncrasy. For the masses, the needs of society tend to dominate, and schooling usually serves to prepare children to become obedient, well-drilled, uncomplaining workers in industry and agriculture. In addition to this dichotomy, there are many national and cultural variations. In France and Germany, for example, there is a strong tendency to see the primary purpose of education as some form of intellectual development, with the French *lycée* and the German *Gymnasium* serving as characteristic institutions. In England, education has usually been seen as primarily serving the purpose of character building. The English public school is sometimes taken as an epitome of this emphasis. In Roman Catholic and Communist countries, moral or religious training is usually the primary purpose of education. Hence, every subject is imbued with religious (or, if it is Marxism–Leninism, quasi-religious) content, and there is a constant attempt to draw moral lessons from educational material. In the United States, where large-scale immigration once brought the fear of social disintegration, preparation for citizenship and development of national consciousness have tended to be emphasized. Even today, nationalistic rituals such as displaying flags and reciting the "Pledge of Allegiance" in classrooms distinguish American schools from those in most other countries.

Needs of the individual and of society

Conceptions of uniform versus differential grouping. One of the controversies that most consistently divides educational theorists is whether students should be clustered together with similar students for instruction or whether

Means of segregating students

instructional groups should be made deliberately diverse. The criteria that are employed for differentiating students are manifold. They include chronological age, mental age, IQ, skin colour, sex, social class, geographical location, parental income, performance ability on various tests, and so on. Some of these criteria, such as skin colour, are used "officially" in some countries, such as South Africa, and "unofficially" in others, such as the United States. Other criteria, such as parental income, are used overtly in some schools (expensive, private schools) and covertly in others (public schools located in expensive suburbs). Some criteria, such as mental age, IQ, and performance ability, often have the effect of separating rich from poor or elites from cultural minorities because of the cultural content of the tests used. The customary arguments used in favour of homogeneous grouping tend to emphasize speed and efficiency in instruction and learning. Arguments in favour of heterogeneous grouping tend to emphasize social outcomes, such as an enhanced understanding of others and tolerance of diversity.

Conceptions of control and discipline. When education is viewed as a transaction between superior (the teacher) and inferior (the student), questions of control invariably arise; *e.g.*, can students be relied upon to do what is best if they are not controlled by teachers? Answers to this question refer back to assumptions about the nature of man as learner. Those theorists who tend to see man as basically untrustworthy will tend to find justification for a regimen of strict control over students, in order to prevent their natural impulses from harming themselves and others. Theorists who view man as basically trustworthy will tend to justify more lenient, loose, and permissive systems of control. Most theorists would agree that an important goal of discipline would be for the student to develop self-discipline. But disagreement arises over whether one best learns self-discipline through practice in obeying others or through practice in commanding oneself.

Conceptions of the need for competition and collaboration. In most cultures of the world, children from an early age seek to compete against and to surpass their fellows. Whether this competitive spirit is innate or whether it is acquired through the culture is at present unverifiable. The question that divides education theorists, however, is whether these existing competitive strivings should be encouraged or reduced. Some argue that, since competition is endemic in the culture and since schooling should assimilate the child to the culture, the school should encourage individual competition and help children to develop the strengths necessary to compete successfully. On the other hand, others argue that competitiveness in the culture is pernicious and that schooling should not serve to adjust children to harmful aspects of the culture but should foster cultural progress. Usually, the school is in conflict over these two attitudes and gives mixed messages to students. As a result, the students often suffer the double strain of trying to compete successfully but to do so without appearing to compete.

Conflict of aims in school

Conceptions of the boundaries of formal education. Considerable disagreement exists over what should legitimately be studied in formal institutions of learning. Usually some topics or subjects are considered taboo. These taboos may come about through interpretations of the law, as in the case of the exclusion of religious teaching from public schools in the United States. Or they may come about as a result of pressure from interest groups, as in the attempts to exclude sex education, teaching about Communism, and sensitivity training from American schools. The major considerations are that young children are not well equipped to resist heavy bias in teaching and that they are usually compelled by law to attend school and thus constitute a captive audience. It is sometimes argued, therefore, that they should be protected against religious and political propaganda and against material or experiences that require greater maturity to be handled creatively. On the other side, the danger in such arguments is that they can be used to keep all controversy out of schools and to render them places characterized by dull uniformity of thought.

(Pa.N.)

BIBLIOGRAPHY

Philosophy of science: ERNEST NAGEL, *The Structure of Science* (1961); and M.W. WARTOFSKY, *Conceptual Foundations of Scientific Thought* (1968), together cover the main approaches to the philosophy of science as a general field of study; whereas ARTHUR DANTO and SIDNEY MORGENBESSER (eds.), *Philosophy of Science* (1962), provides a useful anthology of classic papers on the subject; and SAMUEL B. RAPPORT and HELEN WRIGHT (eds.), *Science: Method and Meaning* (1964), is a stimulating anthology of nontechnical papers. W.V. QUINE, *Theories and Things* (1981), is a collection of papers by an eminent philosopher, on topics of ontology, logic, semantics, epistemology, philosophy of mathematics, identity, existence, etc.; W.H. NEWTON-SMITH, *The Rationality of Science* (1981), is an analysis of interpretations of science; V.V. NALIMOV, *Faces of Science* (1981), is a collection of essays on the nature of science, which originally were published in Soviet philosophical journals; LOREN R. GRAHAM, *Between Science and Values* (1981), is an examination of the relationship of science to values; JEREMY CAMPBELL, *Grammatical Man: Information, Entropy, Language, and Life* (1982), is a study of philosophical aspects of information theory.

The origins and development of natural science and the attitudes of philosophers toward scientific explanation in antiquity and the Middle Ages are well treated in SAMUEL SAMBURSKY, *The Physical World of the Greeks*, 2nd ed. (1962; orig. pub. in Hebrew, 1954); and *The Physical World of Late Antiquity* (1962); JOHN H. RANDALL, JR., *Aristotle* (1960); and in ALISTAIR C. CROMBIE, *Medieval and Early Modern Science*, 2nd ed., 2 vol. (1963). For the 17th and 18th centuries, see EDWIN A. BURTT, *The Metaphysical Foundations of Modern Physical Science*, 2nd ed. (1932); ALEXANDRE KOYRE, *Newtonian Studies* (1965); HERBERT BUTTERFIELD, *The Origins of Modern Science: 1300–1800*, rev. ed. (1965); I. BERNARD COHEN, *Franklin and Newton* (1956); and GOTTFRIED MARTIN, *Immanuel Kant: Ontologie und Wissenschaftstheorie* (1951; Eng. trans., *Kant's Metaphysics and Theory of Science*, 1955). For the 19th century, the most significant contributions have come from scientists writing as philosophers: CLAUDE BERNARD, *Introduction à l'étude de la médecine expérimentale* (1865; Eng. trans., *An Introduction to the Study of Experimental Medicine*, 1927, reprinted 1961); HERMANN VON HELMHOLTZ, *Popular Scientific Lectures* (1962; reprint of a selection of lectures from *Popular Lectures on Scientific Subjects*, 1st and 2nd series, 1881); T.H. HUXLEY, *Science and Culture, and Other Essays* (1882); and WILLIAM WHEWELL, *The Philosophy of the Inductive Sciences, Founded upon Their History*, 2nd ed. (1847, reprinted 1966). RYAN D. TWENEY, MICHAEL E. DOHERTY, and CLIFFORD R. MYNATT (eds.), *On Scientific Thinking* (1981), consists of selections from the writings of Bacon, Newton, Descartes, Darwin, and Einstein; EDWARD GRANT, *Much Ado About Nothing: Theories of Space and Vacuum from the Middle Ages to the Scientific Revolution* (1981), is a study of the origins of modern science and scientific thinking; MORRIS BERMAN, *The Reenchantment of the World* (1981), is a controversial analysis of scientific tradition with alternative suggestions for understanding the modern world.

The background to the 20th-century debate in the philosophy of science, as it developed between 1890 and 1920, can be reconstructed from ERNST MACH, *Die Mechanik in ihrer Entwicklung historisch-kritisch Dargestellt*, 9th ed. (1933; Eng. trans., *The Science of Mechanics*, 1960); and *Die Analyse der Empfindungen und das Verhältnis des Physischen zum Psychischen*, 5th ed. (1906; Eng. trans., *The Analysis of Sensations and the Relation of the Physical to the Psychical*, 1959); KARL PEARSON, *The Grammar of Science*, 3rd ed. (1960); PIERRE DUHEM, *La Théorie physique: son objet et sa structure* (1906; Eng. trans., *The Aim and Structure of Physical Theory*, 1954); HENRI POINCARE, *La Science et l'hypothèse* (1903; Eng. trans., *Science and Hypothesis*, 1905); and HEINRICH HERTZ, *Die Prinzipien der Mechanik* (1895; Eng. trans., *The Principles of Mechanics*, 1889, reprinted 1956), especially the Introduction. The subsequent course of the debate, between 1920 and 1960, is illustrated in CARL G. HEMPEL, *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science* (1965); RUDOLF CARNAP, *Logical Foundations of Probability*, 2nd ed. (1962); KARL R. POPPER, *Logik der Forschung* (1935; 2nd ed., 1966; Eng. trans., *The Logic of Scientific Discovery*, 1959); ROBIN G. COLLINGWOOD, *An Essay on Metaphysics* (1940); NORMAN R. CAMPBELL, *What Is Science?* (1921); and WILLIAM H. WATSON, *On Understanding Physics* (1959). Two significant side-views of the debate from very different formal standpoints may be found in WILLIAM V. QUINE, *From a Logical Point of View*, 2nd ed. rev. (1961); and RONALD A. FISHER, *Statistical Methods for Research Workers*, 14th ed. (1970). PETER ACHINSTEIN and STEPHEN F. BARKER (eds.), *The Legacy of Logical Positivism* (1969), is a first attempt at an historical appraisal of the period. PAUL K. FEYERABEND, *Philosophical Papers*, 2 vol. (1981), is a brilliant examination of historical and philosophical evidence of rationality of science; RICHARD L. GREGORY, *Mind in Science:*

A History of Explanations in Psychology and Physics (1981), is a study of philosophical ideas generated in different historical periods; MANFRED EIGEN and RUTHILD WINKLER, *Laws of the Game: How the Principles of Nature Govern Chance* (1981; trans. from the German), is a discussion of scientific topics from evolution to information theory; LIEBE F. CAVALIERI, *The Double-Edged Helix: Science in the Real World* (1981), deals with sociophilosophical traits of science and technology.

The current debates about conceptual change and the rationale of discovery in science are central topics in KARL R. POPPER, *Conjectures and Refutations*, 2nd rev. ed. (1965); NORWOOD R. HANSON, *Patterns of Discovery* (1958); STEPHEN TOULMIN, *Human Understanding*, vol. 1 (1972); THOMAS S. KUHN, *The Structure of Scientific Revolutions*, 2nd ed. rev. (1970); IMRE LAKATOS and ALAN MUSGRAVE, *Criticism and the Growth of Knowledge* (1970); and MICHAEL POLANYI, *Personal Knowledge* (1958). For broader cultural aspects of the contemporary debate about science, see C.H. WADDINGTON, *The Scientific Attitude*, new ed. (1968); JACOB BRONOWSKI, *Science and Human Values*, rev. ed. (1965); THEODORE ROSZAK, *The Making of a Counter Culture* (1969); and LEWIS MUMFORD, *The Myth of the Machine* (1967). See also PETER B. MEDAWAR, *Pluto's Republic* (1982), a discussion of the relationship between scientists and society; JOHN W. HARRINGTON, *Dance of the Continents* (1983), an explanation of the processes of interpretive scientific thinking; HANS-GEORG GADAMER, *Reason in the Age of Science* (1982; trans. from the German), a collection of essays on hermeneutic philosophy; FRITJOF CAPRA, *The Turning Point: Science, Society, and the Rising Culture* (1982), suggesting a new vision of scientific reality; RICHARD HEALEY (ed.), *Reduction, Time, and Reality: Studies in the Philosophy of the Natural Sciences* (1981), a collection of articles by eminent philosophers; and PETER SMITH, *Realism and the Progress of Science* (1982), examining philosophical foundations of the progress of science.

Philosophy of physics: R. HARRE, "Philosophical Aspects of Cosmology," and W. DAVIDSON, "Philosophical Aspects of Cosmology," *Br. J. Phil. Sci.*, 13:104–119, 120–129 (1962); OTHMAR SPANN, *Naturphilosophie*, 2nd ed. (1963); IVOR LECLERC, *The Nature of Physical Existence* (1972); PASCUAL JORDAN, *Albert Einstein: Sein Lebenswerk und die Zukunft der Physik* (1969), and *Atom und Weltall: Einführung in den Gedankeninhalt der modernen Physik*, 2nd ed. (1960); E.T. WHITTAKER, *From Euclid to Eddington: A Study of Conceptions of the External World* (1949); JACQUES MERLEAU-PONTY and BRUNO MORANDO, *Les Trois Étapes de la cosmologie* (1971; Eng. trans., *The Rebirth of Cosmology*, 1975); LOUIS DE BROGLIE, *La Physique nouvelle et les quanta* (1937; Eng. trans., *The Revolution in Physics*, 1953, reprinted 1969); R.G. COLLINGWOOD, *The Idea of Nature* (1945, reprinted 1960); P.K. FEYERABEND, "Philosophie de la nature," in M.F. SCIACCA (ed.), *Les Grands Courants de la pensée mondiale contemporaine*, part. 2, vol. 2, pp. 901–927 (1961); ERROL E. HARRIS, *The Foundations of Metaphysics in Science* (1965); JAGJIT SINGH, *Great Ideas and Theories of Modern Cosmology*, rev. and enlarged ed. (1970); P.A. SCHILPP (ed.), *Albert Einstein: Philosopher–Scientist*, 2 vol. (1959); PHILIPP FRANK, *Philosophy of Science* (1957); ALFRED NORTH WHITEHEAD, *Process and Reality* (1929, corrected ed. 1978), and *The Concept of Nature* (1920, reprinted 1964); MARY B. HESSE, *Forces and Fields* (1961); CARL FRIEDRICH VON WEIZSAECKER, *Zum Weltbild der Physik*, 12th ed. (1976); ADOLPH GRUENBAUM, *Philosophical Problems of Space and Time*, 2nd enlarged ed. (1973); WERNER HEISENBERG, *Physics and Philosophy* (1958); A.S. EDDINGTON, *The Nature of the Physical World* (1928, reprinted 1958); HENRI POINCARE, *La Science et l'hypothèse* (1903; Eng. trans., *Science and Hypothesis*, 1905); HENRY MARGENAU, *The Nature of Physical Reality* (1950, reprinted 1977), and *Physics and Philosophy* (1978). KARL R. POPPER, *Realism and the Aim of Science* (1983), *The Open Universe: An Argument for Indeterminism* (1982), and *Quantum Theory and the Schism in Physics* (1982), comprise the work of one of the greatest philosophers of the 20th century, who challenges most of the traditional assumptions in modern physics; JEREMY BERNSTEIN, *Science Observed: Essays of My Mind* (1982), is a collection of essays on the process of science; ROGER S. JONES, *Physics as Metaphor* (1982) is an original personal commentary on the nature of physical science; HEINZ R. PAGELS, *The Cosmic Code: Quantum Physics as the Language of Nature* (1982), is a very up-to-date review of modern physics; see also FRITJOF CAPRA, *The Tao of Physics: An Exploration of the Parallels Between Modern Physics and Eastern Mysticism*, 2nd rev. ed. (1983); BENJAMIN GAL-OR, *Cosmology, Physics, and Philosophy* (1983); EDWARD M. MACKINNON, *Scientific Explanation and Atomic Physics* (1982); and P.C.W. DAVIES, *The Accidental Universe* (1982).

Philosophy of biology: FRANCISCO JOSC AYALA and THEODOSIUS DOBZHANSKY (eds.), *Studies in the Philosophy of Biology: Reduction and Related Problems* (1974), a collection of conference papers that discuss definitions and the implications of reductionism; M.O. BECKNER, "Biology," in *The Encyclopedia of Philosophy*, vol. 1, pp. 310–318 (1967), an excellent review of some of the basic philosophical problems in biology; LUDWIG VON BERTALANFFY, *General System Theory* (1968), an attempt to unify the disparate aspects of biology and formulate general principles; BARRY COMMONER, "In Defense of Biology," *Science*, 133:1745–1748 (1961), an argument for the autonomy of biology decrying the emphasis on molecular biology to the neglect of general biology; FRANCIS CRICK, *Of Molecules and Men* (1966), a work that expresses the attitudes and positions taken by many molecular biologists in defense of reductionism; W.M. ELSASSER, *The Chief Abstractions of Biology* (1975), a refutation of reductionist analysis in the life sciences; A.G.N. FLEW, *Evolutionary Ethics* (1967), a significant attempt to relate evolutionary theory to ethics. T.A. GOUDGE, *The Ascent of Life* (1961), a work devoted solely to philosophical problems of evolutionary theory; MARJORIE GRENE, *Approaches to a Philosophical Biology* (1968), a review of the European viewpoints on the philosophy of nature (*Naturphilosophie*); THOMAS H. and JULIAN HUXLEY, *Touchstone for Ethics, 1893–1943* (British title, *Evolution and Ethics, 1893–1943;* 1947, reprinted 1969), opposing views on the same theme by two famous biologists; HANS JONAS, *The Phenomenon of Life: Toward a Philosophical Biology* (1966, reprinted 1979), a collection of essays presenting an existential interpretation of biology in the tradition of *Naturphilosophie*; JACQUES MONOD, *Le Hasard et la nécessité: essai sur la philosophie naturelle de la biologie moderne* (1970; Eng. trans., *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*, 1971), a significant addition to the biological literature by a Nobel Prize-winning biologist; ASHLEY MONTAGU (ed.), *Man and Aggression*, 2nd ed. (1973), a collection of essays and reviews critical of the notion of innate aggressiveness as expounded by Konrad Lorenz and Robert Ardrey; RONALD MUNSON (ed.), *Man and Nature: Philosophical Issues in Biology* (1971), a collection of writings that provides an excellent introduction to current thinking in biophilosophy, with incisive commentary by the editor; JEAN PIAGET, *Biologie et connaissance* (1967; Eng. trans., *Biology and Knowledge*, 1971), concerned primarily with the evolution of mental faculties; V.R. POTTER, *Bioethics* (1971), a collection of articles dealing with the influence of biology on practical moral problems. BERNHARD RENSCH. *Biophilosophie auf erkenntnistheoretischer Grundlage* (1968; Eng. trans., *Biophilosophy*, 1971), a work emphasizing the rise and nature of consciousness within the framework of the evolutionary process; MICHAEL A. SIMON, *The Matter of Life* (1971), a good review of the current approaches to the philosophy of biology; GEORGE GAYLORD SIMPSON, *This View of Life: The World of an Evolutionist* (1964), a collection of essays by a renowned scholar, some of which have shaped the direction of philosophical biology today, and the especially pertinent chapter, "Man's Place in Nature," in his *Meaning of Evolution*, pp. 281–294, rev. ed. (1967); PIERRE TEILHARD DE CHARDIN, *Le Phénomène humain* (1956; Eng. trans., *The Phenomenon of Man*, 1959), a controversial attempt to meld science and religion by presenting a mystical viewpoint of cosmic evolution; C.H. WADDINGTON, *The Ethical Animal* (1960), a defense of the claim that biological facts can provide grounds for moral principles; STEVEN ROSE (ed.), *Against Biological Determinism* (1982) and *Towards a Liberatory Biology* (1982), are two volumes of essays collected by the DIALECTICS OF BIOLOGY GROUP, based on proceedings of meetings of philosophers and biologists and exploring alternatives in modern life science; CHARLES BIRCH and JOHN B. COBB, *The Liberation of Life: From the Cell to the Community* (1982), is a survey of philosophical interpretations of problems of ecology; JOHN C. GREENE, *Science, Ideology, and World View: Essays in the History of Evolutionary Ideas* (1981), is a discussion of the role of philosophy of evolution in Western intellectual history; C. LEON HARRIS, *Evolution: Genesis and Revelations: With Reading from Empodocles and Wilson* (1981), is a survey of the history of evolutionary thought; and U.J. JENSEN and R. HARRÉ (eds.), *The Philosophy of Evolution* (1981), is a collection of essays on evolutionary concepts in different fields; see also ERNST MAYR, *The Growth of Biological Thought: Diversity, Evolution, and Inheritance* (1982).

Philosophy of history: ROBIN G. COLLINGWOOD, *The Idea of History* (1946), a classical contribution to the critical theory of history; KARL R. POPPER, *The Poverty of Historicism* (1957), an influential critique of types of historical speculation; WILLIAM H. WALSH, *An Introduction to Philosophy of History* (1951), a lucid general account; PATRICK L. GARDINER, *The Nature of Historical Explanation*, 3rd ed. (1967); WILLIAM H. DRAY, *Laws and Explanation in History* (1957); PETER WINCH, *The Idea of a Social Science and Its Relation to Philosophy* (1958); and MORTON G. WHITE, *Foundations of Historical Knowledge* (1965), four analytical discussions relating to historical knowledge and understanding; FRANK E. MANUEL, *Shapes of Philosophical History* (1965), a brief but reliable survey of the development

of speculative theories. WILLIAM DRAY, *Perspectives in History* (1980), is an explorative collection of essays on particular theorists of history by a prominent contemporary philosopher; ANTHONY GIDDENS, *A Contemporary Critique of Historical Materialism: Power, Property, and the State* (1982), is a work by a leading British sociologist in which he offers an alternative interpretation of history based upon contemporary anthropological research; HILARY PUTNAM, *Reason, Truth and History* (1981), is a discussion of the nature of truth, knowledge, and order, especially the dichotomy between objective and subjective views of truth; LAWRENCE STONE, *The Past and the Present* (1981), is an examination of historiographic methodology; JOHN W. MILLER, *The Philosophy of History with Reflections and Aphorisms* (1981), is an original personal reflection on history as a human science; see also discussions of special issues in MARK BLITZ, *Heidegger's Being and Time and the Possibility of Political Philosophy* (1981); ALFRED SCHMIDT, *History and Structure: An Essay on Hegelian-Marxist and Structuralist Theories of History* (1982; originally published in German, 1971); EDMUND E. JACOBITTI, *Revolutionary Humanism and Historicism in Modern Italy* (1981); EUGENE WEBB, *Eric Voegelin: Philosopher of History* (1981).

Philosophy of religion: General introductions include H.D. LEWIS, *Philosophy of Religion* ("Teach Yourself Book") (1965); and J. HICK, *Philosophy of Religion* (1963). Introductory books that range over a somewhat narrower field include J.L. GOODALL, *An Introduction to the Philosophy of Religion* (1966); I.T. RAMSEY, *Religious Language* (1957 and 1963); and THOMAS FAWCETT, *The Symbolic Language of Religion* (1971). NINIAN SMART, *Philosophers and Religious Truth*, 2nd ed. (1969), centres discussion of some salient issues around particular philosophers. NINIAN SMART, *The Religious Experience of Mankind* (1969); and EDWARD GEOFFREY PARRINDER, *Comparative Religion* (1962), provide general introductions to the comparative study of religions. In *World Religions* (1966), H.D. LEWIS and R.L. SLATER consider issues in the world religions that are highlighted by contemporary approaches in the philosophy of religion and the comparative study of religions respectively. In the psychology of religion, the classic work of WILLIAM JAMES, *The Varieties of Religious Experience* (1902, reprinted 1952), is probably still the best introduction that might then be followed by a comprehensive survey of the contemporary field, such as L.W. GRENSTED, *The Psychology of Religion* (1952). Most of these books contain excellent bibliographies for further reading. NINIAN SMART, *Historical Selections in the Philosophy of Religion* (1962); and I.T. RAMSEY, *Words About God* (1971), are useful sourcebooks for some classic discussions of topics in the philosophy of religion; an excellent survey of recent thought is given in JOHN MACQUARRIE, *Twentieth-Century Religious Thought: The Frontiers of Philosophy and Theology, 1900–1960* (1963). F.R. TENNANT, *Philosophical Theology*, 2 vol. (1928–30); H.H. FARMER, *The World and God: A Study of Prayer, Providence and Miracle in Christian Experience* (1935); and H.D. LEWIS, *Our Experience of God* (1959), represent different general treatments of the subject. WILFRED C. SMITH, *Towards a World Theology: Faith and the Comparative History of Religion* (1980), is an introduction to analysis of the interrelationship of various religious traditions; JOHN B. COBB and W. WIDICK SCHROEDER (eds.), *Process Philosophy and Social Thought* (1981), is a collection of essays on reshaping of social and political thinking; MICHAEL D. CLARK, *Worldly Theologians: The Persistence of Religion in Nineteenth Century American Thought* (1981), is an examination of the relationship between religion and secular thought; SEYYED HOSSEIN NASR, *Knowledge and the Sacred* (1982), is an exposition of neo-traditionalism in philosophy and religion; FREDERICK COPLESTON, *Religion and the One: Philosophies East and West* (1982), is a comparison of various world philosophies in their treatment of human religious experience. For discussion of particular religious philosophers see TERENCE J. GERMAN, *Hamann on Language and Religion* (1982), ROBERT J. VANDEN BURGT, *The Religious Philosophy of William James* (1981); HENRY S. LEVINSON, *The Religious Investigations of William James* (1981); JEFFREY STOUT, *The Flight from Authority: Religion, Morality, and the Quest for Autonomy* (1981).

Books dealing with specific problems of religious belief are: PETER R. BAELZ, *Prayer and Providence* (1968); J. HICK, *Evil and the God of Love* (1966); A. FARRER, *Love Almighty and Ills Unlimited* (1966); I.T. RAMSEY, *The Problem of Evil* (1972); W.T. STACE, *Mysticism and Philosophy* (1960); JOHN HICK, *God Has Many Names* (1982); LESZEK KOLAKOWSKI, *Religion: If There Is No God: On God, the Devil, Sin, and Other Worries of the So-Called Philosophy of Religion* (1982); HARRY J. AUSMUS, *The Polite Escape: On the Myth of Secularization* (1982); and CHARLES F. KEYES and E. VALENTINE DANIEL (eds.), *Karma: An Anthropological Inquiry* (1983).

Philosophy of art: PAUL ZIFF, "The Task of Defining a Work of Art," *Phil. Rev.*, 62:58–78 (1953); DEWITT PARKER, *The Analysis of Art* (1926); SUSANNE K. LANGER, *Feeling and Form* (1953); JOHN DEWEY, *Art As Experience* (1934); WARREN E. STEINKRAUS, *Philosophy of Art*, rev. ed. (1984); V.A. HOWARD, *Artistry: The Work of Artists* (1982); R.A. SHARPE, *Contemporary Aesthetics: A Philosophical Analysis* (1983); PATRICIA H. WERHANE (ed.), *Philosophical Issues in Art* (1984); and HAROLD ROSENBERG, *The De-definition of Art* (1983).

For a discussion of intention in art interpretation see first the classic attack on intention by WILLIAM K. WIMSATT and MONROE C. BEARDSLEY, "The Intentional Fallacy," in *The Verbal Icon* (1954): for an opposed view see LESLIE A. FIEDLER, "Archetype and Signature: A Study of the Relationship Between Biography and Poetry," *Sewanee Rev.*, 60: 253–273 (1952); and HENRY D. AIKEN, "The Aesthetic Relevance of the Artist's Intentions," *J. Phil.*, 52:742–753 (1955). Classic sources on art as imitation (representation) are ARISTOTLE, *Poetics*, and LONGINUS, *On the Sublime*. Recent works include WALTER ABELL, *Representation and Form* (1936), particularly in visual art; THEODORE M. GREENE, *The Arts and the Art of Criticism* (1940); and STEPHEN C. PEPPER, *The Work of Art* (1955). See also RICHARD KUHNS, *Psychoanalytic Theory of Art: A Philosophy of Art on Developmental Principles* (1983); and HOWARD J. SMAGULA, *Currents, Contemporary Directions in the Visual Arts* (1983).

A classic source on art as expression is LEO TOLSTOY'S *What Is Art?* (1898; Aylmer Maude translation in "Oxford World's Classic Library," 1930, reprinted 1960); defense of forms of expression theory are ROBIN G. COLLINGWOOD, *The Principles of Art* (1938); and CURT J. DUCASSE, *The Philosophy of Art*, rev. ed. (1966). An anthology of readings on artistic expression is JOHN HOSPERS (ed.), *Artistic Expression* (1971). Later monographs include R.L. HELD, *Endless Innovations: Frederick Kiesler's Theory and Scenic Design* (1982); and MARCIA MUELDER EATON, *Art and Nonart* (1983).

For a defense of art as form, see CLIVE BELL, *Art* (1914, reprinted 1958); ROGER FRY, *Transformations* (1927); EDUARD HANSLICK, *Vom Musikalisch-Schönen*, 7th ed. rev. (1885; Eng. trans., *The Beautiful in Music*, 1891, reprinted 1957). Moral and societal aspects of the subject are treated in GEORGE SANTAYANA, *Reason in Art* (1910). ROGER TAYLOR, *Beyond Art: What Art Is and Might Become If Freed from Cultural Elitism* (1981), is a discussion of art as an integral part of life; ARTHUR C. DANTO, *The Transfiguration of the Commonplace: A Philosophy of Art* (1981), is an examination of the relationship between philosophy and art, and of the distinction between art works and commonplace objects. DENNIS DUTTON (ed.), *The Forger's Art: Forgery and the Philosophy of Art* (1983), is a collection of original essays on this special topic.

Histories of the subject include ALBERT HOFSTADTER and RICHARD KUHNS (eds.), *Philosophies of Art and Beauty* (1964); ALEXANDER SESONSKE (ed.), *What Is Art?* (1965); and MONROE C. BEARDSLEY, *Aesthetics from Classical Greece to the Present* (1966). See also EUGENIO TRÍAS, *The Artist and the City* (1982; originally published in Spanish, 1976); MICHAEL F. PALMER, *Paul Tillich's Philosophy of Art* (1984); MIKEL DUFRENNE, *Main Trends in Aesthetics and the Science of Art* (1979); and STEPHEN DAVID ROSS (ed.), *Art and Its Significance: An Anthology of Aesthetic Theory* (1984).

Philosophy of logic: W.V. QUINE, *Philosophy of Logic* (1970), is the best compact introductory exposition. HILARY PUTNAM, *Philosophy of Logic* (1971), is useful as a complement to Quine. Much of the important recent literature, however, is in the form of brief papers rather than monographs. The most successful anthologies of such papers are perhaps L.W. SUMNER and JOHN WOODS (eds.), *Necessary Truth* (1969), on logical truth and analyticity; and LEONARD LINSKY (ed.), *Reference and Modality* (1971), on modal concepts and intensional logic. RAYMOND KLIBANSKY (ed.), *Contemporary Philosophy*, vol. 1 (1968), contains several survey articles covering thoroughly the whole field. Still central are the classical writings of the great modern philosophers of logic: PETER GEACH and MAX BLACK (eds.), *Translations from the Philosophical writings of Gottlob Frege*, 2nd ed. (1960); BERTRAND RUSSELL, *Logic and Knowledge* (1956); LUDWIG WITTGENSTEIN, *Tractatus logico-philosophicus* (Eng. trans., 1922) and *Philosophical Investigations* (Eng. trans., 1953); ALFRED TARSKI, *Logic, Semantics, and Metamathematics* (1956); and RUDOLF CARNAP, *Meaning and Necessity* (1947). The period 1879–1931 is also covered in a magnificent volume by JEAN VAN HEIJENOORT (ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931* (1967), which contains valuable introductions to the different selections and comments on them. Of recent literature, especially noteworthy are the writings of Strawson and Quine: P.F. STRAWSON, *Logico-Linguistic Papers* (1971); W.V. QUINE, *From a Logical Point of View* (1953), *Word and Object* (1960), and *Ontological Relativity and Other Essays* (1969). Quine's ideas are discussed critically in DONALD DAVIDSON and JAAKKO HINTIKKA (eds.), *Words and Objections* (1969). A broad spectrum of new work on the borderline of philosophical logic and

linguistics is represented in DONALD DAVIDSON and GILBERT HARMAN (eds.), *Semantics of Natural Language* (1972). Several problems mentioned above are discussed in JAAKKO HINTIKKA, *Logic, Language-Games, and Information* (1972); and in J.W. DAVIS, D.J. HOCKNEY, and W.K. WILSON (eds.), *Philosophical Logic* (1969). Later monographs include JOHN MARENBON, *From the Circle of Alcuin to the School of Auxerre: Logic, Theology, and Philosophy in the Early Middle Ages* (1981), an examination of the progress of medieval philosophical logic; RUDY RUCKER, *Infinity and the Mind: The Science and Philosophy of the Infinite* (1982), a book on the interface of philosophy and computer science; GEORGE BEALER, *Quality and Concept* (1982), an appraisal of elementary symbolic logic; and OSWALD HANFLING, *Logical Positivism* (1981), a treatment of the philosophical movement.

Philosophy of law: Leading treatises in English include A. ROSS, *Om ret og retfaerdighed* (1953; Eng. trans., *On Law and Justice*, 1958); W. FRIEDMANN, *Legal Theory*, 5th ed. (1967); H. KELSEN, *General Theory of Law and State* (1945); G.W. PATON, *A Text-Book of Jurisprudence*, 3rd ed. by D.P. DERHAM (1964); R. POUND, *Jurisprudence*, 5 vol. (1959); J. STONE, *Legal System and Lawyers' Reasonings* (1964); *Human Law and Human Justice* (1965); *Social Dimensions of Law and Justice* (1966); LEV S. JAWITSCH, *The General Theory of Law: Social and Philosophical Problems* (1981; trans. from the Russian); ROBERT S. SUMMERS, *Instrumentalism and American Legal Theory* (1982). NEIL MACCORMICK, *H.L.A. Hart* (1981), is an analysis of the legal philosophy of the prominent jurist.

Notable works on the history of philosophy of law include E. WOLF, *Griechisches Rechtsdenken*, 3 vol. (1950–54); A. VERDROSS, *Abendländische Rechtsphilosophie*, 2nd ed. (1963); R. MARCIC, *Geschichte der Rechtsphilosophie* (1971); G. FASSO, *Storia della filosofia del diritto*, 2 vol. (1966–68; 2nd ed., vol. 1, 1970). DAVID MILLER, *Philosophy and Ideology in Hume's Political Thought* (1981), is a comprehensive survey of Hume's thought on judgment, justice, political power, and government.

Basic works in the main areas include: (*Analytical jurisprudence*): J. AUSTIN, *Lectures on Jurisprudence* (1832); H.L.A. HART, *The Concept of Law* (1961); W.N. HOHFELD, *Fundamental Legal Conceptions As Applied to Judicial Reasoning* (1923); H. KELSEN, *Reine Rechtslehre*, 2nd ed. (1960; Eng. trans., *Pure Theory of Law*, 1967); A. KOCOUREK, *Jural Relations*, 2nd ed. (1928); I. TAMMELO, *Outlines of Modern Legal Logic* (1969). (*Ethical jurisprudence*): C.K. ALLEN, *Aspects of Justice* (1958); J. BENTHAM, *Introduction to the Principles of Morals and Legislation* (1780); E. BODENHEIMER, *Treatise on Justice* (1967); H. KELSEN, *What is Justice?* (1957); C. PERELMAN, *Justice* (1967); G. RADBRUCH, "Legal Philosophy," in K. WILK (ed.), *The Legal Philosophies of Lask, Radbruch and Dabin*, pp. 47–226 (1950). (*Sociological jurisprudence, including historical jurisprudence*): W. AUBERT (ed.), *Sociology of Law* (1969); E. EHRLICH, *Grundlegung der Soziologie des Rechts* (1913; Eng. trans., *Fundamental Principles of the Sociology of Law*, 1936); L.M. FRIEDMAN and A.S. MACAULAY (eds.), *Law and Behavioral Sciences* (1969); G. GURVITCH, *Sociology of Law* (1942); K.N. LLEWELLYN, *Jurisprudence* (1962); H. MAINE, *Ancient Law* (1861); F. VON SAVIGNY, *Vom Beruf unsrer Zeit für Gesetzgebung und Rechtswissenschaft* (1830; Eng. trans., *Of the Vocation of Our Age for Legislation and Jurisprudence*, 1831); E.M. SCHUR, *Law and Society* (1968); J. STONE, *Law and the Social Sciences in the Second Half-Century* (1966). See also DONALD DAVIDSON, *Essays on Actions and Events* (1980); HANNU TAPANI KLAMI, *Anti-Legalism* (1980); M.A. STEWART (ed.), *Law, Morality, and Rights* (1983); JOHN D. HODSON, *The Ethics of Legal Coercion* (1983); ANTHONY ALLOTT, *The Limits of Law* (1980); B.R. KRISHNA IYER, *Law Versus Justice* (1981).

Philosophy of education: For an excellent example of the analytical approach to philosophy of education, see ISRAEL SCHEFFLER (ed.), *Philosophy and Education*, 2nd ed. (1966). A representative British example is P.H. HIRST and R.S. PETERS, *The Logic of Education* (1970). Several scholars have attempted to group philosophies of education into schools of thought. The best collections are JOHN PAUL STRAIN (comp.), *Modern Philosophies of Education* (1971); and THEODORE B.H. BRAMELD, *Patterns of Educational Philosophy: Divergence and Convergence in Culturological Perspective* (1971). For examples of attempts to focus on a single problem or issue and to combine clarity and commitment, see PAUL NASH, *Authority and Freedom in Education: An Introduction to the Philosophy of Education* (1966); KENNETH D. BENNE, *Education in the Quest for Identity and Community* (1962); and ROBERT ULICH, *The Human Career: A Philosophy of Self-Transcendence* (1955). A good collection of Existentialist statements on education may be found in MAXINE GREENE (comp.), *Existential Encounters for Teachers* (1967). For selections from the classics on conceptions of the educated person, see PAUL NASH (ed.), *Models of Man: Explorations in the Western Educational Tradition* (1968); and ROBERT ULICH (ed.), *Three Thousand Years of Educational Wisdom*, 2nd ed. enl. (1954). Original essays on all of these major figures may be found in PAUL NASH, ANDREAS M. KAZAMIAS, and HENRY J. PERKINSON (eds.), *The Educated Man: Studies in the History of Educational Thought* (1965). Contemporary problems are treated in BRENDA COHEN, *Education and the Individual* (1982), exploring the concept of freedom of education; MORTIMER J. ADLER, *The Paideia Proposal: An Educational Manifesto* (1982); HENRY A. GIROUX, *Ideology, Culture, and the Process of Schooling* (1981), espousing radical Marxist thought; LLOYD DUCK, *Teaching with Charisma* (1981), analyzing psychology of teaching behaviours; R.F. HOLLAND, *Against Empiricism: On Education, Epistemology, and Value* (1980), a collection of essays of broad concern; LAWRENCE KOHLBERG, *The Philosophy of Moral Development: Moral Stages and the Idea of Justice* (1981), stressing an interdisciplinary approach in education; MICHAEL R. MATTHEWS, *The Marxist Theory of Schooling: A Study of Epistemology and Education* (1981); and DAVID NYBERG, *Power Over Power: What Power Means in Ordinary Life, How It Is Related To Acting Freely, and What It Can Contribute to a Renovated Ethics of Education* (1981).

The History of Western Philosophy

This article has three basic purposes: (1) to provide an overview of the history of philosophy in the West; (2) to relate philosophical ideas and movements to their historical background and to the cultural history of their time; and (3) to trace the changing conception of the definition, the function, and the task of philosophy.

The article is divided into the following sections:

The nature of philosophy and the writing of its history 733
- Philosophy in the Western tradition 733
- General considerations 734
  - The writing of the history of philosophy
  - Shifts in the focus and concern of philosophy

Ancient Greek and Roman philosophy 735
- The Pre-Socratic philosophers 735
  - Cosmology and the metaphysic of matter
  - Epistemology of appearance
  - Metaphysic of number
  - Anthropology and relativism
- The seminal thinkers of Greek philosophy 738
  - Socrates
  - Plato
  - Aristotle
- Hellenistic and Roman philosophy 740
  - Stoics
  - Epicureans
  - Skeptics
  - Neo-Pythagoreans and Neoplatonists

Medieval philosophy 742
- Early medieval philosophy 742
- Transition to Scholasticism 743
- The age of the Schoolmen 744
- Philosophy in the late Middle Ages 745

Modern philosophy 746
- The Renaissance and early modern period 746
  - Dominant strands of Renaissance philosophy
  - Rise of Empiricism and Rationalism
  - Literary forms and sociological conditions
- The Enlightenment 751
  - Classical British Empiricism and its basic tasks
  - Nonepistemological movements in the Enlightenment
  - Critical examination of reason in Kant
  - Literary forms
- The 19th century 754
  - German Idealism of Fichte, Schelling, and Hegel
  - Positivism and social theory in Comte, Mill, and Marx
  - Independent and irrationalist movements
- The 20th century 757
  - Individual philosophies of Bergson, Dewey, and Whitehead
  - Marxist thought
  - Analytic philosophy
    - The formalist tradition
    - The informalist tradition
  - Continental philosophy
    - The Phenomenology of Husserl and Heidegger
    - The Existentialism of Jaspers and Sartre
    - Postwar trends

Bibliography 760

The nature of philosophy and the writing of its history

PHILOSOPHY IN THE WESTERN TRADITION

Various conceptions of philosophy

It is a paradox faced by all of those who attempt to write the history of philosophy that the "philosophy" whose history they write probably would not have been defined exactly alike by any two of the major figures whom they judge it fitting to include in their accounts. For throughout its long and varied history in the West, "philosophy" has meant many different things. Some of these have been a search for the wisdom of life (the meaning closest to the Greek words from which the term is derived); an attempt to understand the universe as a whole; an examination of man's moral responsibilities and social obligations; an effort to fathom the divine intentions and man's place with reference to them; an effort to ground the enterprise of natural science; a rigorous examination of the origin, extent, and validity of men's ideas; an exploration of the place of will or consciousness in the universe; an examination of the values of truth, goodness, and beauty; an effort to codify the rules of human thought in order to promote rationality and the extension of clear thinking. Even these do not exhaust the meanings that have been attached to the philosophical enterprise, but they give some idea of its extreme complexity and many-sidedness.

It is difficult to determine whether any common element can be found within this diversity and whether any core meaning can be discovered for philosophy that could serve as a universal and all-inclusive definition. A first attempt in this direction might be to define philosophy either as "a reflection upon the varieties of human experience" or as "the rational, methodical, and systematic consideration of those topics that are of greatest concern to man." Vague and indefinite as such definitions are, they do suggest two important facts about philosophizing: (1) that it is a reflective, or meditative, activity and (2) that it has no explicitly designated subject matter of its own but is a method or type of mental operation (like science or history) that can take any area or subject matter or type of experience as its object. Thus, although there are a few single-term divisions of philosophy of long standing—such as logic, ethics, epistemology (the theory of knowledge), or metaphysics (theory of the nature of Being)—its divisions are probably best expressed by phrases that contain the preposition "of"—such as philosophy of nature, philosophy of mind, philosophy of law, or philosophy of art.

Part of what makes it difficult to find a consensus among philosophers about the definition of their discipline is precisely that they have frequently come to it from different fields, with different interests and concerns, and that they therefore have different areas of experience upon which they find it especially necessary or meaningful to reflect. St. Thomas Aquinas (a Dominican friar of the 13th century), George Berkeley (a bishop of the Irish Church in the 18th century), and Søren Kierkegaard (a Danish divinity student in the 19th century) all saw philosophy as a means to assert the truths of religion and to dispel the Materialistic or Rationalistic errors that, in their opinion, had led to its decline. Pythagoras in ancient south Italy, René Descartes in the late Renaissance, and Bertrand Russell in the 20th century were primarily mathematicians whose views of the universe and of human knowledge were vastly influenced by the concept of number and by the method of deductive thinking. Some philosophers, such as Plato or the British philosophers Thomas Hobbes and John Stuart Mill, were obsessed by problems of political arrangement and social living, so that whatever else they did in philosophy was stimulated by a desire to understand and, ultimately, to change the social and political behaviour of men. And still others—such as the Milesians (the first philosophers of Greece); Francis Bacon, an Elizabethan philosopher; and, in the 20th century, Alfred North Whitehead, a process metaphysician—began with an interest in the physical composition of the natural world, so that their philosophies resemble more closely the generalizations of physical science than those of religion or sociology.

Characteristic oppositions in philosophy

The history of Western philosophy reveals in detail the concentrated activity of a multitude of serious and able men reflecting upon, reasoning about, and considering deeply the nature of their experience. But throughout this diversity certain characteristic oppositions habitually recur, such as the division between monists, dualists, and pluralists in metaphysics; between Materialists and Idealists in cosmological theory; between Nominalists and Realists in the theory of signification; between Rational-

ists and Empiricists in the theory of knowledge; between Utilitarians, self-realizationists, and proponents of duty in moral theory; and between partisans of logic and partisans of emotion in the search for a responsible guide to the wisdom of life.

Many of these fundamental oppositions among philosophers will be treated in the article that follows. But if any single opposition is taken as central throughout the history of Western philosophy at every level and in every field, it is probably that between the critical and the speculative impulses. These two divergent motivations tend to express themselves in two divergent methods: that of analysis and that of synthesis. Plato's *Politeia* (*The Republic*) is an example of the second. The *Principia Ethica* (1903) of G.E. Moore, a founder of linguistic philosophy, is an example of the first. Beginning with a simple question about justice, *Politeia* in its discursiveness slowly but progressively brings more and more areas into the discussion: first ethics, then politics, then educational theory, then theory of knowledge, and finally metaphysics. Starting with one specific question, Plato finally managed to make his discussion as broad as the world. *Principia Ethica* does just the opposite. Beginning with a general question—What is good?—it progressively breaks up this question into a whole series of subordinate questions, analyzing meanings ever more minutely, growing narrower and narrower but always with the utmost modesty and sincerity, striving for increasing simplicity and exactitude.

The analytic, or critical, impulse treats any subject matter or topic by concentrating upon the part, by taking it apart in the service of clarity and precision. It was essentially the method of Aristotle and of Peter Abelard, a medieval Scholastic; of David Hume, a Scottish Skeptic, and of Rudolf Carnap, a 20th-century semantic Positivist; and of Russell and Moore. The synthetic, or speculative, impulse operates by seeking to comprehend the whole, by putting it all together in the service of unity and completeness. It is essentially the method of Parmenides, a Pre-Socratic monist, and of Plato; of St. Thomas and of Benedict de Spinoza, a modern Jewish Rationalist; of G.W.F. Hegel, a German Idealist, and of Whitehead. Throughout philosophy's history, each of the two traditions has made its insistent claim.

There is one philosophical tradition—that of Positivism—that sees philosophy as originating in the obscure mists of religion and coming finally to rest in the pure sunshine of scientific clarity. This represents a necessary progress, because Positivism considers it a scandal when philosophers speak a language that is not accessible to "verification"; it holds that bold and adventuresome philosophical speculation is at best mere self-indulgence, a passing state occurring when philosophical problems are raised prematurely—that is, at a time when philosophy does not possess the means to solve them.

Though Positivism represents a partisan view that it is not necessary to hold, it does express indirectly a basic truth—that the philosophical enterprise has always hovered uncertainly between the lure of religious seriousness and that of scientific exactitude. In the earliest philosophers of Greece it is impossible to separate ideas of divinity and the human soul from ideas about the mystery of being and the genesis of material change, and in the Middle Ages philosophy was acknowledged to be "the handmaiden of theology." But the increased secularization of modern culture has largely reversed this trend, and the Enlightenment emphasis upon the separation of nature from its divine creator has increasingly placed philosophical resources at the disposal of those interested in creating a philosophy of science.

Yet philosophy's continuing search for philosophical truth leads it to hope, but at the same time to profoundly doubt, that its problems are objectively solvable. With respect to a total description of Being or a definitive account of the nature of values, only individual solutions now seem possible; and the optimistic hope for objective answers that secure universal agreement must be given up.

In this respect, philosophy seems less like science than like art and the philosopher more like an artist than a scientist, for his philosophical solutions bear the stamp of his own personality, and his choice of arguments reveals as much about himself as his chosen problem. As a work of art is a portion of the world seen through a temperament, so a philosophical system is a vision of the world subjectively assembled. Plato and Descartes, Immanuel Kant, an 18th-century German Idealist, and John Dewey, a U.S. Pragmatist, have given to their systems many of the quaint trappings of their own personalities.

Comparison with art: living classics

But if philosophy is not true in the same sense as science, it is not false in the same sense either; and this gives to the history of philosophy a living significance, which the history of science does not enjoy. In science, the present confronts the past as truth confronts error; thus, for science, the past, even when important at all, is important only out of historical interest. In philosophy it is different. Philosophical systems are never definitively proved false; they are simply discarded or put aside for future use. And this means that the history of philosophy consists not simply of dead museum pieces but of ever-living classics—comprising a permanent repository of ideas, doctrines, and arguments and a continuing source of philosophical inspiration and suggestiveness to those who philosophize in any succeeding age. It is for this reason that any attempt to separate philosophizing from the history of philosophy is both a provincial act and an unnecessary impoverishment of its rich natural resources.

GENERAL CONSIDERATIONS

The writing of the history of philosophy. The writing of the history of philosophy is itself controlled by a series of cultural habits or conventions.

Ways of ordering the history. The ensuing article on the history of Western philosophy is divided into three sections—ancient, medieval, and modern—and this division is so pervasive today that it is difficult to remember that the threefold distinction is only as old as the end of the 17th century. This distinction—first employed in the writing of European history proper by Georg Horn of Leiden in his *Arca Noae: Sive Historia Imperiorum et Regnorum a Condito Orbe ad Nostra Tempora* (1666; "Noah's Ark; or, The History of Empires and Kingdoms from the Beginning of the World to Our Times") and a generation later by Christophorus Cellarius, a German historian, in 1696—slowly spread to historical writing in all fields and was given definitive influence in philosophical writing through the series of lectures on *Philosophiegeschichte* ("History of Philosophy") that Hegel delivered first at Jena, then at Heidelberg, and finally at Berlin between 1805 and 1830. Since Hegel, it has been taken for granted as standard practice, although a host of cultural assumptions is implied by its use.

Stress on ideas versus stress on men

Treatment of the total field of the history of philosophy has been traditionally subject to two types of ordering, according to whether it was conceived (1) as primarily a history of ideas or (2) as a history of the intellectual products of men. In the first ordering, certain ideas, or concepts, are viewed as archetypal (such as matter or mind or doubt); and the condensations occurring within the flow of thought tend to consist of basic types or schools. This ordering has characterized such works as Friedrich Lange's *Geschichte des Materialismus* . . . (1866; *The History of Materialism*), A.C. Ewing's compilation *The Idealist Tradition: From Berkeley to Blanshard* (1957), or Richard H. Popkin's *History of Scepticism from Erasmus to Descartes* (1960). In the second type of ordering, the historian, impressed by the producers of ideas as much as by the ideas themselves—that is, with philosophers as agents—reviews the succession of great philosophical personalities in their rational achievement. This ordering has produced the more customary histories such as Émile Bréhier's *Histoire de la philosophie* (1926–32), Bertrand Russell's *History of Western Philosophy* (1945), and Karl Jaspers' *Die grossen Philosophen* (1957; *The Great Philosophers*).

These two different types of ordering depend for their validity upon an appeal to two different principles about the nature of ideas, but their incidental use may also be influenced by social or cultural factors. Thus the biographers and compilers of late antiquity (among them, Plutarch, Sextus Empiricus, Philostratus, and Clement of Alexan-

dria), impressed by the religious pluralism of the age in which they lived, thought of philosophers, too, as falling into different sects and wrote histories of the Sophists, the Skeptics, the Epicureans, and other such schools; whereas almost 2,000 years later, Hegel—living in a period of Romantic historiography dominated by the concept of the great man in history—deliberately described the history of philosophy as "a succession of noble minds, a gallery of heroes of thought."

Moving between these two ordering principles, the article below will be eclectic (as has come to be the custom), devoting chief attention to outstanding major figures, while joining more minor figures, wherever possible, into the schools or tendencies that they exemplify.

Factors in writing the history of philosophy. The type of ordering suggested above also has some relationship to the more general problems of method in the writing of the history of philosophy. Here there are at least three factors that must be taken into account: (1) the historian must understand how (at least in part) any philosopher's doctrines depend upon those of his predecessors; (2) he must understand that a man's philosophy occurs at a certain point in history and, thus, how it expresses the effects of certain social and cultural circumstances; and (3) he must understand how in part it stems from the philosopher's own personality and situation in life. This is only to say that the history of philosophy, to be at all comprehensive and adequate, must deal with the mutual interplay of ideas, of cultural contexts, and of agents.

Logical, socio-logical, and individual factors

The first factor may be called logical because a given philosophy is, in part, the intellectual response to the doctrines of its forerunners, taking as central the problems given by the current climate of controversy. Thus, many of the details of Aristotle's ethical, political, and metaphysical systems arise in arguments directed against statements and principles of Plato; much of *An Essay Concerning Human Understanding* (1690) by John Locke, an initiator of the Enlightenment, is directed against current Cartesian presuppositions; and the *Nouveaux essais sur l'entendement humain* (1704; *New Essays Concerning Human Understanding*) by Gottfried Wilhelm Leibniz, a broadly learned Rationalist, is, in turn, specifically directed against Locke.

The second factor may be called sociological because it considers philosophy, at least in part, as a direct form of social expression, arising at a certain moment in history, dated and marked by the peculiar problems and crises of the society in which it flourishes. From this perspective, the philosophy of Plato may be viewed as the response of an aristocratic elitism to the immediate threat of democracy and the leveling of values in 5th-century Athens—its social theory and even its metaphysics servicing the thrust toward an aristocratic restoration in the Greek world. Thus, the philosophy of St. Thomas Aquinas may be viewed as an effort toward doctrinal clarification in support of the institution of the medieval Roman Church, as the saint spent his life obediently fulfilling the philosophical tasks set for him by his superiors in the church and the Dominican order. Thus, the philosophy of Kant, with all of its technical vocabulary and rigid systematization, may be viewed as an expression of the new professionalism in philosophy, a clear product of the rebirth of the German universities during the 18th-century Enlightenment.

The third factor may be called biographical, or individual, because, with Hegel, it recognizes that philosophies are generally produced by men of unusual or independent personality, whose systems usually bear the mark of their creators. And what is meant here by the individuality of the philosopher lies less in the facts of his biography (such as the wealth or poverty, the married state or bachelorhood that he shares with other men) than in the essential form and style of his philosophizing. The cool intensity of Spinoza's geometric search for wisdom, the unswerving (if opaque) discursiveness of Hegel's quest for completeness or totality, the relentless and minute analytic search for distinctions and shades of meaning that marks Moore's master passion ("to be accurate—to get everything exactly right"), these qualities mark the philosophical writings of Spinoza, Hegel, and Moore with an unmistakably individual and original character.

Shifts in the focus and concern of philosophy. Any adequate treatment of individual figures in the history of philosophy tries to utilize this threefold division of logical, sociological, and individual factors; but in a synoptic view of the history of philosophy in the West, one is particularly aware of the various shifts of focus and concern that philosophy has sustained and, indeed, of the often profound differences in the way that it defines itself or visualizes its task from age to age or from generation to generation.

Philosophy among the Greeks slowly emerged out of religious awe into wonder about the principles and elements of the natural world. But as the Greek populations more and more left the land to become concentrated in their cities, interest shifted from nature to social living; questions of law and convention and civic values became paramount. Cosmological speculation partly gave way to moral and political theorizing, and the preliminary and somewhat fragmentary questionings of Socrates and the Sophists turned into the great positive constructions of Plato and Aristotle. With the political and social fragmentation of the succeeding centuries, however, philosophizing once again shifted from the norm of civic involvement to problems of salvation and survival in a chaotic world.

Medieval and modern history

The dawn of Christianity brought to philosophy new tasks. Augustine, the philosophical bishop of Hippo, and the Church Fathers used such resources of the Greek tradition as remained (chiefly Platonism) to deal with problems of the creation, of faith and reason, and of truth. New translations in the 12th century made much of Aristotle's philosophy available and prepared the way for the great theological constructions of the 13th century, chiefly those of the Scholastic philosophers Bonaventure, Albertus Magnus, Thomas Aquinas, Roger Bacon, and Duns Scotus. The end of the Middle Ages saw a new flowering of the opposite tendencies in the Nominalist William of Ockham and the mystic Meister Eckehart.

The Middle Ages gave way to the Renaissance. Universalism was replaced by nationalism. Philosophy became secularized. The great new theme was that of the mystery and immensity of the natural world. The best philosophical minds of the 17th century turned to the task of exploring the foundations of physical science, and the symbol of their success—the great system of Newton's physics—turned the philosophers of the Enlightenment to epistemology and to the examination of the human mind that had produced so brilliant a scientific creation. The 19th century, a time of great philosophical diversity, discovered the irrational and in so doing prepared the way for the oppositions between Analysis and Phenomenology and between Positivism and Existentialism that characterize 20th-century philosophy.

Although the foregoing capsule presentation of the history of philosophy in the West follows a strict chronology, it does not do justice to the constant occurrence and recurrence of dominant strands in the history of thought. It would also be possible to write the philosophical history of the Middle Ages simply by noting the complicated occurrence of Platonic and Aristotelian doctrines, of the Renaissance according to the reappearance of Greek Materialism, Stoicism, and Skepticism, and of the 18th century in terms of the competing claims of Rationalist and Empiricist principles. Thus, chronology and the interweaving of philosophical systems cooperate in a history of philosophy. (A.W.L.)

Ancient Greek and Roman philosophy

THE PRE-SOCRATIC PHILOSOPHERS

Cosmology and the metaphysic of matter. Because the earliest Greek philosophers focused their attention upon the origin and nature of the physical world, they are often called cosmologists or naturalists. Though monistic views (which trace the origins of the world to a single substance) prevailed at first, they were soon followed by several pluralistic theories (which trace it to several ultimate substances).

Thales of Miletus

Monistic cosmologies. There is a consensus, dating back at least to the 4th century BC and continuing to the present, that the first Greek philosopher was Thales of

Miletus, who flourished in the first half of the 6th century BC. At that time the word *philosopher* ("lover of wisdom") had not yet been coined. Thales was counted, however, among the Seven Wise Men (Sophoi), whose name derives from a term that then designated inventiveness and practical wisdom rather than speculative insight. Thales showed these qualities by trying to give the mathematical knowledge that he derived from the Babylonians a more exact foundation and by using it for the solution of practical problems—such as the determination of the distance of a ship as seen from the shore or of the height of the Pyramids. Though he was also credited with predicting an eclipse of the Sun, it is likely that he merely gave a natural explanation of one on the basis of Babylonian astronomical knowledge.

Thales was considered the first Greek philosopher because he was the first to give a purely natural explanation of the origin of the world, free from all mythological ingredients. He upheld that everything had come out of water—an explanation based on the discovery of fossil sea animals far inland. His tendency (and that of his immediate successors) to give nonmythological explanations of the origin of the world was undoubtedly prompted by the fact that all of them lived on the coast of Asia Minor surrounded by a number of nations whose civilizations were much farther advanced than that of the Greeks and whose mythological explanations differed greatly both among themselves and from those of the Greeks. It appeared necessary, therefore, to make a fresh start on the basis of what a person could observe and figure out by looking at the world as it presented itself. This procedure naturally resulted in a tendency to make sweeping generalizations on the basis of rather restricted but carefully checked observations.

Anaximander of Miletus

Thales' disciple and successor, Anaximander of Miletus (mid-6th century), tried to give a more elaborate account of the origin and development of the ordered world (the cosmos). According to him, it developed out of the apeiron, something both infinite and indefinite (without distinguishable qualities). Within this apeiron something arose to produce the opposites of hot and cold. These at once began to struggle with each other and produced the cosmos. The cold (and wet) partly dried up (becoming solid earth), partly remained (as water), and—by means of the hot—partly evaporated (becoming air and mist), its evaporating part (by expansion) splitting up the hot into fiery rings, which surround the whole cosmos. Because these rings are enveloped by mist, however, there remain only certain breathing holes that are visible to men, appearing to them as Sun, Moon, and stars. Anaximander was the first to realize that upward and downward are not absolute but that downward means toward the middle of the Earth and upward away from it, so that the Earth had no need to be supported (as Thales had believed) by anything. Starting from Thales' observations, Anaximander tried to reconstruct the development of life in more detail. Life, being closely bound up with moisture, originated in the sea. All land animals, he held, are descendants of sea animals; because the first humans as newborn infants could not have survived without parents, Anaximander believed that they were born within an animal of another kind—specifically, a sea animal in which they were nurtured until they could fend for themselves. Gradually, however, the moisture will be partly evaporated, until in the end all things will have returned into the undifferentiated apeiron, "in order to pay the penalty for their injustice"—that of having struggled against one another.

Anaximander's successor, Anaximenes of Miletus (second half of the 6th century), taught that air was the origin of all things. His position was for a long time thought to have been a step backward because, like Thales, he placed a special kind of matter at the beginning of the development of the world. But this criticism missed the point. Neither Thales nor Anaximander appear to have specified the way in which the other things arose out of the water or apeiron. Anaximenes, however, declared that the other types of matter arose out of air by condensation and rarefaction. In this way, what to Thales had been merely a beginning became a fundamental principle that remained essentially the same through all of its transmutations. Thus, the term arche, which originally simply meant "beginning," acquired the new meaning of "principle," a term that henceforth played an enormous role in philosophy down to the present. This concept of a principle that remains the same through many transmutations is, furthermore, the presupposition of the idea that nothing can come out of nothing and that all of the comings to be and passings away that men observe are nothing but transmutations of something that essentially remains the same eternally. In this way it also lies at the bottom of all of the conservation laws—those of the conservation of matter, of force, and of energy—that have been basic in the development of physics. Though Anaximenes of course did not realize all of the implications of his idea, its importance can hardly be exaggerated.

The first three Greek philosophers have often been called hylozoists because they seemed to believe in a kind of living matter. But this is hardly an adequate characterization. It is, rather, characteristic of them that they did not clearly distinguish between kinds of matter, forces, and qualities nor between physical and emotional qualities. The same entity is sometimes called fire and sometimes the hot. Heat appears sometimes as a force and sometimes as a quality, and again there is no clear distinction between warm and cold as physical qualities and the warmth of love and the cold of hate. To realize these ambiguities is important to an understanding of certain later developments in Greek philosophy.

Xenophanes and Parmenides

Xenophanes of Colophon (born *c.* 560 BC), a rhapsodist and philosophical thinker who emigrated from Asia Minor to Elea in southern Italy, was the first to bring out more clearly what was implied in Anaximenes' philosophy. He criticized the popular notions of the gods, saying that men made their gods in their own image. But, more importantly, he argued that there could be only one God, the ruler of the universe, who must be eternal. For, being the strongest of all beings, he could not have come out of something less strong, nor could he be overcome or superseded by something else, because nothing could arise that is stronger than the strongest. The argument clearly rested on the axiom that nothing can come out of nothing and that nothing that is can really vanish.

This axiom was made more explicit and carried to its extreme consequences by Parmenides of Elea (first half of the 5th century BC), the founder of the so-called school of Eleaticism, of whom Xenophanes has been regarded as the teacher and forerunner. In a philosophical poem Parmenides insisted that "what is" cannot have come into being and cannot pass away because it would have to have come out of nothing or to become nothing, whereas nothing by its very nature does not exist. There can be no motion either; for it would have to be a motion into something that is—which is not possible since it would be blocked—or a motion into something that is not—which is equally impossible since what is not does not exist. Hence everything is solid immobile being. The familiar world, in which things move around, come into being, and pass away, is a world of mere belief (*doxa*). In a second part of the poem, however, Parmenides tried to give an analytical account of this world of belief, showing that it rested on constant distinctions between what is believed to be positive—*i.e.,* to have real being, such as light and warmth—and what is negative—*i.e.,* the absence of positive being, such as darkness and cold.

It is significant that Heracleitus of Ephesus, a contemporary of Parmenides, whose philosophy was later considered to be the very opposite of Parmenides' philosophy of immobile being, came, in some fragments of his work, near to what Parmenides tried to show: the positive and the negative, he said, are merely different views of the same thing; death and life, day and night, or light and darkness are really one.

Pluralistic cosmologies. Parmenides had an enormous influence on the further development of philosophy. Most of the philosophers of the following two generations tried to find a way to reconcile his thesis that nothing comes into being nor passes away with the evidence presented to men by their senses. Empedocles of Acragas (mid-5th century) declared that there are four material elements

(he called them roots of everything) and two forces, love and hate, that did not come into being and would never pass away or increase or diminish. But the elements are constantly mixed with one another by love and again separated by hate. Thus, through mixture and decomposition composite things come into being and pass away. Because he conceived of love and hate as blind forces, Empedocles had to explain how through random motion living beings could emerge. This he achieved by means of a somewhat crude anticipation of the theory of the survival of the fittest. In the process of mixture and decomposition the limbs and parts of various animals would be formed by chance. But they could not survive. Only when by chance they had come together in such a way that they were able to support and reproduce themselves would they survive. It was in this way that the various species were produced and continued to exist.

Anaxagoras of Clazomenae, a 5th-century pluralist, believed that because nothing can really come into being, everything must be contained in everything, but in the form of infinitely small parts. In the beginning all of these particles had been mixed in an even mixture, in which nothing could be distinguished, much like the indefinite apeiron of Anaximander. But then nous, or intelligence, began at one point to set these particles into a whirling motion, foreseeing that in this way they would become separated from one another and then recombine in the most various ways so as to produce gradually the world in which men live. In contrast to the forces assumed by Empedocles, the nous of Anaxagoras is not blind but foresees and intends the production of the cosmos, including living and intelligent beings; but it does not interfere with the process after having started the whirling motion. This is a strange combination of a mechanical and a nonmechanical explanation of the world.

Leucippus and Democritus

By far of greatest importance for the later development of philosophy and physical science was an attempt by the Atomists Leucippus (mid-5th century) and (in the following generation) Democritus to solve the Parmenidean problem. Leucippus found the solution in the assumption that, contrary to Parmenides' argument, the nothing *does* in a way exist, viz., as empty space. There are then, however, only two fundamental principles of the physical world, empty space and filled space—the latter consisting of atoms that, in contrast to those of modern physics, are real atoms; that is, they are absolutely indivisible because nothing can penetrate to split them. On these foundations, laid by Leucippus, Democritus appears to have built a whole system, aiming at a complete explanation of the varied phenomena of the visible world by means of an analysis of its atomic structure. This system begins with elementary physical problems, such as that of why a hard body can be lighter than a softer one. The explanation is that, although the heavier body contains more atoms, they are equally distributed and of round shape; the lighter body, however, has fewer atoms, most of which have hooks by which they form rigid gratings. The system ends with educational and ethical questions. A sound and cheerful man, useful to his fellowmen, is literally well composed. Although destructive passions involve violent long-distance atomic motions, education can help to contain them, creating a better composure. Democritus also developed a theory of the evolution of culture, which influenced later thinkers. Civilization, he thought, is produced by the needs of life, which compel man to work and to make inventions. When life becomes too easy because all needs are met, there is a danger that civilization will decay as men become unruly and negligent.

Epistemology of appearance. All of the post-Parmenidean philosophers, like Parmenides himself, presupposed that the real world is different from the one that men perceive. Thus the problems of epistemology, or theory of knowledge, arose. According to Anaxagoras, everything is contained in everything. But this is not what people perceive. He solved this problem, however, by assuming that, if there is a much greater amount of one kind of particle in a thing than of all other kinds, the latter are not perceived at all. The observation was then made that sometimes different persons or kinds of animals have different perceptions of the same things. He explained this phenomenon by assuming that like is perceived by like. If, therefore, in the sense organ of one person there is less of one kind of stuff than of another, he will perceive the former less keenly than the latter. This reasoning was also used to explain why some animals see better by night and others by daylight. According to Democritus, the atoms have no sensual qualities such as tastes, smells, or colours at all. Thus, he tried to reduce all of them to tactile qualities (explaining a bright white colour, for instance, as sharp atoms hitting the eye like needles), and he made a most elaborate attempt to reconstruct the atomic structure of things on the basis of their apparent sensual qualities.

Zeno of Elea

Also of very great importance in the history of epistemology was Zeno of Elea (mid-5th century), a younger friend of Parmenides. Parmenides had, of course, been severely criticized because of the strange consequences of his doctrine that in reality there is no motion and no plurality either because there is just one solid being. To support him, however, Zeno tried to show that the assumption that there is motion and plurality leads to consequences that are no less strange. This he did by means of his famous paradoxes, saying that the flying arrow rests since it can neither move in the place in which it is nor in a place in which it is not and that Achilles cannot outrun a turtle because when he has reached its starting point, the turtle will have moved to a further point, and so on ad infinitum—that, in fact, he cannot even start running, for, before traversing the stretch to the starting point of the turtle, he will have to traverse half of it and again half of that and so on ad infinitum. All of these paradoxes are derived from the problem of the continuum. Although they have often been dismissed as logical nonsense, many attempts have also been made to dispose of them by means of mathematical theorems, such as the theory of convergent series or the theory of sets. In the end, however, the difficulties inherent in his arguments have always come back with a vengeance, for the human mind is so constructed that it can look at a continuum in two ways that are not quite reconcilable.

Metaphysic of number. All of the philosophies mentioned so far are in various ways historically akin to one another. Toward the end of the 6th century, however, there arose quite independently another kind of philosophy, which only later entered into interrelation with the developments just mentioned: the philosophy of Pythagoras of Samos. Pythagoras traveled extensively in the East and in Egypt and, after his return to Samos, emigrated to southern Italy because of his dislike of the tyranny of Polycrates. At Croton and Metapontum he founded a philosophical society with strict rules and soon gained considerable political influence. He appears to have brought his doctrine of the transmigration of souls from the East. Much more important for the history of philosophy and science, however, was his doctrine that "all things are numbers," which means that the essences and structures of all things can be determined by finding the numerical relations contained in them. Originally, this, too, was a very broad generalization made on the basis of comparatively few observations: for instance, that the same harmonies can be produced with different instruments—strings, pipes, disks, etc.—by means of the same numerical ratios—1:2, 2:3, 3:4—in one-dimensional extensions; the observation that certain regularities exist in the movements of the celestial bodies; and the discovery that the form of a triangle is determined by the ratio of the lengths of its sides. But because the followers of Pythagoras tried to apply their principle everywhere with the greatest of accuracy, one of them—Hippasus of Metapontum—about 450 BC made one of the most fundamental discoveries of the entire history of science, that of incommensurability, viz., that the quantitative relation between the side and diagonal of such simple figures as the square and the regular pentagon cannot be expressed as a ratio of integers. At first sight this discovery seemed to destroy the very basis of the Pythagorean philosophy, and the school thus split into two sections, one of which engaged in rather abstruse numerical speculations while the other succeeded in overcoming the difficulty by ingenious mathematical inven-

Pythagoras of Samos

tions and laid the foundations of all quantitative science. Pythagorean philosophy also exerted a great influence on the development of Plato's thought in his later years.

The speculations described so far constitute in many ways the most important part of the history of Greek philosophy because all of the most fundamental problems of Western philosophy turned up here for the first time and one finds here the formation of a great many concepts that have continued to dominate Western philosophy and science to the present day.

Anthropology and relativism. In the middle of the 5th century BC, Greek thinking took a somewhat different turn through the advent of the Sophists. The name is derived from the verb *sophizesthai*, "making a profession of being inventive and clever," and aptly described the Sophists, who, in contrast to the philosophers mentioned so far, asked money for their instruction. Philosophically they were, in a way, the leaders of a rebellion against the preceding development, which more and more had resulted in the belief that the real world is quite different from the phenomenal world. "What is the sense of such speculations?" they asked, since men do not live in these so-called real worlds. This is the meaning of the pronouncement of Protagoras of Abdera (mid-5th century) that "Man is the measure of all things, of those which are that they are and of those which are not that they are not." For man the world is what it appears to him to be, not something else; and, though he meant man in general, he illustrated it by pointing out that even in regard to an individual man it makes no sense to tell him that it is really warm when he is shivering with cold, because for him it is cold—for him, the cold exists, is there.

The Sophists

His younger contemporary Gorgias of Leontini, famous for his treatise on the art of oratory, made fun of the philosophers in a book *Peri tou mē ontos ē peri physeōs* ("On that which is not, or on Nature"), in which—referring to the "truly existing world," also called "the nature of things"—he tried to prove (1) that nothing exists, (2) that if something existed, man could have no knowledge of it, and (3) that if nevertheless somebody knew it, he could not communicate his knowledge to others.

The Sophists were not only skeptical of what had by then become a philosophical tradition but also of other traditions. On the basis of the observation that different nations have different rules of conduct even in regard to things considered most sacred—such as the relations between the sexes, marriage, and burial—they concluded that most rules of conduct are conventions. What is really important is to be successful in life and to gain influence on others. This they promised to teach. Gorgias was proud of the fact that, having no knowledge of medicine, he was more successful in persuading a patient to undergo a necessary operation than his brother, a physician, who knew when an operation was necessary. The older Sophists, however, were far from openly preaching immoralism. They, nevertheless, gradually came under suspicion because of their sly ways of arguing. One of the later Sophists, however, Thrasymachus of Chalcedon (late 5th century), was bold enough to declare openly that "right is what is beneficial for the stronger or better one"; that is, for the one able to win the power to bend others to his will.

THE SEMINAL THINKERS OF GREEK PHILOSOPHY

Socrates. By many of his contemporaries, Socrates (5th century BC) was also considered to be a Sophist because of his tricky arguments, though he did not teach for money and his aims were entirely different from theirs. Although there is a late tradition according to which Pythagoras invented the word *philosopher,* it was certainly through Socrates—who insisted that he possessed no wisdom but was striving for it—that the term came into general use and was later applied to all earlier serious thinkers. In fact, all of the records of his life and activity left by his numerous adherents and disciples indicate that he never tried to teach anything directly. But he constantly engaged in conversations with everybody—old and young, high and low—trying to bring into the open by his questions the inconsistencies in their opinions and actions. Though he never taught directly, his whole activity rested on two unshakable premises: (1) the principle never to do wrong nor to participate, even indirectly, in any wrongdoing and (2) the conviction that nobody who really knows what is good and right could act against it. He demonstrated his unshakable adherence to the first principle on various occasions and under different regimes. When, after the Battle of Arginusae, the majority of the Athenian popular assembly demanded death without trial for the admirals, Socrates, who on that day happened to be president of the assembly (an office changing daily), refused to put the proposal to the vote because it was wrong to condemn anyone without a fair trial. He refused to do so even though the people threatened him, shouting that it would be terrible if the sovereign people could not do as they pleased.

When, after the overthrow of democracy, the so-called Thirty Tyrants, who tried to involve everybody in their wrongdoings, ordered him to arrest an innocent citizen whose money they coveted, he simply disobeyed. This he did although at that time such disobedience was still more dangerous than disobeying the sovereign people had been at the time of unrestricted democracy. Likewise, in the time of the democracy he pointed out by his questions the inconsistency of allowing oneself to be swayed by the oratory of a good speaker instead of first inquiring into his capability as a statesman, whereas in private life a sensible citizen would not listen to the oratory of a quack but would try to find the best doctor. When, after the overthrow of democracy, the Thirty Tyrants had many people arbitrarily executed, he asked everybody whether a man was a good shepherd who diminished the number of the sheep instead of increasing it and did not cease doing so when Critias, the leader of the Thirty, warned him to take heed not to diminish the number of the sheep by his own, Socrates', person. But the most fundamental inconsistency that he tried to show up everywhere was that most people by their actions showed that they considered what they found to be good, wonderful, and beautiful in others—such as, for instance, doing right at great danger to oneself—not to be good for themselves, and considered to be good for themselves what they despised and condemned in others. Though all of these stands won him the most fervent admiration of many, especially among the young of all classes, it caused also great resentment among leading politicians, whose inconsistencies were shown up publicly by him and his adherents. Though Socrates had survived unharmed through the regime of the Thirty—partly because it did not last long, partly because he was supported by some close relatives of their leader Critias—it was under the restored democracy that he was accused of impiety and of corrupting the youth and finally condemned to death, largely also in consequence of his intransigent attitude during the trial.

After Socrates' death his influence became a dominating one through the greater part of Greek and Roman philosophy down to the end of antiquity and was more or less noticeable even in all of the rest. Many of his adherents—among them Xenophon, a military man and a historian, and Aeschines of Sphettus, one of those present at his death—tried to preserve his philosophical method by writing Socratic dialogues. Some founded schools or sects that perpetuated themselves over long periods of time, Eucleides of Megara emphasizing the theoretical aspects of Socrates' thought, and Antisthenes stressing the independence of the true philosopher from material wants. The latter, through his disciple Diogenes of Sinope, who carried voluntary poverty to the extreme and emphasized freedom from all conventions, became the founder of the sect of the Cynics. Aristippus of Cyrene, traditional founder of the Cyrenaic school, stressed man's independence from material needs in a somewhat different way, declaring that there is no reason why a philosopher should not enjoy material goods as long as he is completely indifferent to their loss. Though Aristippus renounced his son because he led a dissolute life, the school that he founded (through his daughter and his grandson) was hedonistic, holding pleasure to be the good.

Founding of schools and sects

Plato. By far the most important disciple of Socrates, however, was Plato, a scion of one of the most noble Athenian families, who could trace his ancestry back to

the last king of Athens and to Solon, the great social and political reformer.

Life. As a very young man Plato became a fervent admirer of Socrates in spite of the latter's plebeian origin. Contrary to his master, however, who always concerned himself with the attitudes of individuals, he believed in the importance of political institutions. In his early youth he had observed that the Athenian masses, listening to the glorious projects of ambitious politicians, had engaged in foolhardy adventures of conquest, which led in the end to total defeat in the Peloponnesian War. When, in consequence of the disaster, democracy was abolished, Plato at first set great hopes in the Thirty Tyrants—especially since their leader, Critias, was a close relative. But he soon discovered that—to use his own words—the despised democracy had been gold in comparison with the new terror. When the oligarchy was overthrown and the restored democracy, in 399, adopted a new law code—in fact, a kind of written constitution containing safeguards against rash political decisions—Plato again had considerable hope and was even inclined to view the execution of Socrates as an unfortunate incident rather than a logical consequence of the new regime. It was only some years later, when demagogy appeared to raise its head again, that he "despaired and was forced to say that things would not become better in politics unless the philosophers would become rulers or the rulers philosophers." He wrote a dialogue, the *Gorgias,* violently denouncing political oratory and propaganda and then traveled to southern Italy in order to study political conditions there. Again, however, he found the much vaunted dolce vita of the Greeks there, in which the rich lived in luxury exploiting the poor, much worse than the democracy at Athens. But at Syracuse he met a young man, Dion, brother-in-law of the ruling tyrant, Dionysius I, who listened eagerly to his political ideas and promised to work for their realization if any occasion should arise. On his return to Athens, Plato founded the Academy, an institution for the education of philosophers, and in the following years elaborated, besides other dialogues, his great work, *Politeia* (*The Republic*), in which he drew the outlines of an ideal state. Because it is the passions and desires of men that cause all disturbances in society, the state must be ruled by an elite governed exclusively by reason and supported by a class of warriors entirely obedient to them. Both ruling classes must have no individual possessions and no families and lead an extremely austere life, receiving the necessities of life from the working population, which alone is permitted to own private property. The elite receives a rigid education to fit it for its task. At the death of Dionysius I, Dion induced Plato to come to Syracuse again to try to persuade Dionysius' successor to renounce his power in favour of a realization of Plato's ideals. But the attempt failed, and in his later political works, the *Politicus* (*Statesman*) and the *Nomoi* (*Laws*), Plato tried to show that only a god could be entrusted with the absolute powers of the philosopher-rulers of his *Politeia.* Human rulers must be controlled by rigid laws, he held—though all laws are inevitably imperfect because life is too varied to be governed adequately by general rules. But the *Nomoi* still placed strict restrictions on the ownership of property.

Visit to Syracuse

Philosophy. In the field of theoretical philosophy, Plato's most influential contribution was undoubtedly his theory of Ideas, which he derived from Socrates' method in the following way: Socrates, in trying to bring out the inconsistencies in his interlocutors' opinions and actions, had often asked what it is that makes men say that a certain thing or action is good or beautiful or pious or brave; and he had asked what people are looking at when they make such statements. Plato sometimes made Socrates ask what is the Eidos, or the Idea—*i.e.,* the image—that a person has before him when he calls something "good." A definite answer is never given, however, because no abstract definition would be adequate, the purpose being rather to make the interlocutor aware of the fact that he somehow does look at something undefinable when making such statements.

Theory of Ideas

What was at first simply a way of somehow expressing something that is difficult to express developed into a definite theory of Ideas when Plato made the discovery that something similar could be observed in the field of mathematics. No two things in the visible world are perfectly equal, just as there is nothing that is perfectly good or perfectly beautiful. Yet equality is one of the most fundamental concepts not only in mathematics but also in everyday life—the foundation of all measurement. Hence, like the notion of the good and the beautiful, it appears to come from a different world, a world beyond that of the senses, a world that Plato then called the world of Ideas. Further intimations of such a realm beyond the immediate realm of the senses may be found in the fact that men, in construing a system of knowledge, constantly prefer what is more perfect to what is less perfect; *i.e.,* what is formed and thus recognizable to what is not, what is true to what is false, a sound logical conclusion to a logical fallacy, even an elegant scientific demonstration to a clumsy one, without considering the former as good and the latter as bad.

According to Plato, all of the things that men perceive with their senses appear to be but very imperfect copies of the eternal Ideas. The most important and fundamental one of these is the Idea of the Good. It is "beyond being and knowledge," yet it is the foundation of both. "Being" in this connection does not mean existence, but being something specific—a man, a lion, or a house—being recognizable by its quality or shape.

Knowledge begins with a perception of these earthly shapes, but it ascends from there to the higher realm of Ideas, which is approachable to the human mind. In the famous myth of the cave in the seventh book of *Politeia,* Plato likened the ordinary person to a man sitting in a cave looking at a wall on which he sees nothing but the shadows of the real things that are behind his back, and he likened the philosopher to a man who has got out in the open and seen the real world of the Ideas. Coming back, he may be less able to distinguish the shades because he has been blinded by the light outside; but he is the only one who knows reality, and he conducts his life accordingly.

Myth of the cave

In his later thinking, in *Theaetetus,* Plato criticized a sensualist theory of knowledge, anticipating the explanations of the 17th-century English sensualists, such as Thomas Hobbes. In the *Timaeus,* he tried to build up a complete system of physics, partly employing Pythagorean ideas. Most modern Positivists do not take Plato seriously any longer. But one of the greatest physicists of the 20th century, Werner Heisenberg, has insisted that the modern physicist still has to learn a good deal from Plato concerning the foundation of his science.

Aristotle. After Plato's death the Academy continued to exist for many centuries under various heads. When Plato's nephew, Speusippus, was elected as his successor, his greatest disciple, Aristotle, left to go first to Assus and then to the island of Lesbos, where he met Theophrastus, who became his most gifted disciple. But soon thereafter he was called to the Macedonian court at Pella to become the educator of the crown prince, who was later to become Alexander the Great. After the latter had become king, Aristotle returned to Athens and opened there a school of his own, whose members became known as the Peripatetics.

Philosophy. Aristotle had become a member of the Academy at the age of 17, in the year 367 (during Plato's absence in Sicily), under the acting chairmanship of Eudoxus of Cnidus, a great mathematician and geographer. It is a controversial question as to how far Aristotle, during the 20 years of his membership in the Academy, developed a philosophy of his own differing from that of his master. But two things can be considered as certain: (1) that he soon raised certain objections to Plato's theory of Ideas, for one of the arguments against it attributed to him is discussed in Plato's dialogue *Parmenides,* which Plato must have written soon after his return from Sicily, and (2) that it was during his membership in the Academy that Aristotle began and (to a considerable extent) elaborated his theoretical and formal analysis of the arguments used in various Socratic discussions—an enterprise that (when completed) resulted in the corpus of his works on logic,

a new science, which Aristotle himself claimed to have originated and about which (until rather recent times) it used to be said that he completed it in such a way that hardly anything could be added.

Certainly quite some time before his return to Athens to open a school of his own, Aristotle declared that it is not necessary to assume the existence of a separate realm of transcendent Ideas of which the individual things that men perceive with their senses are but imperfect copies; that the world of perceived things is the real world; and that it is necessary merely to be able to say that something is generally true of certain types or groups of things in order to build up a system of knowledge about them. Thus, it would be wrong to say that, having abandoned the theory of Ideas, Aristotle was left with a completely contingent world. The last chapters of his *Analytica posteriora* (*Posterior Analytics*) show, on the contrary, that he merely replaced Plato's transcendent Ideas with something (*katholou*) corresponding to them that the human mind can grasp in individual things.

Teleology and perfection

Aristotle retained another important element of the theory of Ideas in his teleology, or doctrine of purposiveness. According to Plato, individual things are imperfect copies of perfect Ideas. Aristotle pointed out, however, that all living beings develop from an imperfect state (from the seed, the semen, through the germinating plant, or embryo, to the child and young adult), to the more perfect state of the fully developed plant or the full grown mature animal or man—after which they again decay and finally die, having reproduced themselves. But not all individuals reach the same degree of relative perfection. Many of them die before reaching it; others are retarded or crippled or maimed in various ways in the process. It is, therefore, of the utmost importance for man to find out what the best conditions are for reaching the most perfect state possible. This is what the gardener tries to do for the plants; but it is even more important for man to do it in regard to himself. The first question, then, is what kind of perfection a human being as human can reach. In answering this question Aristotle observed that man, being the social animal par excellence, can reach as an individual only some of the perfections possible for man as such. Cats are more or less all alike in their functions; thus each can fend for itself. With bees and termites, however, it is different. They are by nature divided into worker bees, drones, and queen bees or worker termites, soldier termites, and queens. With human beings the differentiation of functions is much more subtle and varied. Men can lead satisfactory lives only on the basis of a division of labour and distribution of functions. Some human individuals are born with very great talents and inclinations for special kinds of activity. They will be happy and will make their best possible contribution to the life of the community only if they are permitted to follow this inclination. Others are less one-sidedly gifted and more easily adaptable to a variety of functions. These people can be happy shifting from one activity to another. That this is so is an enormous advantage the human species has over all other animals because it enables it to adapt to all sorts of circumstances. But the advantage is paid for by the fact that no human individual is able to develop all of the perfections that are possible for the race as a whole.

There is another possible and, in its consequences, real disadvantage to such adaptability; the other animals, tightly confined to the limits set by nature, are crippled almost exclusively by external factors; but man, in consequence of the freedom of choice granted to him through the variety of his gifts, can and very often does cripple and harm himself. All human activities are directed toward the end of a good and satisfactory life. But there are many subordinate aims that are sensible ends only as far as they serve a superior end. There is, for example, no sense in producing or acquiring more shoes than can possibly be worn. This is self-evident. With regard to money, however, which has become exchangeable against everything, the illusion arises that it is good to accumulate it without limit. By doing so, man harms both the community and himself because, concentrating on such a narrow aim, he deprives his soul and spirit of larger and more rewarding experiences. Similarly, an individual especially gifted for large-scale planning needs power to give orders to those capable of executing his plans. Used for such purposes, power is good. But coveted for its own sake, it becomes oppressive to those subdued by it and harmful to the oppressor because he thus incurs the hatred of the oppressed. Because of his imperfection man is not able to engage in serious and fruitful activities without interruption. He needs relaxation and play, or amusement. Because the necessities of life frequently force man to work beyond the limit within which working is pleasant, the illusion arises that a life of constant amusement would be the most pleasant and joyful. In reality nothing is more tedious.

Aristotle's teleology seems to be based entirely on empirical observation. It has nothing to do with a belief in divine providence and is not, as some modern critics believe, at variance with the law of causality. It forms the foundation, however, of Aristotle's ethics and political theory. Aristotle was an avid collector of empirical evidence. He induced his students, for instance, to make collections of the laws and political institutions (and their historical developments) of all known cities and nations in order to find out how they worked and at what points their initiators had been mistaken regarding the way in which they would work. In later times, Aristotle came to be considered (and by many is still considered) a dogmatic philosopher because the results of his inquiries were accepted as absolutely authoritative. In reality, however, he was one of the greatest Empiricists of all times.

Disciples and commentators. After Aristotle's death his immediate disciples carried on the same kind of work, especially in the historical field: Theophrastus wrote a history of philosophy and works on botany and on mineralogy, Eudemus of Rhodes wrote histories of mathematics and of astronomy, Meno a history of medicine, and Dicaearchus of Messene a history of civilization and a book on types of political constitutions. The next two generations of Peripatetics spread out in two different directions: literary history, in the form of histories of types of poetry, epic, tragedy, and comedy, and of biographies of famous writers, and physical science, Straton of Lampsacus creating a new kind of physics based on experiments, and the great astronomer Aristarchus of Samos inventing the heliocentric system. The school then went for some time into eclipse until, in the 1st century AD, after the rediscovery of Aristotle's lecture manuscripts, there arose a great school of commentators on his works, which had an enormous influence on medieval philosophy.

HELLENISTIC AND ROMAN PHILOSOPHY

The period after the death of Aristotle was characterized by the decay of the Greek city-states, which then became pawns in the power game of the Hellenistic kings who succeeded Alexander. Life became troubled and insecure. It was in this environment that two dogmatic philosophical systems came into being, the Stoic and the Epicurean, which were destined to give their adherents something to hold onto and to make them independent of the external world.

Zeno of Citium

Stoics. The Stoic system was created by a Syrian, Zeno of Citium (about the turn of the 3rd century BC), who went to Athens as a merchant but lost his fortune at sea. Zeno was consoled by the Cynic philosopher Crates, who taught him that material possessions were of no importance whatsoever for a man's happiness. He therefore stayed at Athens, heard the lectures of various philosophers, and—after he had elaborated his own philosophy—began to teach in a public hall, the Stoa Poikile (hence the name Stoicism).

Zeno's thought comprised, essentially, a dogmatized Socratic philosophy, with added ingredients derived from Heraclejtus. The basis of human happiness, he said, is to live "in agreement" (with oneself), a statement that was later replaced by the formula "to live in agreement with nature." The only real good for man is the possession of virtue; everything else (wealth or poverty, health or illness, life or death) is completely indifferent. All virtues are based exclusively on right knowledge—self-control (*sōphrosynē*) being the knowledge of the right

choice, fortitude the knowledge of what must be endured and what must not, and justice the right knowledge "in distribution." The passions, which are the cause of all evil, are the result of error in judging what is a real good and what is not. Because it is difficult to see, however, why murder, fraud, and theft should be considered evil if life and possessions are of no value, the doctrine was later modified by making among the "indifferent things" distinctions between "preferable things," such as having the necessities of life and health; "completely indifferent things"; and "anti-preferable things," such as lacking the necessities of life or being ill—while insisting still that the happiness of the truly wise man could not be impaired by illness, pain, hunger, or any deprivation of external goods. In the beginning, Zeno also insisted that either a man is completely wise, in which case he would never do anything wrong and would be completely happy, or he is a fool. Later he made the concession, however, that there are men not completely wise but progressing toward wisdom. Though the latter might even have true insight, they are not certain that they have it, whereas the truly wise man is also certain of having true insight. The world is governed by divine Logos—a word originally meaning "word" or "speech," then (with Heracleitus) also a speech that expresses the laws of the universe, and, finally, "reason." This Logos keeps the world in perfect order. Man can deviate from or rebel against this order, but by doing so he cannot disturb it but can only do harm to himself.

Development of Zeno's philosophy

Zeno's philosophy was further developed by Cleanthes, the second head of the school, and by Chrysippus, its third head. Chrysippus elaborated a new kind of logic, which did not receive much attention, however, outside the Stoic school until in recent times (under the name of "propositional logic") it has been hailed by some logicians as superior to the "conceptual logic" of Aristotle. In the mid-2nd century BC, Panaetius of Rhodes adapted Stoic philosophy to the needs of the Roman aristocracy (whose members were then governing the known world) and made a great impression on some of the leading men of the time, who tried to follow his moral precepts. In the following century, in the time of the decay of the Roman Republic, of civil war, and of slave rebellions, Poseidonius of Apamea, who was also one of the most brilliant historians of all times, taught that the Stoic takes a position above the rest of mankind, looking down on men's struggles as on a spectacle. In the periods of the rising monarchy and of its established rule, Stoicism became the religion of the republican opposition. The most famous Stoic was the younger Cato, who committed suicide after the victory of Julius Caesar. It was also the guiding philosophy of Seneca the Younger, the educator and (for a long time) the adviser of Nero, who tried to keep Nero on the path of virtue but failed and finally had to commit suicide on the orders of the Emperor. In spite of the oddities of Zeno's original doctrine, Stoicism gave consolation, composure, and fortitude in times of trouble to many proud men to the end of antiquity and beyond.

Epicurus

Epicureans. The thought of Zeno's contemporary Epicurus also comprised a philosophy of defense in a troubled world. It has been (and still is) considered—in many respects justly—the opposite of Zeno's. Whereas Zeno had proclaimed that the wise man would try to learn from everybody and would always acknowledge his debt to earlier philosophers, Epicurus insisted that everything he taught was the result of his own thinking, though it is obvious that his physical explanation of the universe is a simplification of Democritus' Atomism. And whereas the Stoics had taught that pleasure and pain are of no importance for a man's happiness, Epicurus made pleasure the very essence of a happy life. Moreover, the Stoics from the beginning had acted as advisers of kings and statesmen. Epicurus, on the other hand, lived in the retirement of his famous Garden, cultivating intimate friendships with his adherents but warning against participation in public life. The Stoics believed in divine providence; Epicurus taught that the gods pay no attention whatsoever to human beings. Yet in spite of these contrasts, the two philosophies had some essential factors in common. Though Epicurus made pleasure the criterion of a good life, he was far from advocating a dissolute life and debauchery; he insisted that it was the simple pleasures that made life happy. When in his old age he suffered terrible pains from prostatitis, he asserted that philosophizing and the memory and love of his distant friends made pleasure prevail even in the grips of such pain. Nor was Epicurus an atheist. His Roman admirer, the poet Lucretius Carus (*c.* 95–55 BC), in his poem *De rerum natura* (*On the Nature of Things*), praised Epicurus enthusiastically as the liberator of mankind from all religious fears; and Epicurus himself had affirmed that this had been one of the aims of his philosophy. But although he taught that the gods are much too superior to trouble themselves with paying attention to mortals, he said—and, as his language clearly shows, sincerely believed—that it is important for human beings to look at the gods as perfect beings, since only in this way could men approach perfection. It was only in Roman times that people began to misunderstand Epicureanism, holding it to be an atheistic philosophy justifying a dissolute life, so that a man could be called "a swine from the herd of Epicurus." Seneca recognized the true nature of Epicureanism, however, and in his *Epistulae morales* (*Moral Letters*) deliberately interspersed through his Stoic exhortations maxims from Epicurus.

Lucretius Carus

Skeptics. There was still another Hellenistic school of philosophy, the Skeptic school initiated by another of Zeno's contemporaries—Pyrrhon of Elis—a school that was destined to become of great importance for the preservation of a detailed knowledge of Hellenistic philosophy in general. Pyrrhon had come to the conviction that no man can know anything for certain nor ever be certain that the things he perceives with his senses are real and not illusory. He is said to have carried the practical consequences of his conviction so far that, when walking in the streets, he paid no attention to the vehicles and other obstacles, so that his faithful disciples always had to accompany him to see that he came to no harm. Pyrrhon's importance for the history of philosophy lies in the fact that one of the later adherents of his doctrine, Sextus Empiricus (2nd–3rd century AD), wrote a large work, *Pros dogmatikous* ("Against the Dogmatists"), in which he tried to refute all of the philosophers who had a more positive philosophy, and in so doing he quoted extensively from their works, thus preserving much that would otherwise have been lost. It is a noteworthy fact that the British sensualists of the 18th century, such as David Hume, and also Immanuel Kant derived most of their knowledge of ancient philosophy from Sextus.

Pyrrhon of Elis and Sextus Empiricus

Neo-Pythagoreans and Neoplatonists. All of the philosophical schools and sects of Athens that had originated in the 4th century BC continued into late antiquity, most of them until the emperor Justinian I ordered all of them closed in AD 529 because of their pagan character. Within that whole period of nearly 1,000 years only two new schools were added; even these, however—the Neo-Pythagorean and the Neoplatonic schools—drew their inspiration from early Greek philosophy, though only the latter was of importance for the history of philosophy. Neoplatonism began with Ammonius Saccas (first half of the 3rd century AD), who had been brought up as a Christian but had abandoned his religion for the study of Plato and developed his own kind of Platonic philosophy. Because he wrote nothing, his philosophy is known only through his famous disciple, Plotinus. But Plotinus did not publish anything either. His philosophy is known, however, through the *Enneads,* a collection of his writings arranged by his disciple Porphyry, who also wrote a biography of Plotinus.

Plotinus and Iamblichus

The philosophy of Plotinus (and Ammonius) was derived from the study of Plato. It, moreover, used many philosophical terms first coined by Aristotle and adopted some elements of Stoic philosophy as well. Yet it is essentially a new philosophy, agreeing with the religious and mystical tendencies of its time. Plotinus assumed the existence of several levels of Being, the highest being that of the One, or the Good, which are identical but indescribable and indefinable in human language. The next lower level is that of the nous, or pure intellect or reason; the third is that of the soul or souls. There then follows the world

perceivable by the senses and, finally, at the lowest level there is matter, which is the cause of all evil. The highest bliss for man is union with the One, or Good, attained by contemplation and purification. That this is not a lasting state attained once for all—like the status of the Stoic wise man, who was supposed never to lose his wisdom again—is shown by the fact that Porphyry, in his *Vita Plotini*, said that Plotinus had experienced this supreme bliss seven times in his life, whereas he, Porphyry, had experienced it only once.

The further history of Neoplatonism is extremely complicated. While Porphyry had emphasized the ethical element in Plotinus' philosophy, his disciple Iamblichus of Chalcis in Syria (died *c.* AD 330), founder of a Syrian branch of the sect, mingled Neoplatonism with Neo-Pythagoreanism, writing on the Pythagorean way of life and on number theory. Above all, he multiplied the levels of being, or the emanations from the One, which enabled him to incorporate the traditional Greek gods into his system. Another branch of the school was founded in Pergamum, in western Asia Minor, by his disciple Aedesius, who with his disciple Maximus tried to revive the ancient Greek mystery religions, such as Orphism. All of these developments became of great importance in the 4th century when the emperor Julian attempted to revive paganism. In the following century the Athenian school reached a new high point when Proclus combined ideas of his predecessors into one comprehensive system. When in 529 Justinian closed all of the philosophical schools in Athens, however, a branch continued to exist in Alexandria. The Athenian Neoplatonists found refuge at the court of the Persian king Khosrow, and in 535 they were permitted to return to Athens. But gradually pagan philosophy as such died out, though it continued to exist as an influence in the development of Christian philosophy and theology.

(K.v.Fr.)

Medieval philosophy

Medieval philosophy designates the philosophical speculation that occurred in western Europe during the Middle Ages; *i.e.*, from the fall of the Roman Empire in the 4th and 5th centuries AD to the Renaissance of the 15th century. Philosophy of the medieval period remained in close conjunction with Christian thought, particularly theology, and the chief philosophers of the period were churchmen, particularly churchmen who were teachers. Philosophers who strayed from the close relation were chided by their superiors. Greek philosophy ceased to be creative after Plotinus in the 3rd century AD. A century later Christian thinkers such as Ambrose, Victorinus, and Augustine began to assimilate Neoplatonism into Christian doctrine in order to give a rational interpretation of Christian faith. Thus, medieval philosophy was born of the confluence of Greek (and to a lesser extent of Roman) philosophy and Christianity. Plotinus' philosophy was already deeply religious, having come under the influence of Middle Eastern religion. Medieval philosophy continued to be characterized by this religious orientation. Its methods were at first those of Plotinus and later those of Aristotle. But it developed within faith as a means of throwing light on the truths and mysteries of faith. Thus, religion and philosophy fruitfully cooperated in the Middle Ages. Philosophy, as the handmaiden of theology, made possible a rational understanding of faith. Faith, for its part, inspired Christian thinkers to develop new philosophical ideas, some of which became part of the philosophical heritage of the West.

Toward the end of the Middle Ages, this beneficial interplay of faith and reason started to break down. Philosophy began to be cultivated for its own sake, apart from, and even in contradiction to, Christian religion. This divorce of reason from faith, made definitive in the 17th century by Francis Bacon and René Descartes, marked the birth of modern philosophy.

EARLY MEDIEVAL PHILOSOPHY

The early medieval period, which extended to the 12th century, saw the barbarian invasions of the Roman Empire, the collapse of its civilization, and the gradual building of a new, Christian culture in western Europe. Philosophy in these troubled and darkened times was cultivated by late Roman thinkers such as Augustine (354–430) and Boethius (*c.* 480–*c.* 525), then by monks such as Anselm (1033–1109). The monasteries became the main centres of learning and education and retained their preeminence until the founding of the cathedral schools and universities in the 11th and 12th centuries.

Augustine. During these centuries philosophy was heavily influenced by Neoplatonism; Stoicism and Aristotelianism played only a minor role. Augustine was awakened to the philosophical life by reading Cicero, but the Neoplatonists most decisively shaped his philosophical methods and ideas. To them he owed his conviction that beyond the world of the senses there is a spiritual, eternal realm of truth that is the object of the human mind and the goal of all man's striving. This truth he identified with the God of Christianity. Man encounters this divine world of truth and beauty not through his senses but by turning inward to his mind, and above his mind to the intelligible light, in which he sees the truth. The Augustinian demonstration of the existence of God coincides with the proof of the existence of necessary, immutable Truth. Augustine considered the truths of both mathematics and ethics to be necessary, immutable, and eternal. These truths cannot come from the world of contingent, changing, and temporal things, nor from the mind itself, which is also contingent, mutable, and temporal. They are due to the illuminating presence in man's mind of eternal and immutable Truth, or God. Any doubt that man knows the truth with certainty was dispelled for Augustine by the certitude that even if he is deceived in many cases, man cannot doubt that he exists and knows and loves.

Augustine conceived of man as a composite of two substances, body and soul, of which the soul is by far the superior. The body, nevertheless, is not to be excluded from human nature, and its eventual resurrection from the dead is assured by Christian faith. The soul's immortality is proved by its possession of eternal and unchangeable Truth.

Augustine's *Confessions* (*c.* 400) and *De Trinitate* (400–416; *On the Trinity*) abound with penetrating psychological analyses of knowledge, perception, memory, and love. His *De civitate Dei* (413–426; *The City of God*) presents the whole drama of human history as a progressive movement of humanity, redeemed by God, to its final repose in its Creator.

Boethius. One of the most important channels by which Greek philosophy was transmitted to the Middle Ages was Boethius. He began to translate into Latin all the philosophical works of the Greeks, but his imprisonment and death by order of Theodoric, king of the Ostrogoths, cut short this project. He translated only the logical writings of Porphyry (a 3rd-century-AD Neoplatonist) and Aristotle. These translations and his commentaries on them brought to the thinkers of the Middle Ages the rudiments of Aristotelian logic. They also raised important philosophical questions, such as the nature of universals (terms that can be applied to more than one particular thing). Are universals real or only mental concepts? If real, are they corporeal or incorporeal; if incorporeal, do they exist in the sensible world or apart from it? Medieval philosophers debated at length these and other problems relating to universals. In his logical works Boethius presents the Aristotelian doctrine of universals, that they are only mental abstractions. In his *De consolatione philosophiae* (*c.* 525; *Concerning the Consolation of Philosophy*), however, he adopts the Platonic notion that they are innate ideas and their origin is in the remembering of knowledge in a previous existence. This book was extremely popular and influential in the Middle Ages. It contains not only a Platonic view of knowledge and reality but also a lively treatment of providence, divine foreknowledge, chance, fate, and human happiness.

The question of the nature of universals

Greek Fathers of the Church and Erigena. Another stream from which Greek philosophy, especially Neoplatonic thought, flowed into the Middle Ages was the Greek Fathers of the Church, notably Origen (*c.* 185–*c.* 254),

Gregory of Nyssa (*c.* 335–*c.* 394), Nemesius (*c.* 400), Pseudo-Dionysius the Areopagite (*c.* 500), and Maximus the Confessor (*c.* 580–662). In the 9th century, John Scotus, called Erigena ("belonging to the people of Erin") because he was born in Ireland, a master at the Carolingian court of Charles the Bald, translated into Latin some of the writings of these Greek theologians, and his own major work, *De divisione naturae* (862–866; *On the Division of Nature*), is a vast synthesis of Christian thought organized along Neoplatonic lines. For him, God is the primal unity, unknowable and unnameable in himself, from which the multiplicity of creatures flows. He so far transcends his creatures that he is most appropriately called superreal and supergood. Creation is the process of division whereby the many derive from the One. The One descends into the manifold of creation and reveals himself in it. By the reverse process the multiplicity of creatures will return to their unitary source at the end of time, when everything will be absorbed in God.

Anselm. After the breakdown of the Carolingian Empire in the 10th century, intellectual speculation was at a low ebb in western Europe. In the next century some political stability was achieved by Otto I, who reestablished the empire, and Benedictine monasteries were revitalized by reformers such as Peter Damian. Like Tertullian, a Christian writer of the 2nd and 3rd centuries, Damian mistrusted secular learning and philosophy as harmful to the faith. Other monks showed a keen interest in dialectic and philosophy. Among the latter was Anselm, an Italian who became abbot of the French monastery of Bec and later archbishop of Canterbury.

Like Augustine, Anselm used both faith and reason in his search for truth. Faith comes first, in his view, but reason should follow, giving reasons for what men believe.

Anselm's proofs of the existence of God

Anselm's monks asked him to write a model meditation on God in which everything would be proved by reason and nothing on the authority of Scripture. He replied with his *Monologium* (1077; "Monologue"; Eng. trans., *Monologium*), the original title of which was "A Meditation on the Reasonableness of Faith." It contains three proofs of the existence of God, all of which are based on Neoplatonic thought. The first proof moves from the awareness of a multiplicity of good things to the recognition that they all share or participate more or less in one and the same Good, which is supremely Good in itself, and this is God. The second and third proofs are similar: beginning with their awareness of a multiplicity of beings that have more or less of being, and more or less of perfection, men recognize that they share in One who is supremely Being and perfect.

Anselm's later work, the *Proslogium* (1077/78; "Allocution" or "Address"; Eng. trans., *Proslogium*), also entitled "Faith Seeking Understanding," contains his most famous proof of the existence of God. This begins with a datum of faith: men believe God to be the being than which none greater can be thought. Some, like the fool in the Psalms, say there is no God; but even the fool, on hearing these words, understands them, and what he understands exists in his intellect, even though he does not grant that such a being exists in reality. But it is greater to exist in reality and in the understanding than to exist in the understanding alone. Therefore it is contradictory to hold that God exists only in the intellect, for then the being than which none greater can be thought is one than which a greater can be thought, namely, one that exists both in reality and in the understanding. Philosophers still debate the meaning and value of this so-called ontological argument for God's existence.

Bernard of Clairvaux and Abelard. Anselm's inquiry into the existence and nature of God, as also his discussion of truth, love, and human liberty, aimed at fostering monastic contemplation. Other monks, such as the Cistercian Bernard of Clairvaux (1090–1153), were suspicious of the use of secular learning and philosophy in matters of faith. Bernard complained of the excessive indulgence in dialectic displayed by contemporaries such as Peter Abelard (1079–1142) and Gilbert de La Porrée. He himself developed a doctrine of mystical love, the influence of which lasted through the centuries. The monks of the Parisian Abbey of Saint-Victor were no less intent on fostering mystical contemplation, but they cultivated the liberal arts and philosophy as an aid to it. In this spirit, Hugh of Saint-Victor wrote his *Didascalicon* (*c.* 1127; "Teaching"; Eng. trans., *Didascalicon*), a monumental treatise on the theoretical and practical sciences and the trivium (grammar, rhetoric, dialectic) and the quadrivium (arithmetic, music, geometry, astronomy). During the same period the School of Chartres, attached to the famous cathedral near Paris, was the focus of Christian Neoplatonism and humanism.

Urban development in the 12th century shifted the centre of learning and education from the monasteries to the towns. Abelard founded several urban schools near Paris and taught in them. A passionate logician, he pioneered a method in theology that contributed to the later Scholastic method. His *Sic et non* (1115–17; *Yes and No*) cites the best authorities on both sides of theological questions in order to reach their correct solution. In philosophy his main interest was logic. On the question of universals he agreed with neither the Nominalists nor the Realists of his day. His Nominalist teacher Roscelin held that universals, such as "man" and "animal," are nothing but words, or names (*flatus vocis*). Abelard argued that this does not take into account the fact that names have meaning. His Realist teacher William of Champeaux taught that universals are realities apart from the mind. For Abelard, only individuals are real; universals are indeed names or mental concepts, but they have meaning because they refer to individuals. They do not signify an essence common to individuals, as the Realists maintained (*e.g.*, the essence "humanity" shared by all men), but signify instead the individuals in their common condition, or status, of being in a certain species, which results from God's creating them according to the same divine idea.

TRANSITION TO SCHOLASTICISM

In the 12th century a cultural revolution took place that influenced the whole subsequent history of Western philosophy. The old style of education, based on the liberal arts and emphasizing grammar and the reading of the Latin classics, was replaced by new methods stressing logic, dialectic, and all the scientific disciplines known at the time. John of Salisbury, of the School of Chartres, witnessed this radical change:

> Behold, everything was being renovated: grammar was being made over, logic was being remodeled, rhetoric was being despised. Discarding the rules of their predecessors, [the masters] were teaching the quadrivium with new methods taken from the very depths of philosophy.

Growing interest in Aristotelianism

In philosophy itself, there was a decline in Platonism and a growing interest in Aristotelianism. This change was occasioned by the translation into Latin of the works of Aristotle in the late 12th and early 13th centuries. Until then, only a few of his minor logical treatises were known. Now his *Topica, Analytica priora,* and *Analytica posteriora* were rendered into Latin, giving the schoolmen (the teachers of Western Christian philosophy in the 13th and 14th centuries) access to the Aristotelian methods of disputation and science, which became their own techniques of discussion and inquiry. Many other philosophical and scientific works of Greek and Arabic origin were translated at this time, creating a "knowledge explosion" in western Europe.

Arabic thought. Among the translations from Arabic were some of the writings of Avicenna (980–1037). This Islāmic philosopher had an extraordinary impact on the medieval schoolmen. His interpretation of Aristotle's notion of metaphysics as the science of *ens qua ens* ("being as being"), his analysis of many metaphysical terms, such as "being," "essence," "existence," and his metaphysical proof of the existence of God were often quoted, with approval or disapproval, in Christian circles. Also influential were his psychology, logic, and natural philosophy. His *al-Qānūn fī aṭ-ṭibb* (*Canon of Medicine*) was an authority on the subject until modern times. The *Maqāṣid al-Falāsifah* (1094; "The Aims of the Philosophers") of the Arabic theologian al-Ghazālī, known in Latin as Algazel (died *c.* 1111), an exposition of Avicenna's philosophy written in

order to criticize it, was read as a complement of Avicenna's works. The anonymous *Liber de causis* ("Book of Causes") was also translated into Latin from Arabic. This work, excerpted from Proclus' *Stoicheiōsis theologikē* (*Elements of Theology*), was often ascribed to Aristotle, and it gave a Neoplatonic cast to his philosophy until its true origin was discovered by Thomas Aquinas.

The commentaries of the Arabic philosopher Averroës were translated along with Aristotle's works. As Aristotle was called "the Philosopher" by the medieval philosophers, Averroës was dubbed "the Commentator." These two taught the Scholastics philosophy as a purely rational discipline, divorced from revealed religion. The Christian schoolmen often attacked Averroës as the archenemy of Christianity for his rationalism and his doctrines of the eternity of the world and the unity of the intellect for all men; *i.e.*, the doctrine that intellect is a single, undifferentiated form with which men become reunited at death. This was anathema to the Christian schoolmen because it contravened the Christian doctrine of individual immortality.

Jewish thought. Of considerably less influence on the Scholastics was medieval Jewish thought. Ibn Gabirol, known to the Scholastics as Avicebron or Avencebrol, was thought to be an Arab or Christian, though in fact he was a Spanish Jew. His chief philosophical work, written in Arabic and preserved only in a Latin translation entitled *Fons vitae* (*c.* 1050; *Fountain of Life*), stresses the unity and simplicity of God. All creatures are composed of form and matter, either the gross corporeal matter of the sensible world or the spiritual matter of angels and human souls. Some of the schoolmen were attracted to the notion of spiritual matter and also to Ibn Gabirol's analysis of a plurality of forms in creatures, according to which every corporeal being receives a variety of forms by which it is given its place in the hierarchy of being—for example, a dog has the forms of a corporeal thing, a living thing, an animal, a dog.

Maimonides, or Moses ben Maimon, was known to Christians of the Middle Ages as Rabbi Moses. His *Dalālat al-hā'irīn* (*c.* 1190; *Guide of the Perplexed*) helped them to reconcile Greek philosophy with revealed religion. For Maimonides there can be no conflict between reason and faith because both come from God; an apparent contradiction is due to a misinterpretation of either the Bible or the philosophers. Thus, he showed that creation is reconcilable with philosophical principles and that the Aristotelian arguments for an eternal world are not conclusive because they ignore the omnipotence of God, who can create a world of either finite or infinite duration.

The universities and Scholasticism

While Western scholars were assimilating the new treasures of Greek, Islāmic, and Jewish thought, universities that became the centres of Scholasticism were being founded. Of these the most important were Paris and Oxford (formed 1150–70 and 1168, respectively). Scholasticism is the name given to the theological and philosophical teachings of the schoolmen in the universities. There was no one Scholastic doctrine; each of the Scholastics developed his own, which was often in disagreement with that of his fellow teachers. They had in common a respect for the great writers of old, such as the Fathers of the Church, Aristotle, Plato, Boethius, the Pseudo-Dionysius, and Avicenna. These they called "authorities." Their interpretation and evaluation of the authorities, however, frequently differed. They also shared a common style and method that developed out of the teaching practices in the universities. Teaching was done by lecture and disputation (a formal debate). A lecture consisted of the reading of a prescribed text followed by the teacher's commentary on it. Masters also held disputations in which the affirmative and negative sides of a question were thoroughly argued by students and teacher, before the latter resolved the problem.

THE AGE OF THE SCHOOLMEN

Robert Grosseteste and Roger Bacon. The newly translated Greek and Arabic treatises had an immediate effect on the University of Oxford. Its first chancellor, Robert Grosseteste (*c.* 1168–1253), commented on some of Aristotle's works and translated the *Ethica Nichomachea* (*Nicomachean Ethics*) from Greek to Latin. He was deeply interested in scientific method, which he described as both inductive and deductive. By the observation of individual events in nature, man advances to a general law, called a "universal experimental principle," that accounts for these events. Experimentation either verifies or falsifies a theory by testing its empirical consequences. For Grosseteste the study of nature is impossible without mathematics. He cultivated the science of optics (*perspectiva*), which measures the behaviour of light by mathematical means. His studies of the rainbow and comets employ both observation and mathematics. His treatise *De luce* (1215–20; *On Light*) embodies a metaphysics of light, presenting light as the basic form of all things and God as the primal uncreated light.

Grosseteste's pupil Roger Bacon (*c.* 1220–*c.* 1292) made the mathematical and experimental methods the key to natural science. The term experimental science was popularized in the West through his writings. For him, man acquires knowledge through reasoning and experience, but without the latter he can have no certitude. Man gains experience through the senses and also through an interior divine illumination that culminates in mystical experience. Bacon was critical of the methods of Parisian theologians such as Albertus Magnus and Thomas Aquinas. He strove to create a universal wisdom embracing all the sciences and organized by theology. He also proposed the formation of a single worldwide society, or "Christian republic," that would unite all men under the leadership of the pope.

William of Auvergne. At the University of Paris, William of Auvergne (*c.* 1180–1249) was one of the first to feel the impact of the philosophies of Aristotle and Avicenna. As a teacher, and then as bishop of Paris, he was concerned with the threat of pagan and Islāmic thought to the Christian faith. He opposed the Aristotelian doctrine of the eternity of the world as contrary to the Christian notion of creation. His critique of Avicenna centred around the latter's conception of God and creation. The God of Avicenna, who creates the universe eternally and necessarily, through the mediation of 10 Intelligences, was opposed by William of Auvergne with the Christian notion of a God who creates the world freely and directly. Creatures are radically contingent and dependent on God's creative will. Unlike God, they do not exist necessarily; indeed, their existence is distinct from their essence and accidental to it. God has no essence distinct from his existence; he is pure existence. In stressing the essential instability and nonetemity of the world, he attributed true existence and causality to God alone. William of Auvergne was a follower of Augustine, but, like others at the time, he was compelled to rethink the older Augustinian notions in terms of the newer Aristotelian and Avicennian philosophies.

Opposition to Aristotelianism

Bonaventure. The Franciscan friar Bonaventure (*c.* 1217–74) reacted similarly to the growing popularity of Aristotle and his Arabic commentators. He admired Aristotle as a natural scientist, but he preferred Plato and Plotinus, and, above all, Augustine, as metaphysicians. His main criticism of Aristotle and his followers was that they denied the existence of divine ideas. As a result, Aristotle was ignorant of exemplarism (that is, God's creation of the world according to ideas in his mind) and also of divine providence and government of the world. This involved Aristotle in a threefold blindness: he taught that the world is eternal, that all men share one agent intellect (the active principle of understanding in man), and that there are no rewards or punishments after death. Plato and Plotinus avoided these mistakes, but, because they lacked Christian faith, they could not see the whole truth. For Bonaventure, faith alone enables men to avoid error in these important matters.

Bonaventure did not confuse philosophy with theology. Philosophy is the knowledge of the things of nature and the soul innate in man or acquired by his own efforts, whereas theology is the knowledge of heavenly things based on faith and divine revelation. Bonaventure, however, rejected the practical separation of philosophy from theology. Philosophy needs the guidance of faith; far from being self-sufficient, it is but a stage toward the higher knowledge that culminates in the vision of God.

For Bonaventure, every creature to some degree bears the mark of its Creator. The soul has been made in the very image of God. Thus, the universe is like a book in which the triune God is revealed. His *Itinerarium mentis in Deum* (1259; *The Soul's Journey into God*) follows Augustine's path to God, from the external world to the interior world of the mind, and then above the mind from the temporal to the eternal. Throughout this journey, men are aided by a moral and intellectual divine illumination. The mind has been created with an innate idea of God, so that, as Anselm pointed out, man cannot think that God does not exist. In a terse reformulation of the Anselmian argument for God's existence, Bonaventure states that if God is God, he exists.

Albertus Magnus. The achievement of the Dominican friar Albertus Magnus (*c.* 1200–80) was of vital importance for the development of medieval philosophy. A man of immense erudition and intellectual curiosity, he was one of the first to recognize the true value of the newly translated Greco-Arabic scientific and philosophical literature. Everything he considered valuable in it, he included in his encyclopaedic writings. He set out to teach this literature to his contemporaries and in particular to make the philosophy of Aristotle, whom he considered to be the greatest philosopher, understandable to them. He also proposed to write original works in order to complete what was lacking in the Aristotelian system. In no small measure, the triumph of Aristotelianism in the 13th century can be attributed to him.

Triumph of Aristotelianism

Albertus' observations and discoveries in the natural sciences advanced botany, zoology, and mineralogy. In philosophy he was less original and creative than his famous pupil Thomas Aquinas. Albertus produced a synthesis of Aristotelianism and Neoplatonism, blending together the philosophies of Aristotle, Avicenna, and Ibn Gabirol, and, among Christians, Augustine and Pseudo-Dionysius.

Thomas Aquinas. Albertus Magnus' Dominican confrere and pupil Thomas Aquinas (1224/25–1274) shared his master's great esteem for the ancient philosophers, especially Aristotle, and also for the more recent Arabic and Jewish thinkers. He welcomed truth wherever he found it and used it for the enrichment of Christian thought. For him reason and faith cannot contradict each other because they come from the same divine source. In his day conservative theologians and philosophers regarded Aristotle with suspicion and leaned toward the more traditional Christian Neoplatonism. Thomas realized that their suspicion was due, in part, to the fact that Aristotle's philosophy had been distorted by his Arabic commentators; so he wrote his own commentaries on Aristotle to show the essential soundness of his system and to convince contemporaries of its value for Christian theology.

Thomas' own philosophical views are best expressed in his theological works, especially his *Summa theologiae* (1265/66–1273; Eng. trans., *Summa theologiae*) and *Summa contra gentiles* (1258–64; *Summa Against the Gentiles*). In these works he clearly distinguishes between the domains and methods of philosophy and theology. The philosopher seeks the first causes of things, beginning with data furnished by the senses; the subject of the theologian's inquiry is God as revealed in sacred Scripture. In theology, appeal to authority carries most weight; in philosophy, it carries least.

Thomas found Aristotelianism and, to a lesser extent, Platonism useful instruments for Christian thought and communication; but he transformed and deepened everything he borrowed from them. For example, he took over Aristotle's proof of the existence of a primary unmoved mover, but the primary mover at which Thomas arrives is very different from that of Aristotle; it is in fact the God of Judaism and Christianity. He also adopted Aristotle's teaching that the soul is man's form and the body is his matter, but for Aquinas this does not entail, as it does for the Aristotelians, the denial of the immortality of the soul or the ultimate value of the individual. Thomas never compromised Christian doctrine by bringing it into line with the current Aristotelianism; rather, he modified and corrected the latter whenever it clashed with Christian belief. The harmony he established between Aristotelianism and Christianity was not forced but achieved by a new understanding of philosophical principles, especially the notion of being, which he conceived as the act of existing (esse). For him, God is pure being, or the act of existing. Creatures participate in being according to their essence; for example, man participates in being, or the act of existing, to the extent that his humanity, or essence, permits. The fundamental distinction between God and creatures is that creatures have a real composition of essence and existence, whereas God's essence is his existence.

Averroists. A group of masters in the Faculty of Arts at Paris welcomed Aristotle's philosophy and taught it in disregard of its possible opposition to the Christian faith. They wanted to be philosophers, not theologians, and to them this meant following the Aristotelian system. Because Averroës was the recognized commentator on Aristotle, they generally interpreted his thought in an Averroistic way. Hence, in their own day they were known as "Averroists"; today they are often called "Latin Averroists" because they taught in Latin. Their leader, Siger de Brabant, taught as rationally demonstrated certain Aristotelian doctrines that contradicted the faith, such as the eternity of the world and the oneness of the intellect for all men. They were accused of holding a "double truth"—of maintaining the existence of two contradictory truths: one commanded by faith, the other taught by reason. Although Siger never proposed as true philosophical conclusions contrary to faith, other members of this group upheld the right and duty of the philosopher to follow human reason to its natural conclusions, even when they contradicted the truths of faith.

This growing rationalism confirmed the belief of theologians of a traditionalist cast that the pagan and Muslim philosophies would destroy the Christian faith. They attacked these philosophies in treatises such as Giles of Rome's *Errores philosophorum* (1270; *The Errors of the Philosophers*). In 1277 the Bishop of Paris condemned 219 propositions based on the new trend toward rationalism and naturalism. These included even some of Thomas' Aristotelian doctrines. The same year, the Archbishop of Canterbury made a similar condemnation at Oxford. These reactions to the novel trends in philosophy did not prevent the Averroists from treating philosophical questions apart from religious considerations. Theologians, on their part, were increasingly suspicious of the philosophers and less optimistic about the ultimate reconciliation of philosophy and theology.

Condemnation of Aristotelian doctrines

PHILOSOPHY IN THE LATE MIDDLE AGES

In the late Middle Ages earlier ways of philosophizing were continued and formalized into definite schools of thought. In the Dominican order, Thomism (theology and philosophy of Thomas Aquinas) was made the official teaching, though the Dominicans did not always adhere to it rigorously. Averroism, cultivated by philosophers such as John of Jandun (died *c.* 1328), remained a live, though sterile, movement into the Renaissance. In the Franciscan order, the Englishmen John Duns Scotus and William of Ockham developed new styles of theology and philosophy that vied with Thomism throughout the late Middle Ages.

Duns Scotus. John Duns Scotus (*c.* 1265–1308) opposed the rationalists' contention that philosophy is self-sufficient and adequate to satisfy man's desire for knowledge. In fact, he claimed, a pure philosopher, such as Aristotle, could not truly understand the human condition because he was ignorant of the Fall of man and his need for grace and redemption. Unenlightened by Christian revelation, Aristotle mistook man's present fallen state, in which all his knowledge comes through the senses, for his natural condition, in which the object of his knowledge would be coextensive with all being, including the being of God. The limitation of Aristotle's philosophy is apparent to Duns Scotus in the Aristotelian proof of the existence of God as the primary mover of the universe. More adequate than this physical proof, he contended, is his own very intricate metaphysical demonstration of the existence of God as the absolutely primary, unique, and infinite being. He incorporated the Anselmian argument into this demonstration. For Duns Scotus, the notion of

infinite being, not that of primary mover or being itself, is man's most perfect concept of God.

In opposition to the Greco-Arabic view of the government of the universe from above by necessary causes, Duns Scotus stressed the contingency of the universe and its total dependence on God's infinite creative will. He adopted the traditional Franciscan voluntarism, elevating the will above the intellect in man.

Defense of the doctrine of God's will

Duns Scotus' doctrine of universals justly earned him the title "Doctor Subtilis." Universals, in his view, exist only as abstract concepts, but they are based on common natures, such as humanity, which exist, or can exist, in many individuals. Common natures are real, and they have a real unity of their own distinct from the unity of the individuals in which they exist. The individuality of each individual is due to an added positive reality that makes the common nature to be this individual; for example, humanity to be Socrates. Duns Scotus calls such a reality an "individual difference," or "thisness" (*haecceitas*). It is an original development of the earlier medieval realism of universals.

William of Ockham. In the late 14th century, Thomism and Scotism were called the "old way" (*via antiqua*) of philosophizing in contrast to the "modern way" (*via moderna*) begun by such men as William of Ockham (*c.* 1285–1347). Ockham, no less than Duns Scotus, wanted to defend the Christian doctrine of the freedom and omnipotence of God and the contingency of creatures against the necessitarianism of Greco-Arabic philosophy. But for him the freedom of God is incompatible with the existence of divine ideas as positive models of creation. God does not use preconceived ideas when he creates, as Duns Scotus maintained, but he fashions the universe as he wishes. As a result, creatures have no natures or essences in common. There are no realities but individual things, and these have nothing in common. They are more or less like each other, however, and on this basis men can form universal concepts of them and talk about them in general terms.

The absolute freedom of God was often used by Ockham as a principle of philosophical and theological explanation. Because the order of nature has been freely created by God, it could have been different: fire, for example, could cool, as it now heats. If he wishes, he can give us the sight, or "intuitive knowledge," of a star without the reality of the star. The moral order could also have been different. God could have made hating him meritorious instead of loving him. It was typical of Ockham not to put too much trust in the power of human reason to reach the truth. For him, philosophy must often be content with probable arguments, for example, in establishing the existence of the Christian God. Faith alone gives certitude in this and in other vital matters. Another principle invoked by Ockham is that a plurality is not to be posited without necessity. This principle of the economy of thought, later stated as beings are not to be multiplied without necessity, is called "Ockham's razor."

Ockhamism was censured by a papal commission at Avignon, Fr., in 1326, and in 1474 it was forbidden to be taught at Paris: it spread widely in the late Middle Ages, nevertheless, and rivaled Thomism and Scotism in popularity. Other Scholastics in the 14th century shared Ockham's basic principles and contributed with him to skepticism and probabilism in philosophy. John of Mirecourt (*c.* 1345) stressed the absolute power of God and the divine will to the point of making him the cause of man's sin. Nicholas of Autrecourt (*c.* 1347) adopted a skeptical attitude regarding such matters as man's ability to prove the existence of God and the reality of substance and causality. Rejecting Aristotelianism as inimical to the Christian faith, he advocated a return to the Atomism of the ancient Greeks as a more adequate explanation of the universe.

Trend away from Aristotelianism

Meister Eckehart. The trend away from Aristotelianism was accentuated by the German Dominican Meister Eckehart (*c.* 1260–1327/28), who developed a speculative mysticism of both Christian and Neoplatonic inspiration. Eckehart depicts the ascent of the soul to God in Neoplatonic terms: by gradually purifying itself from the body, the soul transcends being and knowledge until it is absorbed in the One. The soul is then united with God at its highest point, or "citadel." God himself transcends being and knowledge. Sometimes Eckehart describes God as the being of all things. This language, which was also used by Erigena and other Christian Neoplatonists, leaves him open to the charge of pantheism (the doctrine that the being of creatures is identical with that of God); but for Eckehart there is an infinite gulf between creatures and God. Eckehart means that creatures have no existence of their own but are given existence by God, as the body is made to exist and is contained by the soul. Eckehart's profound influence can be seen in the flowering of mysticism in the German Rhineland in the late Middle Ages.

Nicholas of Cusa. Nicholas of Cusa (1401–64) also preferred the Neoplatonists to the Aristotelians. To him, the philosophy of Aristotle is an obstacle to the mind in its ascent to God because its primary rule is the principle of contradiction, which denies the compatibility of contradictories. But God is the "coincidence of opposites." Because he is infinite, he embraces all things in perfect unity; he is at once the maximum and the minimum. Nicholas uses mathematical symbols to illustrate how, in infinity, contradictories coincide. If a circle is enlarged, the curve of its circumference becomes less; if a circle is infinite, its circumference is a straight line. As for man's knowledge of the infinite God, he must be content with conjecture or approximation to the truth. The absolute truth escapes man; his proper attitude is "learned ignorance."

For Nicholas, God alone is absolutely infinite. The universe reflects this divine perfection and is relatively infinite. It has no circumference, for it is limited by nothing outside of itself. Neither has it a centre; the Earth is neither at the centre of the universe nor is it completely at rest. Place and motion are not absolute but relative to the observer. This new, non-Aristotelian conception of the universe anticipated some of the features of modern theories.

Thus, at the end of the Middle Ages, some of the most creative minds were abandoning Aristotelianism and turning to newer ways of thought. The philosophy of Aristotle, in its various interpretations, continued to be taught in the universities, but it had lost its vitality and creativity. Christian philosophers were once again finding inspiration in Neoplatonism. The Platonism of the Renaissance was in direct continuity with the Platonism of the Middle Ages.

(A.A.Ma.)

Modern philosophy

THE RENAISSANCE AND EARLY MODERN PERIOD

The philosophy of a period arises as a response to social need, and the development of philosophy in the history of Western civilization since the Renaissance has, thus, reflected the process in which creative philosophers have responded to the unique challenge of each stage in the development of Western culture itself.

The career of philosophy—how it views its tasks and functions, how it defines itself, the special methods it invents for the achievement of philosophical knowledge, the literary forms it adopts and utilizes, its conception of the scope of its subject matter, and its changing criteria of meaning and truth—hinges on the mode of its successive responses to the challenges of the social structure within which it arises. Thus, Western philosophy in the Middle Ages was primarily a Christian philosophy, complementing the divine revelation, reflecting the feudal order in its cosmology, devoting itself in no small measure to the institutional tasks of the Roman Catholic Church. It was no accident that the major philosophical achievements of the 13th and 14th centuries were the work of churchmen who also happened to be professors of theology at the universities of Oxford and Paris.

Cultural impact of new inventions and nations

The Renaissance of the late 15th and 16th centuries presented a different set of problems and therefore suggested different lines of philosophical endeavour. What is called the European Renaissance followed upon the introduction of three novel mechanical inventions from the East: gunpowder, block printing from movable type, and the compass. The first was used to explode the massive fortifications of the feudal order and thus became an

agent of the new spirit of nationalism that threatened the rule of churchmen—and, indeed, the universalist emphasis of the church itself—with a competing secular power. The second, printing, made the propagation of knowledge widespread, secularized learning, reduced the intellectual monopoly of an ecclesiastical elite, and restored the literary and philosophical classics of Greece and Rome. The third, the compass, increased the safety and scope of navigation, produced the voyages of discovery that opened up the Western Hemisphere, and symbolized a new spirit of physical adventure and a new scientific interest in the structure of the natural world.

Each of these inventions with its wider cultural consequences presented new intellectual problems and novel philosophical tasks within a changed political and social environment. For, as the power of a single religious authority was slowly eroded under the influence of the Protestant Reformation and as the prestige of the universal Latin language gave way to vernacular tongues, philosophers became less and less identified with their positions in the ecclesiastical hierarchy and more and more identified with their national origins. The works of Albertus Magnus, St. Thomas Aquinas, St. Bonaventure, and John Duns Scotus had been basically unrelated to the countries of their birth; but the philosophy of Niccolò Machiavelli was directly related to Italian experience, that of Sir Francis Bacon and Thomas Hobbes was English to the core, and that of René Descartes set the standard and tone of French intellectual life for 200 years.

Dominant strands of Renaissance philosophy. Knowledge in the contemporary world is conventionally divided between the natural sciences, the social sciences, and the humanities. In the Renaissance, however, fields of learning had not yet become so sharply departmentalized: in fact, each of these divisions arose in the comprehensive and broadly inclusive area of Renaissance philosophy. For, as the Renaissance mounted its revolt against the reign of religion and therefore reacted against the church, against authority, against Scholasticism, and against Aristotle, there was a sudden blossoming of interest in problems centring on civil society, man, and nature. These three interests found exact representation in the three dominant strands of Renaissance philosophy: (1) political theory, (2) humanism, and (3) the philosophy of nature.

Political theory. As secular authority replaced ecclesiastical authority and as the dominant interest of the age shifted from religion to politics, it was natural that the rivalries of the national states and their persistent crises of internal order should raise with renewed urgency philosophical problems, practically dormant since pre-Christian times, about the nature and the moral status of political power. This new preoccupation with national unity, internal security, state power, and international justice stimulated the growth of political philosophy in Italy, France, England, and Holland.

Machiavelli and power

In early 16th-century Italy, Niccolò Machiavelli, sometime state secretary of the Florentine republic, explored in *Il principe* (written 1512–13; *The Prince*) and in his *Discorsi sopra la prima deca di Tito Livio* (completed by 1521; "Discourses on the First Ten Books of Livy" in *Discourses*) the techniques for the seizure and retention of power in ways that seemed to exalt "reasons of state" above morality and codified the actual practices of Renaissance diplomacy for the next 100 years. In fact, Machiavelli was motivated by patriotic hopes for the ultimate unification of Italy and by the conviction that the low estate of Italian Renaissance morality needed to be elevated by restoring the ancient Roman virtues. More than half a century later in France, Jean Bodin, magistrate of Laon and a member of the Estates-General, insisted that the state must possess a single, unified, and absolute power; he thus developed in detail the doctrine of national sovereignty in all of its administrative consequences and in its role as the source of all legal legitimacy.

In the 17th century in England, Thomas Hobbes, who was to become tutor to the future Charles II, developed the fiction that in the "state of nature" that preceded civilization life was "nasty, brutish, and short" with "every man's hand raised against every other," and that a "social contract" was thus agreed upon to convey all private rights to a single sovereign in return for general protection and for the institution of a reign of law. Because law is simply "the command of the sovereign," Hobbes at once turned justice into a by-product of power and denied any right of rebellion except when the sovereign becomes too weak to protect the commonwealth and hold it united.

In Holland, a prosperous and tolerant commercial republic in the 17th century, the issues of political philosophy took a different form. Thus, when the Dutch East India Company commissioned a great jurist, Hugo Grotius (1583–1645), to provide a defense of their trade rights and of their free access to the seas, the resulting two treatises, *Mare Liberum* (1609; *The Freedom of the Seas*) and *De Jure Belli ac Pacis* (1625; *On the Law of War and Peace*), were the first significant codifications of international law. Their philosophical originality lay, however, in the fact that, in defending the rights of a small, militarily weak nation against the powerful absolutisms of England, France, and Spain, Grotius was led to a preliminary investigation of the sources and validity of the concept of "natural law"—the notion that inherent in human reason and immutable even against the willfulness of sovereign states are imperative considerations of natural justice and moral responsibility, which must serve as a check against the arbitrary exercise of vast political power.

Ambivalence of political philosophers

In general, the political philosophy of the Renaissance was dualistic: it was haunted, even confused, by the conflict between political necessity and general moral responsibility. Machiavelli, Bodin, and Hobbes asserted claims that justified the actions of Italian despotism and the absolutisms of the Bourbon and Stuart dynasties. Yet Machiavelli was obsessed with the problem of human virtue; Bodin insisted that even the sovereign ought to obey the law of nature—that is, to govern in accordance with the dictates of natural justice; and Hobbes himself found in natural law the rational motivation that causes a man to seek for security and peace. In the end, though Renaissance political necessity required that the philosophical doctrines of Thrasymachus (who held that right is what is in the interest of the strong) be implemented, it could never finally escape a twinge of Socratic conscience.

Humanism. The Renaissance was characterized by the renewed study of mathematics, medicine, and classical literature. The first two sparked the scientific revolution of the 16th and 17th centuries; the last became the foundation of the philosophy of Renaissance humanism. From its origin, humanism—suspicious of science and generally indifferent to religion—emphasized anew the centrality of man in the universe, his supreme value and importance. Characteristic of this emphasis was the famous *Oratio de hominis dignitate* (written 1486; *Oration on the Dignity of Man*) of a late 15th-century Platonist, Pico della Mirandola, a leading member of Lorenzo de' Medici's Platonic Academy of Florence. But the new emphasis upon man's personal responsibility and on the possibility of his self-creation as a work of art was in no small part a consequence of the rediscovery of a series of crucial classical texts, which served to reverse the trends of medieval learning. Renaissance humanism was predicated upon the victory of rhetoric over dialectic and of Plato over Aristotle, as Quintilian and Cicero had triumphed over Abelard and as the cramped format of Scholastic philosophical method gave way to a Platonic discursiveness.

Much of this had been prepared by Italian scholarly initiative in the early 15th century. The recently discovered manuscript of Quintilian was used by Lorenzo Valla, an antiauthoritarian humanist, for the creation of modern rhetoric and the principles of textual criticism. But even more important was the rebirth of an enthusiasm for the philosophy of Plato in Medicean Florence and at the cultivated court of Urbino, a dukedom east of Florence. Precisely to service this enthusiasm, Marsilio Ficino, head of the Platonic Academy, had translated the entire Platonic corpus into Latin by the end of the century.

Enthusiasm for Plato

Except for Pico and Giordano Bruno, a late 16th-century Italian philosopher, the direct influence of Platonism upon Renaissance metaphysics is difficult to trace. The Platonic account of the moral virtues, however, was admirably

adapted to the requirements of Renaissance education and gave new support to the Renaissance ideal of the courtier and the gentleman. But Plato also represented the philosophical importance of mathematics and the Pythagorean attempt to discover the secrets of the heavens, the earth, and the world of nature in terms of number and exact calculation; and this aspect of Platonism spilled over from humanism into the domain of Renaissance science. The scientists Nicolaus Copernicus, Johannes Kepler, and Galileo owe more to the general climate of Pythagorean confidence in the explanatory power of number than does Renaissance metaphysics.

But Platonism also had the effect of influencing the literary form in which Renaissance philosophy was written. Although the very early medieval Platonists St. Augustine and John Scotus Erigena occasionally had used the dialogue form, later Scholasticism had abandoned it in favour of the formal treatise, of which the great "Summas" of Alexander of Hales and St. Thomas Aquinas are pristine examples. The Renaissance rediscovery of the Platonic dialogues suggested the literary charm of this conversational method to humanists, scientists, and political theorists alike. The humanist philosopher Bruno put forth his central insights in a dialogue, *De la causa, principio e uno* (1584; *Concerning the Cause, Principle, and One*); Galileo presented his novel mechanics in his *Dialogo sopra i due massimi sistemi del mondo, tolemaico e copernicano* (1632; *Dialogue Concerning the Two Chief World Systems—Ptolemaic and Copernican*); and even the politician Machiavelli wrote *Dell'arte della guerra* (1521; *The Art of War*) as a genteel conversation taking place in a quiet Florentine garden.

Renaissance humanism was primarily a moral and a literary, rather than a narrowly philosophical, movement. And it flowered in figures with broadly philosophical interests, such as Erasmus of Rotterdam, the erudite citizen of the world; Sir Thomas More, the learned but unfortunate chancellor of Henry VIII; and, in the next generation, in the great French essayist and mayor of Bordeaux, Michel de Montaigne. But the recovery of the Greek and Latin classics, which was the work of humanism, had profound effects upon the entire field of Renaissance philosophy and science through the ancient schools of philosophy to which it once more directed attention. In addition to Platonism, the most notable of these were Greek Atomistic Materialism, Greek Skepticism, and Roman Stoicism.

Revival of other ancient Greek schools

The discovery of the manuscript of Lucretius (and the Atomistic doctrines of Democritus) finally came to influence Galileo, Bruno, and, later, Pierre Gassendi, a modern Epicurean, through the insights into nature reflected in this work. The recovery of the manuscript of Sextus Empiricus, with its carefully argued Skepticism presented in a printed text in 1562, produced "a skeptical crisis" in French philosophy, which dominated the period from Montaigne to René Descartes. And the Stoicism of Seneca and Epictetus became almost the official ethics of the Renaissance—to appear prominently in the *Essais* (1580–88) of Montaigne, in the letters that Descartes wrote to the princess Elizabeth of Bohemia and to Queen Christina of Sweden, and in the later sections of the *Ethics* (first published 1675) of the Rationalist Benedict de Spinoza.

Philosophy of nature. Philosophy in the modern world is a self-conscious discipline. It has managed to define itself narrowly, so as to differentiate itself on the one hand from religion and on the other from exact science. But this narrowing of focus came about very late in its history—certainly not before the 18th century. The earliest philosophers of Greece were theorists of the physical world; Pythagoras and Plato were at once philosophers and mathematicians; and in Aristotle no clear distinction between philosophy and natural science can be maintained. The Renaissance continued this breadth of conception characteristic of the Greeks. Galileo and Descartes were mathematicians, physicists, and philosophers at once; and physics retained the name of "natural philosophy" at least until the death of Sir Isaac Newton in 1727.

Had the Renaissance been painstakingly self-aware in the matter of definition (which it was not), it might have defined philosophy, on the basis of its actual practice, as "the rational, methodical, and systematic consideration of man, civil society, and the natural world." The areas of its interests would in no case have been in doubt. But exactly what constitutes "rational, methodical, and systematic consideration" would have been extremely controversial. For knowledge advances through the discovery and advocacy of new philosophical methods; and, because the diverse methods advocated depend for their validity upon the acceptance of different philosophical criteria of truth, meaning, and importance, the crucial philosophical quarrels of the 16th and 17th centuries were at bottom quarrels in the advocacy of methods. It is this issue rather than any disagreement over subject matter or areas of attention that separated the greatest Renaissance philosophers—such as Francis Bacon, René Descartes, and Thomas Hobbes.

The great new fact that confronted the Renaissance was the immediacy, the immensity, and the uniformity of the natural world. But what was of primary importance was the new perspective in which this fact was interpreted. To the Middle Ages the universe was hierarchical, organic, and God-ordained. To the Renaissance it was pluralistic, machinelike, and mathematically ordered. In the Middle Ages scholars thought in terms of purposes, of ends, of divine intentions; in the Renaissance they thought in terms of forces, mechanical agencies, and physical causes. All of this had become clear by the end of the 15th century. Within the early pages of the *Notebooks* of Leonardo da Vinci, the great Florentine artist, scientist, humanist, and mechanical genius, occur the following three propositions:

Leonardo's three theses

(1) Since experience has been the mistress of whoever has written well, I take her as my mistress, and to her on all points make my appeal.
(2) Instrumental or mechanical science is the noblest and above all others the most useful, seeing that by means of it all animated bodies which have movement perform all their actions.
(3) There is no certainty where one can neither apply any of the mathematical sciences, nor any of those which are based upon the mathematical sciences.

Here are enunciated respectively: (1) the principle of Empiricism, (2) the advocacy of mechanistic science, and (3) the faith in mathematical explanation; and it is upon these three formulations, as upon a rock, that the science and philosophy of the Renaissance built their foundations. From each of Leonardo's theses descended one of the great streams of Renaissance philosophy: from the empirical principle the work of Francis Bacon; from mechanism the work of Thomas Hobbes; and from mathematical explanation that of René Descartes.

Any adequate philosophical treatment of scientific method surely contains both an empirical principle and a faith in mathematical explanation; and, in Leonardo's thinking, as in scientific procedure generally, there need be no conflict between them. Yet they do represent two poles of emphasis, each capable of excluding the other. Moreover, the peculiar accidents of Renaissance scientific achievement did present some evidence for their mistaken separation: for the revival of medical studies on the one hand and the novel blooming of mathematical physics on the other emphasized opposite virtues in scientific methodology. This polarity was represented by the opposing figures of Vesalius and Galileo.

In the mid-16th century Andreas Vesalius, a Belgian physician, was astounding all of Europe with the unbelievable precision of his anatomical dissections and drawings. Having invented new tools for this precise purpose, he successively laid bare the vascular, the neural, and the musculature systems of the human body; and this procedure seemed to demonstrate the virtues of empirical method, of physiological experiment, and of the precision and disciplined skill in sensory observation that made his demonstrations classics of inductive procedure.

Galileo's experiential dynamics

Only slightly later the Italian physicist Galileo, following in the tradition already established by Copernicus and Kepler, founders of modern astronomy (but without their more mystical and metaphysical eccentricities), attempted to do for terrestrial and sidereal movements what Vesalius had managed for the structure of the human body—creating his experiential dynamics, however, with the help

of hypotheses supplied by the quantitative calculations of mathematics. In Galileo's work all of the most original scientific directions of the Renaissance came to a head: the revival of Alexandrian mathematics, the experimental use of new instruments, like the lens and the telescope, the search for certainty in physics based upon the undoubted applicability of mathematical theory, and the underlying faith that the search for absolute certainty in science was reasonable because matter in motion accorded with a model of mathematical simplicity. Galileo's work also deals with some of the recurrent themes of 16th- and 17th-century philosophy: an atomism that relates changes in the relations of physical bodies to the corpuscular motion of their parts, the reduction of all qualitative differences to quantitative reasons, and the resultant important distinction between "primary" and "secondary" qualities. The former—including shape, extension, and specific gravity—were considered to be in fact a constituent part of nature and therefore "real." The latter—such as colour, odour, taste, and relative position—were taken to be simply the effect of bodily movements upon perceiving minds and therefore ephemeral, "subjective," and essentially irrelevant to the nature of physical reality.

Rise of Empiricism and Rationalism. The scientific contrast between Vesalius' rigorous observational techniques and Galileo's reliance upon mathematical theory received further expression in the contrast between the respective philosophies of Francis Bacon and René Descartes. And, indeed, in its more abstract formulation as the contrast between Rationalism and Empiricism, it was to dominate the philosophical controversies of the 17th and 18th centuries and to present a dilemma hardly to be resolved before the advent of Immanuel Kant.

The Empiricism of Francis Bacon. Flourishing about the turn of the 17th century, Sir Francis Bacon (1561–1626) was the outstanding apostle of Renaissance Empiricism. Less an original metaphysician or cosmologist than the advocate of a vast new program for the advancement of learning and the reformation of scientific method, Bacon conceived of philosophy as a new technique of reasoning that should reestablish natural science upon a firm foundation. In the *Advancement of Learning* (1605) he charted the map of knowledge: history, which depends upon the human faculty of memory; poetry, upon that of imagination; and philosophy, upon man's reason. To reason, however, Bacon assigned a completely experiential function. Fifteen years later, in his *Novum Organum*, he made this clear: because, he said, "we have as yet no natural philosophy which is pure, . . . the true business of philosophy must be . . . to apply the understanding . . . to a fresh examination of particulars." A technique for "the fresh examination of particulars" thus constituted his chief claim to philosophical distinction.

Examination of particulars in tables of discovery

Bacon's hope for a new birth of science hinged not only upon vastly more numerous and varied experiments but chiefly upon "an entirely different method, order, and process for advancing experience." This method consisted in the construction of what he called "tables of discovery." He distinguished three kinds: tables of presence, of absence, and of degree (*i.e.,* in the case of any two properties, such as heat and friction, instances in which they appeared together, instances in which one appeared without the other, and instances in which their amounts varied proportionately); and the ultimate purpose of these tables was to order facts in such a way that the true causes of phenomena (the subject of physics) and the true "forms" of things (the subject of metaphysics—the study of the nature of Being) could be inductively established.

Bacon's was no raw Empiricism; his profound sense of fact and his belief in the primacy of observation led him to elicit laws and generalizations. Also, his conception of forms was quite un-Platonic: a form for him was not an essence but a permanent geometric or mechanical structure. His enduring place in the history of philosophy lies, however, in his single-minded advocacy of experience as the only source of valid knowledge and in his profound enthusiasm for the perfection of natural science. It is in this sense that "the Baconian spirit" continued as a source of inspiration: his elaborate classification of the sciences inspiring the French Encyclopaedists of the 18th century and his Empiricism inspiring the English philosophers of science of the 19th century.

The Materialism of Thomas Hobbes. The English political philosopher Thomas Hobbes (1588–1679) was acquainted with both Bacon and Galileo. With the first he shared a strong concern for philosophical method, with the second an overwhelming interest in matter in motion. His philosophical efforts, however, were more inclusive and more complete than those of either of these contemporaries. He was a comprehensive thinker within the scope of an exceedingly narrow set of presuppositions, and he produced one of the most systematic philosophies of the early modern period—an almost completely consistent description of nature, man, and civil society according to the tenets of mechanistic Materialism.

Hobbes's system of physics and moral and civil philosophy

Hobbes's account of what philosophy is and ought to be clearly distinguished between content and method. As method, philosophy is simply reasoning or calculating by the use of words as to the causes of phenomena. When a man reasons forward from causes to effects, he reasons synthetically, and, when he reasons from effects backward to causes, he does so analytically. (His strong deductive and geometric bias favoured the former.) Hobbes's dogmatic metaphysical presupposition was that the basic reality is matter in motion. The real world is a corporeal universe in constant movement, and the phenomena, the causes and effects of which it is the business of philosophy to lay bare, are either the mutual action of bodies or the quaint effects of bodies upon minds. From this assumption follows Hobbes's classification of the fields that form the content of philosophy: (1) physics, (2) moral philosophy, and (3) civil philosophy. Physics is the science of the motions and actions of natural bodies conceived in terms of cause and effect. Moral philosophy (or, more accurately, psychology) is the detailed study of "the passions and perturbations of the mind"—that is, how minds are "moved" by desire, aversion, appetite, fear, anger, and envy. And civil philosophy concerns the concerted actions of men in a commonwealth—how in detail the wayward wills of men are constrained by power (force) in the prevention of civil disorder and the maintenance of peace.

Hobbes's philosophy was a bold Renaissance restatement of Greek Atomistic Materialism with applications to the realities of Renaissance politics that would have seemed strange to its ancient originators. But there are also elements in it that make it characteristically English. For Hobbes's conventionalist account of language led him to a nominalistic position—that is, to a position denying the reality of universals (general or common concepts)—in much the same fashion as did Bacon's exaggerated emphasis on particulars. Moreover, Bacon's general emphasis upon experience also had its analogue in Hobbes's sensationalist theory of knowledge: the notion that all knowledge has its origin in sense impressions and that all sensations are caused by the action of external bodies upon the organs of sense. Empiricism has been a basic and recurrent expression of British mentality, and its nominalistic and sensationalist roots were already clearly evident in both Bacon and Hobbes.

Rationalism of Descartes. But it was not their philosophy that was to dominate the last half of the 17th century but rather that of René Descartes (1596–1650), a French gentleman who signed himself "Lord of Perron" and who lived the 20 most productive years of his life in the tolerant and hospitable Dutch republic. Descartes, a crucial figure in the history of philosophy, combined (however unconsciously or even unwillingly) the influences of the past into a synthesis that was striking in its originality and yet congenial to the scientific temper of the age. In the minds of all later historians he counts as the progenitor of the modern spirit in philosophy.

Synthesis of past and present thought

From the past there seeped into the Cartesian synthesis doctrines about God from Anselm and Aquinas, a theory of the will from Augustine, a deep sympathy with the Stoicism of the Romans, and a skeptical method taken indirectly from the ancient Skeptics Pyrrho and Sextus Empiricus. But Descartes was also a great mathematician, who invented analytic geometry, who made many physical

and anatomical experiments, who knew and profoundly respected the work of Galileo, and who withdrew from publication his own cosmological treatise *Le Monde* ("The World") after Galileo's condemnation by the Inquisition in 1633.

Each of the maxims of Leonardo, which constitute the Renaissance worldview, found its place in Descartes: the Empiricism of his physiological researches described in his *Discours de la méthode* (1637; *Discourse on Method*), the mechanistic interpretations of the physical world and human action detailed in the *Principia Philosophiae* (1644; *Principles of Philosophy*) and *Les Passions de l'âme* (1649; *The Passions of the Soul*), and the mathematical bias that dominates his theory of method in the *Regulae ad Directionem Ingenii* (published 1701; *Rules for the Direction of the Mind*) and his metaphysics in the *Meditationes de Prima Philosophia* (2nd ed. 1642; *Meditations on the First Philosophy*). But of these three, it is the mathematical strain that clearly predominates.

Bacon and Descartes, the founders of modern Empiricism and Rationalism, respectively, shared two pervasive Renaissance tenets: an enormous enthusiasm for physical science; and the belief that knowledge means power—that the ultimate purpose of theoretical science is to serve the practical needs of men.

In his *Principia* Descartes defined philosophy as "the study of wisdom" or "the perfect knowledge of all one can know." Its chief utility is "for the conduct of life" (morals), "the conservation of health" (medicine), and "the invention of all the arts" (mechanics). He expressed the relation of philosophy as theoretical inquiry to practical consequences in the famous metaphor of the tree of philosophy whose root is metaphysics, whose trunk is physics, and whose branches are, respectively, morals, medicine, and mechanics. The metaphor is revealing for it indicates that, for Descartes (as for Bacon and Galileo), the major concern was for the trunk (physics) and that he busied himself with the roots only in order to provide a firm foundation for the trunk. Thus the *Discours de la méthode,* which provides a synoptic view of the Cartesian philosophy, shows it to be not (as with Aristotle or Whitehead) a metaphysics founded upon physics but rather—that more characteristic product of the 17th century—a physics founded upon metaphysics.

Elements in Cartesian synthesis

Descartes's mathematical bias was expressed in his determination to ground natural science not in sensation and probability (as did Bacon) but in a principle of absolute certainty. Thus his metaphysics in essence consisted of three principles:

1. To employ the procedure of complete and systematic doubt to eliminate every belief that does not pass the test of indubitability (skepticism);

2. To accept no idea as certain that is not clear, distinct, and free of contradiction (mathematicism);

3. To found all knowledge upon the bedrock certainty of self-consciousness, so that "I think, therefore I am" becomes the only innate idea unshakable by doubt (subjectivism).

From the indubitability of the self, Descartes deduced the existence of a perfect God; and, from the fact that a perfect being is incapable of falsification or deception, he made the inference that those ideas about the corporeal world that he has implanted within man must be true. The achievement of certainty about the natural world was thus guaranteed by the perfection of God and by the clear and distinct ideas that are his gift.

The Cartesian metaphysics is the fountainhead of Rationalism in modern philosophy, for it suggests that the mathematical criteria of clarity, distinctness, and absence of contradiction among ideas are the ultimate test of meaningfulness and truth. This stance is profoundly antiempirical. Bacon, who had said that "reasoners resemble spiders who make cobwebs out of their own substance," might well have said so of Descartes, for the Cartesian self is just such a substance from which the idea of God originates and with which all deductive reasoning begins. Yet for Descartes the understanding is vastly superior to the senses, and, in the question of what constitutes truth in science, only man's reason can ultimately decide.

Cartesianism after Descartes

Cartesianism was to dominate the intellectual life of the Continent until the end of the 17th century. It was a fashionable philosophy, appealing alike to learned gentlemen and highborn ladies; and it was one of the few philosophical alternatives to the decadent Scholasticism still being taught in the universities. Precisely for this reason it constituted a serious threat to established religious authority. In 1663 the Roman Catholic Church placed Descartes's works on the Index of Forbidden Books, and the University of Oxford forbade the teaching of his doctrines. Only in the liberal Dutch universities, such as Groningen and Utrecht, did Cartesianism make serious headway.

Certain features of the Cartesian philosophy made it an important starting point for subsequent philosophical speculation. Being the meeting ground of the medieval and the modern worldviews, it accepted the doctrines of Renaissance science while attempting to ground them metaphysically in the medieval notions of God and the human mind. Thus a certain dualism between God the Creator and the mechanistic world of his creation and between mind as a spiritual principle and matter as mere spatial extension was inherent in the Cartesian position; and a whole generation of French Cartesians (among them Arnold Geulincx, Nicolas Malebranche, and Pierre Bayle) wrestled with the resulting problems of the interaction and reconciliation between the counterposed entities.

Rationalism of Spinoza and Leibniz. Two philosophers of genius carried on the tradition of continental Rationalism: the Dutch Jew Benedict de Spinoza (1632–77) and his younger contemporary Gottfried Wilhelm Leibniz (1646–1716), a Leipzig scholar and polymath. Bacon's philosophy had been a search for method in science, and Descartes's basic aim had been the achievement of scientific certainty; but Hobbes and Spinoza provided the most comprehensively worked out speculative systems of the early modern period. In certain respects they had much in common: a mechanistic picture of the world, with its events guided by a strict determinism, and even a political philosophy in each case looking for political stability based upon centralized power. Yet Spinoza introduced a conception of philosophizing that was new to the Renaissance: Philosophy became a personal and moral quest for the wisdom of life and for the achievement of human perfection.

Spinoza's aim, terminology, and method

In conducting this search, Spinoza borrowed much of the basic apparatus of Descartes: the aim at a rational understanding of principles, the terminology of "substance" and of "clear and distinct ideas," and a mathematical method that seeks to convert philosophical knowledge into a complete deductive system using the geometric model of Euclid's *Elements.* Spinoza viewed the universe pantheistically as a single infinite substance, which he called "God," with the dual attributes (or aspects) of thought and extension, and which he differentiated into plural "modes" (or particular things); and he attributed to this world as a whole the properties of a timeless logical system—of a complex of completely determined causes and effects. In so doing Spinoza was simply seeking for man the series of "adequate" ideas that furnish the intellect and constitute human freedom. For ultimately, for Spinoza, the wisdom that philosophy seeks is achieved when one perceives the universe in its wholeness, through the "intellectual love of God," which merges the finite individual with the eternal unity and provides the mind with the pure joy that is the final achievement of its search.

Whereas the basic elements of the Spinozistic worldview are given in his one great work, the *Ethics,* Leibniz' philosophy has to be pieced together from numerous brief expositions or fragments, which seem to be mere intermissions, or philosophical interludes, in an otherwise busy life. But the philosophical form is deceptive. Leibniz was a mathematician and jurist (inventor of the infinitesimal calculus and codifier of the laws of Mainz), diplomat, historian to royalty, and court librarian in a princely house; yet he was also one of the most original philosophers of the early modern period. His chief contributions were in the fields of logic, in which he was a truly brilliant innovator, and metaphysics, in which he provided a third alternative to the Rationalist constructions of Spinoza and Descartes. Leibniz saw logic as a mathematical calculus.

He was the first to distinguish "truths of reason" from "truths of fact" and to contrast the "necessary" propositions of logic and mathematics (which express identities), which hold for all possible worlds, with the "contingent" (or empirical) propositions of science, which hold only for certain existential conditions; and he saw clearly that, as "the principle of contradiction" controls the first, so "the principle of sufficient reason" governs the second.

A pluralism of monads

In metaphysics Leibniz espoused pluralism (as opposed to the dualism of Descartes's thought and extension and the monism of Spinoza's single substance, which is God). There were for him an infinite number of spiritual substances (which he called "monads"), each different, each a percipient of the universe around it, and each mirroring that universe from its own point of view. The chief significance of Leibniz, however, lies not in his differences from Descartes and Spinoza but in the extreme Rationalism that all three shared. In the *Principes de la nature et de la grâce fondés en raison* (1714; "Principles of Nature and of Grace Founded in Reason"), he stated the maxim that can stand for the entire school:

> True reasoning depends upon necessary or eternal truths, such as those of logic, numbers, geometry, which establish an indubitable connection of ideas and unfailing consequences.

Literary forms and sociological conditions. The literary forms in which philosophical exposition was couched in the early modern period ranged from the scientific aphorisms of Bacon and the autobiographical meditations of Descartes to the systematic prose of Hobbes and the episodic propositional format of Leibniz. Two basic tendencies, however, can be discerned:

1. The early Renaissance commitment to the dialogue form (already noted), inspired by the rediscovery of the Platonic dialogues;

2. The later prevalence of the systematically ordered treatise, undoubtedly influenced by the enormous prestige of deductive mathematics.

The concept of serial order stressed by geometry, in which the reasoner passes from more universal axioms to more specific derivative propositions, influenced in turn the style of Hobbes, Descartes, and Spinoza. The order of presentation in Hobbes's *Leviathan* and in Descartes's *Principia Philosophiae* reflects this serial concern, while Spinoza's *Ethics* utilizes the Euclidean method so formalistically as almost to constitute an impenetrable barrier to the basic lucidity of his thought.

Medieval philosophy with its texts, readings, learned authorities, Disputed Questions, Quodlibetal Questions (brief academic discussions), and its Summas was characteristically associated with the medieval university. It is a singular fact, therefore, that from the birth of Bacon in 1561 to the death of Hume in 1776—*i.e.,* for 200 years—not one first-rate philosophical mind in Europe was permanently associated with a university.

The fact is that, as the age of the saint passed into that of the gentleman, philosophers too reflected this profound change in their titles, their social status, and their economic situation. Sir Francis Bacon was a lawyer, judge, and attendant upon the royal court. Thomas Hobbes was the tutor and companion of young noblemen. Renê Descartes, son of a noble family, traveled and studied at leisure, retiring to Holland to live out his life on inherited income. Gottfried Wilhelm Leibniz, courtier, diplomat, and scholar, became a privy councillor and baron of the Holy Roman Empire. Thus philosophers often belonged to the lesser nobility or were closely associated with the nobility, to whom—like poets—they dedicated their works; and they lived not by philosophy but for it—either independently, by pensions, gifts, inherited income, or in the households of the nobility.

Thus, philosophy in the 16th and 17th centuries was clearly the preoccupation of a widely scattered elite; and this meant that, despite the printed essay, much philosophical communication took place within a small but at the same time loose and informal circle. Treatises were circulated in manuscript; comments and objections were solicited; and a vast polemical correspondence was built up. Prior to its publication, Descartes prudently sent his *Meditationes* to the theologians of the Sorbonne for comment; and, after its publication, his friend Mersenne sent it to Hobbes, Antoine Arnauld, and Pierre Gassendi, among others, who returned formal "objections," to which Descartes in turn replied. In addition, the 17th century possessed a rich repository of philosophical correspondence, such as the letters that passed between Descartes and the scientist Christiaan Huygens, between Spinoza and Henry Oldenburg (one of the first secretaries of the Royal Society), and between Leibniz and Arnauld. But philosophers were also familiar with the great monarchs and administrators of the age: Descartes gave philosophical instruction to Queen Christina of Sweden, Leibniz was an intimate of Queen Sophia Charlotte of Prussia, and Spinoza enjoyed the personal friendship of the Dutch politician Johan de Witt.

The chief fact, however, remains the sharp separation of creative philosophers from the formal centres of learning: Hobbes expressed extreme contempt for the decadent Aristotelianism of Oxford; Descartes, despite his prudence, scorned the medievalists of the Sorbonne; and Spinoza refused the offer of a professorship of philosophy at Heidelberg with polite aversion. It was to be another 100 years before philosophy returned to the universities.

THE ENLIGHTENMENT

Although they both lived and worked in the late 17th century, Sir Isaac Newton and John Locke were the true fathers of the Enlightenment. Newton was the last of the scientific geniuses of the age, and his great *Philosophiae Naturalis Principia Mathematica* (1687; *Mathematical Principles of Natural Philosophy*) was the culmination of the entire movement that had begun with Copernicus and Galileo—the first great physical synthesis based upon the application of mathematics to nature in every detail. The basic idea of the authority and autonomy of reason, which dominated all philosophizing in the 18th century, was, at bottom, the consequence of Newton's work.

Seminal influence of Newton's *Principia*

Copernicus, Kepler, Bacon, Galileo, and Descartes—scientists and methodologists of science—performed like men urgently attempting to persuade nature to reveal her secrets. Newton's comprehensive mechanistic system made it seem as if at last she had done so. It is impossible to exaggerate the enormous enthusiasm that this assumption kindled in all of the major thinkers of the 18th century from Locke to Kant. The new enthusiasm for reason that they all instinctively shared was based not upon the mere advocacy of propagandists like Descartes and Leibniz but upon the conviction that for the intellectual conquest of the natural world reason had really worked.

Classical British Empiricism and its basic tasks. Two major philosophical problems remained: to account for the genetic origins of the reason that had proved so successful and to shift its application from external nature to man. John Locke's *Essay Concerning Human Understanding* (1690) was devoted to the first; and David Hume's *Treatise of Human Nature* (1739–40), "being an attempt to apply the method of experimental reasoning to moral subjects," was devoted to the second.

Shift to epistemological and Empiricist philosophy

These two basic tasks reflect a shift away from the direction in which philosophizing had moved in the late Renaissance. The Renaissance preoccupation with the natural world had represented a certain "realistic" bias. Hobbes and Spinoza had each produced a metaphysics. They had been interested in the real constitution of the physical world. Moreover, the Renaissance bias in favour of mathematics had generated the profound interest in rational principles, necessary propositions, and innate ideas that was so prevalent in the philosophies of Leibniz and Descartes. On the other hand, the Enlightenment, in turning from the realities of nature to account for the structure of the mind that knows it so successfully and in attempting an experiential account of the furnishings of that mind, settled on the sensory components of knowledge rather than on the merely mathematical. Thus the school of so-called British Empiricism (John Locke, George Berkeley, and David Hume) dominated the perspective of Enlightenment philosophy until the time of Kant. And this school philosophized in terms of ideas rather than things

and of experience rather than innate necessary principles. Whereas the philosophy of the late Renaissance had been metaphysical and Rationalistic, that of the Enlightenment was epistemological and Empiricist.

Origin and nature of reason in Locke and Berkeley. Locke's *Essay Concerning Human Understanding* thus marked a decisively new direction for modern philosophizing because it proposed what amounts to a new criterion of truth. The design of his essay was "to inquire into the origin, certainty, and extent of human knowledge," which involved three tasks:

1. To discover the origin of men's ideas;
2. To exhibit their certainty and evidential value;
3. To examine the claims of all knowledge that is less than certain.

What was crucial for Locke, however, was that the second task is dependent upon the first. Following the general Renaissance custom, Locke defined an "idea" as a mental content, as "whatever is the object of the understanding when a man thinks"; but, whereas for Descartes and the entire Rationalistic school the certainty of ideas had been a function of their self-evidence—*i.e.,* of their clarity and distinctness—for Locke their validity hinged expressly upon the mode and manner of their origin. A genetic criterion of truth and validity replaced an intrinsic one.

Locke's exhaustive survey of mental contents is useful, if elaborate. Though he distinguished between ideas of sensation and ideas of reflection, the whole thrust of his efforts and those of his Empiricist followers was to reduce the latter to the former, to minimize the originative power of the mind in favour of its passive receptivity to the sensory impressions received from without. Locke's classification of ideas into simple and complex was an attempt to distinguish mental contents such as blueness or solidity, which come from a single sense like sight or touch, and those such as figure, space, extension, rest, and motion, which are the product of several senses combined, on the one hand, from those complicated and compounded (complex) ideas of universals such as triangle or gratitude, of substances, and of relations such as identity, diversity, and cause and effect, on the other.

Locke's effort to build a total world

Locke's *Essay* was a dogged attempt to produce the total world of man's conceptual experience out of a set of elementary sensory building blocks, moving always from sensation toward thought and from the simple to the complex. The basic outcome of his epistemology was therefore:

1. That the ultimate source of men's ideas is sensation;
2. That all mental operations are a combining and compounding of simple sensory materials into complex conceptual tools.

It was a theory of knowledge based upon a kind of sensory atomism, which sees the mind as an agency of discovery rather than of creation and views its ideas as "like" the objects that are the sources of the sensations it receives. But it sees also that an important distinction must still be made between those "primary qualities" such as solidity, figure, extension, motion, and rest, which are the actual characteristics of objects themselves, and those "secondary qualities" such as colour, taste, and smell, which are simply the internal consequences of how the mind is affected by them. This important Lockean distinction (already found in Galileo) was only a mirroring in the theory of knowledge of the physics of Sir Isaac Newton, which was its contemporary.

It was precisely this dualism presupposed by the science of the time between primary qualities, which belong to matter, and secondary qualities, which belong to mind, that Locke's successor George Berkeley (1685–1753) sought to overcome. Though Berkeley was a bishop in the Anglican Church who wanted to combat "atheistic Materialism" and to sustain the power of God and spirituality, his importance for the theory of knowledge lies rather in the way in which he demonstrated that, in the end, primary qualities are reducible to secondary qualities. His Empiricism led to a denial of abstract ideas because he believed that general notions are simply fictions of the mind and vehemently denied that one may validly distinguish between objects and the sensory impressions that repose in the mind. Science, he argued, can easily dispense with the concept of matter: nature is simply that which men perceive by their senses, and this means that sense data can be considered as "objects for the mind" rather than as "qualities adhering in a substance." A thing is simply a recurrent group of sense qualities. With this important reduction of substance to quality, Berkeley thus became the true father of the epistemological position known as phenomenalism, which has remained an important influence in British philosophizing to the present day.

Basic science of man in Hume. The third, and in many ways the most famous, of the British Empiricists was the Skeptic David Hume (1711–76). Hume's philosophical intention was to reap, humanistically, the harvest sowed by Newtonian physics, to apply the method of natural science to human nature, and to create a basic science of man. The paradoxical result of this admirable purpose, however, was to create a skeptical crisis more devastating than that of the early French Renaissance and to reduce human certainty once more to the state that it was in before Descartes had reached the dogmatic halting point in his procedure of methodical doubt.

Sense impressions and association of ideas

Hume followed Locke and Berkeley in approaching the problem of knowledge from a psychological perspective. He too found the origin of knowledge in sense impressions. But whereas Locke had found a certain trustworthy order in the compounding power of the mind, and Berkeley had found mentality itself expressive of a certain spiritual power, Hume's relentless analysis discovered as much contingency in mind as in the external world. All uniformity in perceptual experience, he held, comes from "an associating quality of the mind." The "association of ideas" is a fact, but the relations of resemblance, contiguity, and cause and effect that it produces have no intrinsic validity because they are the product of an inexplicable "mental habit." Thus the causal principle upon which all knowledge rests indicates no necessary connections between things but is simply the accident of their constant conjunction in men's minds. Moreover, the mind itself, far from being an independent power, is simply "a bundle of perceptions" without unity or cohesive quality. Hume's denial of a necessary order of nature on the one hand and of a substantial or unified self on the other precipitated a philosophical crisis from which Enlightenment philosophy was not to be definitely rescued until the work of Kant.

Nonepistemological movements in the Enlightenment. Though the school of British Empiricism represented the mainstream of Enlightenment philosophy until the time of Kant, it was by no means the only type of philosophy that the 18th century produced. The Enlightenment, which was based upon a few great fundamental ideas—such as the dedication to reason, the belief in intellectual progress, the confidence in nature as a source of inspiration and value, and the search for tolerance and freedom in political and social institutions—produced many crosscurrents of intellectual and philosophical expression.

Materialism and scientific discovery. The profound influence of Locke spread to France, where it not only produced the skeptical Empiricism of Voltaire but also united with the mechanistic side of the teachings of Descartes to create an entire school devoted to a sensationalistic Materialism. Julien de La Mettrie in *L'Homme-machine* (1747; *Man a Machine*), Étienne de Condillac in his *Traité des sensations* (1754; *Treatise on the Sensations*), and Paul, baron d'Holbach, in his *Système de la nature* (1770; *The System of Nature*) represented a limited worldview in which matter in motion is the only reality, determining the functioning of the human brain and its sensations according to ironclad and necessary laws. This position even found its way into many of the articles of the great French encyclopaedia edited by Denis Diderot and Jean d'Alembert, which was almost a complete compendium of the scientific and humanistic accomplishments of 18th-century intellectual life.

Optimism and the beginning of many disciplines

Though the Middle Ages and the Renaissance had not referred to themselves by these names, the 18th century called itself "the Enlightenment" with self-conscious enthusiasm and pride. It was an age of optimism with a sense of new beginnings. Great strides were made in chemistry and biological science. Jean-Baptiste, chevalier

de Lamarck, Georges Cuvier, and Georges-Louis Leclerc, comte de Buffon, were perfecting a system of animal classification. And, in the eight years between 1766 and 1774, Henry Cavendish discovered hydrogen; Daniel Rutherford, nitrogen; and Joseph Priestley, oxygen. It was the period when foundations were being laid in psychology and the social sciences and in ethics and aesthetics. Anne-Robert-Jacques Turgot, the marquis de Condorcet, and Montesquieu in France, Giambattista Vico in Italy, and Adam Smith in England marked the beginning of history, economics, sociology, and jurisprudence as sciences. Hume, Jeremy Bentham, and the British moral sense philosophers were turning ethics into a specialized field of philosophical inquiry; and Anthony Ashley, 3rd earl of Shaftesbury, Edmund Burke, Johann Gottsched, and Alexander Baumgarten were laying the foundations for a systematic aesthetics.

Social and political philosophy. But, outside of the theory of knowledge, the most significant contribution of the Enlightenment came in the field of social and political philosophy, as Locke's *Two Treatises of Civil Government* (1690) and Jean-Jacques Rousseau's *Du contrat social* (1762; *The Social Contract*) proposed a justification of political association grounded in the newer political requirements of the age. The Renaissance political philosophies of Machiavelli, Bodin, and Hobbes had centred on the absolute power of kings and rulers. But the Enlightenment theories of Locke and Rousseau turned instead to the freedom and equality of citizens. It was a natural historical transformation. The 16th and 17th centuries had constituted the age of absolutism; the political problem had been largely that of internal order, and political theory had been presented in the language of national sovereignty. But the 18th century was the age of the democratic revolutions (including the French Revolution); the political problem was that of freedom and the revolt against injustice, and political theory was expressed in the idiom of natural and inalienable rights.

Freedom, equality, justice, rights

Locke's political theory was an express denial of the divine right of kings and the absolute power of the sovereign as contained in the doctrines of Hobbes. Instead, he insisted that all men have a natural right to freedom and equality. The state of nature in which men originally live is not, as in Hobbes, intolerable, but it has certain inconveniences. Therefore men band together to form society, as Aristotle had taught, "not simply to live, but to live well." Political power can never be exercised apart from its ultimate purpose, which is the common good; for men enter the political contract in order to preserve life, liberty, and property.

Locke thus stated one of the fundamental principles of the liberal tradition: that there can be no subjection to power without consent, although once political society has been founded, there is an obligation to submit to the decisions of the majority. It is the legislature that makes these decisions, although the ultimate power of choosing the legislature rests with the people; and even the powers of the legislature are not absolute, because the laws of nature remain as a permanent standard and as principles of protection against any arbitrary authority.

Rousseau's more radical political doctrines were built upon Lockean foundations. For him, too, the convention of the social contract formed the basis of all legitimate authority among men, although his conception of citizenship was much more organic and much less individualistic than Locke's. The surrender of natural liberty for civil liberty means that all individual rights (among them property rights) become subordinate to the general will. For Rousseau the state is a moral person whose life is the union of its members, whose laws are acts of the general will, and whose end is the liberty and equality of the citizens. It follows that, when any government usurps the power of the people, the social compact is broken; and not only are the citizens no longer compelled to obey but they also have an obligation to rebel. Rousseau's defiant collectivism was clearly a revolt against Locke's systematic individualism; for him the fundamental category is not "natural person" but "citizen." Nevertheless, however much they differed, in these two social theorists of the Enlightenment is to be found the germ of all modern liberalism: its faith in representative democracy, in civil liberties, and in the basic dignity of man.

Professionalization of philosophy. In his *Éléments de philosophie* (1759; "Elements of Philosophy"), Jean d'Alembert, an 18th-century French mathematician and Encyclopaedist, wrote:

> Our century is the century of philosophy par excellence. If one considers without bias the present state of our knowledge, one cannot deny that philosophy among us has shown progress.

D'Alembert was calling attention to that reflective self-examination for which the 18th century is famous, and he was undoubtedly referring to the activities of mathematicians like himself, jurists, economists, and amateur moralists rather than to narrow philosophical specialists. But the 18th century was clearly the "century of philosophy par excellence" in a more technical sense also. For it was the period in which philosophizing first began to pass from the hands of gentlemen and amateurs into those of true professionals. The chief sign of this shift was the return of reputable philosophy to the universities.

University of Halle and Christian Wolff

This transformation first occurred in Germany and is chiefly associated with the University of Halle (founded 1694). In the time of the generation that lies between Leibniz (died 1716) and Kant (born 1724), the philosophical climate changed profoundly. The chief representative of this change was Christian Wolff, who taught philosophy at Halle and was the intermediary between the ideas of Leibniz and those of Kant and the profoundly different conceptions of philosophizing for which they stood.

Kant later called Wolff "the real originator of the spirit of thoroughness in Germany"; and Wolff was indeed a pioneer in those techniques that transform philosophy into a professional discipline—the self-conscious adoption of a systematic approach and the creation of a specialized philosophical vocabulary. Wolff carefully distinguished the various fields of philosophy; wrote textbooks in each of them, which were used in the German universities for many years; and created many of the specialized philosophical terms that have survived to the present day.

The German Enlightenment was the first modern period to produce "specialists in philosophy." In England philosophizing in the universities did not become serious until well after the time of Hume, but already philosophical fields had been sufficiently distinguished to be represented by distinct professorships. The titles professor of mental or moral or metaphysical philosophy, as they arose at Oxford and Cambridge, were the product of the late 18th and early 19th centuries.

Journals and histories of philosophy

Two additional factors of the German Enlightenment are relevant: (1) the founding of the first professional journals and (2) the rising concern of philosophy with its own history. The learned journal, like the scientific society, was an innovation of the 17th century. But what had begun as a general intellectual endeavour became in 18th-century Germany a specifically philosophical enterprise. Journals were published in great numbers; *e.g., Acta Philosophorum* (Halle, 1715–26), *Der Philosophische Büchersaal* (Leipzig, 1741–44; "The Philosophical Book Room"), and the short-lived *Neues philosophisches Magazin* (Leipzig, 1789–91), devoted exclusively to the philosophy of Kant.

More interesting still is the flowering of voluminous German histories of philosophy after 1740. By the 18th century there was already a vast accumulation of historical materials, and the self-conscious feeling that philosophy constitutes a specific field the past of which is worth examination and careful ordering combined with the German spirit of thoroughness to produce a series of massive histories such as Johann Brucker's *Historia Critica Philosophiae,* 6 vol. (1742–44; *The History of Philosophy*); Johann Buhle's *Lehrbuch der Geschichte der Philosophie,* 8 vol. (1796–1804; "Textbook on the History of Philosophy"); Dietrich Tiedemann's *Geist der spekulativen Philosophie von Thales bis Berkeley,* 6 vol. (1791–97; "The Spirit of Speculative Philosophy from Thales to Berkeley"); and Gottlieb Tennemann's *Geschichte der Philosophie,* 11 vol. (1789–1819; excerpted in *A Manual of the History of Philosophy*).

Critical examination of reason in Kant. All of these developments led directly to Immanuel Kant (1724–1804), professor at Königsberg, the greatest philosopher of the modern period, whose works mark the true culmination of the philosophy of the Enlightenment. Historically speaking, Kant's great substantive contribution was to relate both the sensory and the a priori elements in knowledge and thus to mend the breach between the extreme Rationalism of Leibniz and the extreme Empiricism of Hume. But in addition to the brilliant content of his philosophical doctrines, Kant was responsible for three crucial philosophical innovations: (1) a new definition of philosophy, (2) a new conception of philosophical method, and (3) a new structural model for the writing of philosophy.

Kant's definition of philosophy culminated the Enlightenment, for it took reason to be the very heart of the philosophical enterprise. Philosophy's sole task, in his view, is to determine what reason can and cannot do. Philosophy, he said, "is the science of the relation of all knowledge to the essential ends of human reason"; and its true aim is both constructive ("to outline the system of all knowledge arising from pure reason") and critical ("to expose the illusions of a reason that forgets its limits"). Philosophy is thus a calling of great dignity, for its aim is wisdom, and its practitioners are themselves "lawgivers of reason." But in order for philosophy to be "the science of the highest maxims of reason," the philosopher must be able to determine the source, the extent, and the validity of human knowledge and the ultimate limits of reason. And these tasks require a special philosophical method.

Sometimes Kant called this the "transcendental method," but more often the "critical method"; for his purpose was to reject the dogmatic assumptions of the Rationalist school, and his wish was to return to the semiskeptical position with which Descartes had begun before his dogmatic pretensions to certainty took hold. Kant's method was to conduct a critical examination of the powers of an a priori judging reason and to inquire what reason can achieve when all experience is removed. The method was based upon a doctrine that he himself called "a Copernican revolution" in philosophy (by analogy with the change from geocentric to heliocentric cosmology): the assumption not that man's knowledge must conform to objects but that objects must conform to man's apparatus of knowing. The question then became: What is the exact nature of that knowing apparatus?

The transcendental, or critical, method of examining reason

Unlike Descartes, Kant could not question that knowledge exists. Mathematics and Newtonian physics were too real for any Enlightenment mentality to entertain this doubt. Kant's methodological question was rather: How is mathematical and physical knowledge possible? What must be the structure of man's knowing process to have made these sciences secure? The attempt to answer these questions was the task of Kant's great *Kritik der reinen Vernunft* (1781; *Critique of Pure Reason*).

But Kant's great aim was to examine reason not merely in one of its domains but in each of its employments, according to the threefold structure of the human mind that he had taken over from Wolff. Thus the critical examination of reason in thinking (science) is undertaken in the *Kritik der reinen Vernunft*, that of reason in willing (ethics) in the *Kritik der praktischen Vernunft* (1788; *Critique of Practical Reason*), and that of reason in feeling (aesthetics) in *Kritik der Urteilskraft* (1790; *Critique of Judgment*).

Literary forms. The literary format of Enlightenment philosophizing was essentially simple and straightforward. Except for an occasional reversion to the dialogue form, as in Berkeley's *Three Dialogues Between Hylas and Philonous* (1713) and in Hume's *Dialogues Concerning Natural Religion* (1779), it consisted of inquiries, discourses, treatises, dissertations, and essays, generally well written in clear, relatively nontechnical prose. But Kant not only introduced a formidable technical philosophical terminology into his works but was, in fact, the originator of a new philosophical form—the "critique" or "critical examination"—which had its own special architectonics. Each of Kant's three critiques consists of the same division into three parts: (1) an "Analytic," or analysis of reason's right functioning; (2) a "Dialectic," or logic of error, showing the pitfalls into which a careless reason falls; and (3) a "Methodology," which is an arrangement of rules for practice. It is a form that was unique to Kant, but it raised certain problems of "oppositional" thinking, to which 19th-century philosophers such as Hegel, Schopenhauer, and Kierkegaard were subsequently to turn.

THE 19TH CENTURY

The end of the Enlightenment

Kant's death in 1804 formally marked the end of the Enlightenment. The 19th century ushered in new philosophical problems and new conceptions of what philosophy ought to do. It was a century of great philosophical diversity. In the Renaissance the chief intellectual fact had been the rise of mathematics and natural science, and the tasks that this fact imposed upon philosophy determined its direction for two centuries. In the Enlightenment attention had turned to the character of the mind that had so successfully mastered the natural world, and Rationalists and Empiricists had contended for mastery until the Kantian synthesis. As for the 19th century, however, if one single feature of its thought could be singled out for emphasis, it might be called the discovery of the irrational. But many philosophical schools were present, and they contended, one with another, in a series of distinct and powerful oppositions: Pragmatism against Idealism; Positivism against irrationalism; Marxism against liberalism.

Politically the 19th century began with the consulate of Napoleon and ended with the Golden Jubilee of Queen Victoria; but it is the intellectual and social changes that fell in between that have philosophical consequences. These changes were chiefly the Romantic movement of the early 19th century, which was a poetic revolt against reason in favour of feeling; the maturation of the Industrial Revolution, which caused untold misery within society and called forth a multitude of philosophies of social reform; the Revolution of 1848 in Paris, Germany, and Vienna, which symbolized class divisions and first implanted in the European consciousness the concepts of "the bourgeoisie" and "the proletariat"; and, finally, the great surge in biological science with Darwin and the publicizing of the idea of biological evolution. Romanticism influenced both German Idealism and philosophers of irrationalism. Experiences of economic discord and social unrest produced the ameliorative social philosophy of English Utilitarianism and the revolutionary doctrines of Karl Marx. And the developmental ideas of Darwin provided the prerequisites for American Pragmatism.

A synoptic view of philosophy in the 19th century reveals an interesting chronology. The early century was dominated by the German school of absolute Idealism (Johann Fichte, Friedrich Schelling, and G.W.F. Hegel). The midcentury was marked by a rebirth of interest in science and its methods (Auguste Comte in France and John Stuart Mill in England) and by liberal (Mill) and radical (Marx) social theory. The late century saw a second flowering of Idealism, this time in England (T.H. Green, F.H. Bradley, and Bernard Bosanquet) and, with Charles Sanders Peirce and William James, the rise of American Pragmatism. The new philosophies of the irrational in the highly individual thinkers Søren Kierkegaard, Arthur Schopenhauer, and Friedrich Nietzsche ran through the century in its entirety.

German Idealism of Fichte, Schelling, and Hegel. The Enlightenment, inspired by the example of natural science, had accepted certain bounds to the possibility of knowledge; that is, it had recognized certain limits to reason's ability to penetrate ultimate reality because that would require methods that surpass the boundaries of scientific method. In this particular modesty, the philosophies of Hume and Kant were much alike. But the early 19th century marked a resurgence of the metaphysical spirit at its most ambitious and extravagant extreme. German Idealism reinstated the speculative pretensions of Leibniz and Spinoza at their height. This turn was partly a consequence of the Romantic influence but, more importantly, of a new alliance of philosophy not with science but with religion. It was not accidental that all of the great German Idealists were university professors whose fathers were Protestant pastors or who had themselves studied theology: Fichte at Jena and Leipzig (1780–84); Schelling

Resurgence of religiously inspired metaphysics

and Hegel at the Tübingen seminary (1788–95). And it is probably this circumstance that gave to German Idealism its intensely serious, its quasi-religious, and its dedicated character.

The consequence of this religious alignment was that philosophical interest shifted from Kant's *Critique of Pure Reason* (in which he had attempted to account for natural science and denied the possibility of certainty in metaphysics) to his *Critique of Practical Reason* (in which he had explored the nature of the moral self) and his *Critique of Judgment* (in which he had treated of the purposiveness of the universe as a whole). For absolute Idealism was based upon three premises:

1. That the chief datum of philosophy is the human self and its self-consciousness;

2. That the world as a whole is spiritual through and through, that it is, in fact, something like a cosmic Self;

3. That, in both the self and the world, it is not primarily the intellectual element that counts but, rather, the volitional and the moral.

Thus, to understand the self, self-consciousness, and the spiritual universe became for Idealistic metaphysics the task of philosophy.

From the point of view of doctrine, Fichte, Schelling, and Hegel had much in common. Fichte (1762–1814), professor of philosophy at the newly founded University of Berlin (1809–14) and a great symbol of German patriotism through the Napoleonic Wars, combined in a workable unity the subjectivism of Descartes, the cosmic monism of Spinoza, and the moral intensity of Kant. He saw human self-consciousness as the primary metaphysical fact through the analysis of which the philosopher finds his way to the cosmic totality that is "the Absolute." And, just as the moral will is the chief characteristic of the self, so also is it the activating principle of the world. Thus Fichte provided a new definition of philosophizing that made it central in dignity in the intellectual world. The sole task of philosophy is "the clarification of consciousness." And the highest degree of self-consciousness is achieved by the philosopher because he alone recognizes "Mind," or "Spirit," as the central principle of reality.

Hegel: "the rational is the real"

This line of thought was carried further by Georg Wilhelm Friedrich Hegel (1770–1831), Fichte's successor at Berlin and perhaps the single most comprehensive and influential thinker of the 19th century. Kant's problem had been the critical examination of reason's role in human experience. For Hegel, too, the function of philosophy is to discover the place of reason in nature, in experience, and in reality; to understand the laws according to which reason operates in the world. But whereas Kant had found reason to be the form that mind imposes upon the world, Hegel found it to be constitutive of the world itself—not something that mind imposes but that it discovers. As Fichte had projected consciousness from mind into reality, so Hegel projected reason; and the resultant Hegelian dictates—that "the rational is the real" and that "the truth is the whole"—although they express an organic and a totalitarian theory of truth and reality, tend to blur the usual distinctions that previous philosophers had made between logic and metaphysics, between subject and object, and between thought and existence. For the basic tenet of Idealism, that reality is spiritual, generates just such a vague inclusiveness.

To the Fichtean foundations, however, Hegel added one crucial corollary: that the Absolute, or Whole, which is a concrete universal entity, is not static but undergoes a crucial development in time. Hegel called this evolution "the dialectical process." By stressing it, Hegel accomplished two things: (1) he indicated that reason itself is not eternal but "historical," and (2) he thereby gave new meaning and relevance to the changing conditions of human society in history—which added to the philosophical task a cultural dimension that it had not possessed before.

The philosopher's vocation, in Hegel's view, was to approach the Absolute through consciousness, to recognize it as Spirit expressing and developing itself ("realizing itself" was his own phrase) in all of the manifold facets of human life. For struggle is the essence of spiritual existence, and self-enlargement is its goal. For these reasons the various branches of intellect and culture become stages in the unfolding of the World-Spirit:

1. The psychological characteristics of man (habit, appetite, judgment) representing "Subjective Spirit";

2. His laws, social arrangements, and political institutions (the family, civil society, the state) expressing "Objective Spirit";

3. His art, religion, and philosophy embodying "Absolute Spirit."

What began, therefore, in Hegel as a metaphysics of the Absolute ended by becoming a total philosophy of human culture.

Positivism and social theory in Comte, Mill, and Marx. The absolute Idealists wrote as if the Renaissance methodologists of the sciences had never existed. But if in Germany the Empirical and scientific tradition in philosophy lay dormant, in France and in England in the middle of the 19th century it was very much alive. In France, Auguste Comte (1798–1857) wrote his great philosophical history of science, *Cours de philosophie positive* (1830–42; *The Positive Philosophy of Auguste Comte* [abridged]) in six volumes. Influenced by Bacon and the entire school of British Empiricism, by the doctrine of progress put forward by Turgot and Condorcet during the 18th century, and by the very original social reformer Henri de Saint-Simon, Comte called his philosophy "Positivism," by which he meant a philosophy of science so narrow that it denied any validity whatsoever to "knowledge" not derived through the accepted methods of science. But the *Cours de philosophie positive* made its point not by dialectic but by an appeal to the history of thought, and here Comte presented his two basic ideas:

Comte's *Philosophie positive*

1. The notion that the sciences have emerged as sciences in strict order, beginning with mathematics and astronomy, followed by physics, chemistry, and biology in that order, and culminating in the new science of sociology, to which Comte was the first to ascribe the name;

2. The so-called "law of the three stages," which views thought in every field as passing progressively from superstition to science by first being (*a*) religious, then (*b*) abstract, or metaphysical, and finally (*c*) positive, or scientific. Comte's permanent contribution was to initiate an antireligious and an antimetaphysical bias in the philosophy of science that has passed into the 20th century.

In mid-19th-century England the chief representative of the Empirical tradition from Bacon to Hume was John Stuart Mill (1806–73). Mill's theory of knowledge, best presented in his *Examination of Sir William Hamilton's Philosophy* (1865), was not particularly original but rather a judicious combination of the doctrines of Berkeley and Hume; but it symbolized his mistrust of vague metaphysics, his denial of the a priori element in knowledge, and his determined opposition to any form of intuitionism. It is in his enormously influential *System of Logic* (1843), however, that Mill's chief theoretical ideas are to be found.

This work, as part of its subtitle, *the Principles of Evidence and the Methods of Scientific Investigation,* shows, was concerned less with formal logic than with scientific methodology. Mill made here the fundamental distinction between deduction and induction, defined induction as the process by which men discover and prove general propositions, and presented his "four methods of experimental inquiry" as the heart of the inductive method. These methods were, in fact, only an enlarged and refined version of Francis Bacon's "tables of discovery." But the most significant section of *A System of Logic* was its conclusion: book 6, "On the Logic of the Moral Sciences."

Mill had taken men's experience of the uniformity of nature as the warrant of induction. Here he reaffirmed the belief of Hume that it is possible to apply the principle of causation and the methods of physical science to moral and social phenomena. These may be so complex as to yield only "conditional predictions," but in this sense there are "social laws." Thus Comte and Mill agreed upon the possibility of a true social science.

Mill's *Logic* was extremely influential, and it continued to be taught at Oxford and Cambridge well into the 20th century; but in the end his importance lay less in logic and

theory of knowledge than in ethics and political theory. For Mill was the great apostle of political liberalism in the 19th century, a true follower of John Locke. And, just as Locke and Rousseau had represented the liberal and the radical wings of social theory in the early modern period, so Mill and Karl Marx represented the liberal and the radical approaches to social reform 100 years later.

Mill's political liberalism and ethics

Raised by social reformers (his father, James Mill, and Jeremy Bentham) to be a social reformer himself, Mill's social theory was an attempt, by gradual means arrived at democratically, to combat the evils of the Industrial Revolution. His ethics, expressed in his *Utilitarianism* (1861), followed the formulations of Bentham in finding the end of society to consist in the production of the greatest quantity of happiness for its members, but he gave to Bentham's cruder (but more consistent) doctrines a humanistic and individualistic slant. Thus, the moral self-development of the individual becomes the ultimate value in Mill's ethics.

This trend was also expressed in his essay *On Liberty* (1859) and *Considerations on Representative Government* (1861). In the former he stated the case for the freedom of the individual against "the tyranny of the majority," presented strong arguments in favour of complete freedom of thought and discussion, and argued that no state or society has the right to prevent the free development of human individuality. In the latter he provided a classic defense for the principle of representative democracy, asked for the adequate representation of minorities, urged renewed public participation in political action for necessary social reforms, and pointed up the dangers of class-oriented or special-interest legislation.

Marx and social revolution

A radical counterbalance to Mill's liberal ideas was provided by Karl Marx (1818–83), a German philosopher, social revolutionary, and political economist. Taking over from Hegel the idea of estrangement (which Hegel had used in a metaphysical sense), Marx used this notion prior to 1848 to indicate the alienation of the worker from the enjoyment of the products of his work, the crass treatment of human labour as a mere commodity (and man as a thing), and, in fact, the general dehumanization of man in a selfish, profit-seeking capitalist society.

In the famous *Communist Manifesto* (1848), Marx, yielding to the revolutionary temper of the times, called (as Rousseau had done before the French Revolution) for the violent overthrow of the established order. All of history, Marx said, is the struggle between exploiting minorities and the underlying population, that is, between bourgeois and proletarians; and he advocated the formation of a Communist Party to stimulate proletarian class consciousness toward the seizure of power and the institution of a just and democratically managed Socialist society.

Marx's revolutionary fervour may have tended to dampen his philosophical reputation in the West, and his philosophical achievement remains a controversial point; but certain of his ideas (some Hegelian in inspiration, some original) have endured. Among these are the ideas:

1. That society is a moving balance (dialectic) of antithetical forces that produce social change;
2. That there is no conflict between a rigid economic determinism and a program of revolutionary action;
3. That ideas (including philosophical theories) are not purely rational and thus independent of external circumstance but depend upon the nature of the social order in which they arise.

Independent and irrationalist movements. The end of the 19th century saw a flowering of many independent philosophical movements. After Hegel was almost forgotten in Germany, his influence migrated to England, where T.H. Green, F.H. Bradley, and Bernard Bosanquet initiated a new Hegelian renaissance. Bradley's *Appearance and Reality* (1893) constituted the high-water mark of the rediscovery of Hegel's dialectical method. In America a strong reaction against Idealism led to the beginnings of the Pragmatic movement initiated by Charles Sanders Peirce and William James. Peirce was an exact logician who recognized that the function of all inquiry is to eradicate doubt and that the meaning of a concept consists in the practical consequences that the concept might be said to have. James transformed Peirce's pragmatic theory of meaning into a pragmatic theory of truth and, in his *Will to Believe* (1897), asserted that men have a right to believe even in the face of inconclusive evidence and that, because knowledge is an instrument for the sake of life, the true test of a belief is the practical consequences that it entails. Meanwhile, in Austria, Franz Brentano, who taught at Vienna from 1874 to 1895, and Alexius Meinong, who taught at Graz, were developing an empirical psychology and a theory of objects, which were to have considerable influence upon the new philosophy of Phenomenology.

Irrationalism in Kierkegaard and Schopenhauer

It was not, however, any of these late 19th-century developments but, rather, the emphasis upon the irrational, which started almost at the century's beginning, that gave the philosophy of the period its peculiar flavour. Hegel, despite his commitment to systematic metaphysics, had, nonetheless, carried on the Enlightenment tradition of faith in human rationality. But soon his influence was challenged from two different directions. The Danish Christian thinker Søren Kierkegaard (1813–55) challenged the logical pretensions of the Hegelian system; and one of his contemporaries, Arthur Schopenhauer (1788–1860), himself a German Idealist and constructor of a bold and imaginative system, contradicted Hegel by asserting that the irrational is the truly real.

Kierkegaard's criticism of Hegel was an appeal to the concrete as against the abstract. He satirized Hegelian Rationalism as a perfect example of "the academic in philosophy"—of detached, objective, abstract theorizing and system building, blind to the realities of human existence, to its subjective, living, emotional character. What a man requires in life, said Kierkegaard, is not infinite inquiry but the boldness of resolute decision and commitment. Man's essence is not contained in thinking but in the existential conditions of his emotional life, in his anxiety and despair. The titles of three of Kierkegaard's books—*Frygt og baeven* (1843; *Fear and Trembling*), *Begrebet angest* (1844; *The Concept of Dread*), *Sygdommen til døden* (1849; *The Sickness unto Death*)—indicate his preoccupation with states of consciousness quite unlike the usual philosophical concentration on cognition and the validity of knowledge.

Schopenhauer, even though, for a short time, he competed unsuccessfully with Hegel at the University of Berlin, soon withdrew and thereafter waged a lifelong battle against academic philosophy. His own system, though orderly and carefully worked out, was written in a vivid and engaging style. Schopenhauer agreed with Kant that the world of appearances, of phenomena, is governed by the conditions of space, time, and causality. But he held that science, which investigates this world, cannot itself penetrate the real world behind appearances, which is dominated by a strong, blind, striving, universal cosmic Will that expresses itself in the vagaries of human instinct, in sexual striving, and in the wild uncertainties of all animal behaviour. Everywhere in nature one sees strife, conflict, and inarticulate impulse; and these, rather than rational processes or intellectual clarity, are man's true contacts with ultimate reality.

Nietzsche's Dionysian irrationalism

Schopenhauer's great work *Die Welt als Wille und Vorstellung* (*The World As Will and Idea*) was published in 1819, and Kierkegaard's uneven masterpieces appeared between 1843 and 1849. But Friedrich Nietzsche (1844–1900), the third of the irrationalist triumvirate, wrote between 1872 and 1889. A prolific but unsystematic writer, presenting his patchwork of ideas in swift atoms of thought, Nietzsche saw the task of the philosopher as that of a man who destroys old values, creates new ideals, and through them a new civilization. He agreed with Schopenhauer that mind is an instrument of instinct to be used in the service of life and power, and he held that illusion is as necessary to man as truth. Nietzsche spent much time in the analysis of such states as resentment, guilt, bad conscience, and self-contempt.

Kierkegaard, Schopenhauer, and Nietzsche provided for the 19th century a new, nonrational conception of human nature, and they viewed the mind not with the rational clarity of Locke and Hume but as something dark, obscure, hidden, deep. But, above all, they initiated a new

style of philosophizing. Schopenhauer wrote like an 18th-century essayist; Kierkegaard was a master of the methods of irony and paradox; and Nietzsche used aphorism and epigram with conscious literary intention. For these three, the philosopher should be less a crabbed academician than a man of letters.

THE 20TH CENTURY

Professionalism in philosophy

Despite the tradition of philosophical "professionalism" established during the Enlightenment by Wolff and Kant, philosophy in the 19th century was still largely created outside the universities. Comte, Mill, Marx, Kierkegaard, and Schopenhauer were not professors, and only the German Idealist school was rooted in academic life. Since the mid-20th century, however, most well-known philosophers have been associated with academia. Philosophers more and more employ a technical vocabulary and deal with specialized problems, and they write not for a broad intellectual public but for one another. Professionalism also has sharpened the divisions between philosophical schools and made the question of what philosophy is and what it ought to be a matter of the sharpest controversy.

The most significant divisions in 20th-century philosophy were influenced and intensified by geographic and cultural differences. The tradition of clear logical analysis, inaugurated by Locke and Hume, dominated the English-speaking world, whereas a speculative and broadly historical tradition, begun by Hegel but later diverging radically from him, held sway on the European Continent. From the early decades of the century the substantive as well as stylistic differences between the two approaches—known after World War II as Analytic and Continental philosophy, respectively—gradually became more pronounced, and until the 1990s few serious attempts were made to find a common ground between them.

Other less significant currents in 20th-century philosophy were the speculative philosophies of Henri Bergson of France, John Dewey of the United States, and Alfred North Whitehead of England, as well as the philosophical Marxism practiced in the former Soviet Union and eastern Europe until the collapse of communism there in 1990–91.

Individual philosophies of Bergson, Dewey, and Whitehead. In his *Introduction à la métaphysique* (1903; *An Introduction to Metaphysics*) and in *L'Évolution créatrice* (1907; *Creative Evolution*), Bergson distinguished between two ways of knowing: the method of analysis, which is characteristic of science, and the method of intuition, a kind of intellectual sympathy through which it is possible to enter into objects and other persons and identify with them. All basic metaphysical truths, Bergson held, are grasped by philosophical intuition.

For Whitehead, philosophy was primarily metaphysics, or "speculative philosophy," the effort "to frame a coherent, logical, necessary system of general ideas in terms of which every element of our experience can be interpreted." Whitehead's philosophy was, thus, an attempt to survey the world with a large generality of understanding.

Whereas Bergson and Whitehead were principally metaphysicians and philosophers of culture, Dewey was a generalist who stressed the unity and interrelationship of all forms of philosophical knowledge. He is chiefly notable for the fact that his conception of philosophy stressed so powerfully the notions of practicality and moral purpose. One of the guiding aims of Dewey's philosophizing was the effort to find the same warranted assertibility for ethical and political judgements as for scientific ones.

Dewey's solution to the social problems of the 20th century, unlike that of Lenin, emphasized not revolution but the continuous application of intelligence to social affairs. He believed in social planning, in conscious intelligent intervention to produce desirable social change; and he proposed a new "experimentalism" as a guide to enlightened public action to promote the aims of a democratic community. His pragmatic social theory is the first major political philosophy produced by modern liberal democracy.

Marxist thought. The framework of 19th-century Marxism, augmented by philosophical suggestions by Lenin, served as the starting point of all philosophizing in the Soviet Union and its eastern European satellites. Much of Lenin's thinking was also devoted to more practical issues, however, such as tactics of violence and the role of the Communist Party in bringing about and consolidating the proletarian revolution. Later Marxism continued this practical concern. Marxism (like Pragmatism) assimilated theoretical issues to practical needs. It asserted the basic unity of theory and practice by finding that the function of the former was to serve the latter. Marx and Lenin both held that theory was always in fact expressive of class interests; consequently, they wished philosophy to be transformed into a tool for furthering the class struggle. The task of philosophy was not abstractly to discover the truth but concretely to forge the intellectual weapons of the proletariat. Thus, philosophy became inseparable from ideology.

Theory and practice in Marxism

Analytic philosophy. It is difficult to give a precise definition of Analytic philosophy, since it is not so much a specific doctrine as an overlapping set of approaches to problems. Its 20th-century origin is often attributed to the work of G.E. Moore (1873–1958). In *Principia Ethica* (1903), Moore argued that the predicate "good," which defines the sphere of ethics, is "simple, unanalyzable, and indefinable." His contention was that many of the difficulties in ethics, and indeed in philosophy generally, arise from an "attempt to answer questions, without first discovering precisely *what* question it is which you desire to answer." These questions thus require analysis for their clarification. Philosophers in this tradition have generally agreed with Moore that the purpose of analysis is the clarification of thought. Their varied methods have included the creation of symbolic languages as well as the close examination of ordinary speech, and the objects to be clarified have ranged from concepts to natural laws and from notions that belong to the special sciences—such as mass, force, and testability—to ordinary terms such as "responsibility" and "see." From its inception, Analytic philosophy has also been highly problem-oriented. There is probably no major philosophical problem that its practitioners have failed to address.

Influence of symbolic logic

The development of Analytic philosophy was significantly influenced by the creation of symbolic (or mathematical) logic at the beginning of the century. Although there are anticipations of this kind of logic among the Stoics, its modern forms are without exact parallel in Western thought, a fact that is made apparent by its close affinities with mathematics and science. Many philosophers thus regarded the combination of logic and science as a model that philosophical inquiry should follow. Still others rejected this model or minimized its usefulness for dealing with philosophical problems. The century thus witnessed the development of two distinct streams of analysis, one of them emphasizing formal (logical) techniques, and the other informal (ordinary-language) ones. There were, of course, many philosophers whose work was influenced by both approaches. Although analysis can in principle be applied to any subject matter, its central focus for most of the century was language, and especially the notions of meaning and reference. Ethics, aesthetics, religion, and law were also fields of interest, though to a lesser degree. In the last quarter of the century there was a profound shift in emphasis from the topics of meaning and reference to issues about the human mind, including the nature of mental processes such as thinking, judging, perceiving, believing, and intending, as well as the products or objects of such processes, including representations and visual images. At the same time, intensive work continued on the theory of reference, and the results obtained in that domain were transferred to the analysis of mind. Both formalist and informalist approaches exhibited this shift in interest.

The formalist tradition. (1) Logical Atomism. The first major development in the formalist tradition was a metaphysical theory known as Logical Atomism, which was derived from work in mathematical logic by the English philosopher Bertrand Russell (1872–1970). Russell's work in turn was based in part on early notebooks written before World War I by Russell's former pupil Ludwig Wittgenstein (1889–1953). In "The Philosophy of Logical Atomism," a monograph published in 1918, Russell gave credit to Wittgenstein for supplying "many of the theories" con-

tained in it. Wittgenstein had joined the Austrian army when the war broke out, and Russell had been out of contact with him since that time. Wittgenstein thus did not become aware of Russell's version of Logical Atomism until after the war. Wittgenstein's own polished and very sophisticated version appeared in the *Tractatus-Logico Philosophicus,* which he wrote during the war but did not publish until 1922.

Both Russell and Wittgenstein believed that mathematical logic could reveal the basic structure of reality, a structure that is hidden beneath the cloak of ordinary language. In their view, the new logic showed that the world is made up of simple, or "atomic," facts, which in turn are made up of particular objects. Atomic facts are mind-independent features of reality, such as the fact that a particular rock is white or the fact that the Moon is a satellite of the Earth. As Wittgenstein says in the *Tractatus,* "The world is determined by the facts, and by their being *all* the facts." Both writers held that the basic propositions of logic, which Wittgenstein called "elementary propositions," refer to atomic facts. There is thus an immediate connection between formal languages, such as the logical system of Russell's *Principia Mathematica* (written with Alfred North Whitehead and published between 1910 and 1913), and the structure of the real world: Elementary propositions represent atomic facts, which are constituted by particular objects, which are the meanings of logically proper names. Russell differed from Wittgenstein in that he held that the meanings of proper names are "sense data," or immediate perceptual experiences, rather than particular objects. Further, for Wittgenstein, but not for Russell, elementary propositions are connected to the world by being structurally isomorphic to atomic facts—*i.e.,* by being a "picture" of them. Wittgenstein's view thus came to be known as the picture theory of meaning.

The picture theory of meaning

Logical Atomism rested upon a number of theses. It was Realistic, as distinct from Idealistic, in its contention that there are mind-independent facts. But it presupposed that language is mind-dependent—*i.e.,* that language would not exist unless there were sentient beings who used sounds and marks to refer and to communicate. Logical Atomism was thus a dualistic metaphysics that described both the structure of the world and the conditions that any particular language must satisfy in order to represent it. Although its career was brief, its guiding principle—that philosophy should be scientific and grounded in mathematical logic—was widely acknowledged throughout the century.

(2) Logical Positivism. Logical Positivism was developed in the early 1920s by a group of Austrian intellectuals, mostly scientists and mathematicians, who named their association the Vienna Circle (*Der Wiener Kreis*). The Logical Positivists accepted the Logical Atomist conception of philosophy as properly scientific and grounded in mathematical logic. By "scientific," however, they had in mind the classical Empiricism handed down from Locke and Hume, and in particular the view that all factual knowledge is based on experience. Unlike Logical Atomism, the Logical Positivists held that only the special sciences can make statements that are meaningful, or cognitively significant. They thus regarded metaphysical, religious, ethical, literary, and aesthetic pronouncements as literally nonsense. Since Logical Atomism was a metaphysics purporting to convey true information about the structure of reality, it too was disavowed. The Positivists also held that there is a fundamental distinction to be made between "analytic" statements (such as "All husbands are married"), which can be known to be true independent of any experience, and "synthetic" statements (such as "The snow is deep here"), which are knowable only through observation.

The main proponents of Logical Positivism—Rudolf Carnap, Herbert Feigl, Philipp Frank, and Gustav Bergmann—all immigrated to the United States from Germany and Austria to escape the rise of Nazism. Their influence on American philosophy was profound, and with various modifications Logical Positivism was still a vital force on the American scene at the end of the century.

(3) Naturalized Epistemology. The philosophical psychology and philosophy of mind developed from the 1950s by the American philosopher W.V.O. Quine (1908–2000), known generally as Naturalized Epistemology, was influenced both by Russell's work in logic and by Logical Positivism. Quine's philosophy forms a comprehensive system that is scientistic, Empiricist, and behaviourist. Indeed, for Quine, the basic task of an Empiricist philosophy is simply to describe how our scientific theories about the world—as well as our prescientific, or intuitive, picture of it—are derived from experience. As he writes: "The stimulation of his sensory receptors is all the evidence anybody has had to go on, ultimately, in arriving at his picture of the world. Why not just see how this construction really proceeds? Why not settle for psychology?" Although Quine shared the Logical Postivists' scientism and Empiricism, he crucially differed from them in rejecting the traditional analytic-synthetic distinction. For Quine, this distinction is ill-founded because it is not required by any adequate psychological account of how scientific (or prescientific) theories are formulated. Quine's views had an enormous impact on 20th-century thought, and at the end of the century he was generally regarded as the dominant figure in Analytic philosophy.

Rejection of the analytic-synthetic distinction

(4) Identity theory, functionalism, and eliminative materialism. Logical Positivism and Naturalized Epistemology were forms of Materialism. Beginning about 1970, these approaches were applied to the human mind and gave rise to three general viewpoints: the identity theory, which held that mental states are identical with physical states of the brain; functionalism, which contended that psychology is not reducible to physics or biology but that each instance of a mental state is numerically identical with a physical state in a particular physical system; and eliminative materialism, which held that the familiar concepts we use to describe our everyday mental lives—concepts such as belief, desire, and intention—do not refer to anything real. According to this view, there are no such things as beliefs, desires, and intentions; instead, there is simply neural activity in the brain. A mature science of the mind has as much need for such entities as modern chemical theory has for the discarded notion of phlogiston. A complete account of so-called "mentation" can be achieved simply by describing how the brain operates.

The informalist tradition. Generally speaking, philosophers in the informalist tradition viewed philosophy as an autonomous activity that should acknowledge the importance of logic and science but not treat either or both as models for dealing with conceptual problems. The 20th century witnessed the development of three such approaches, each of which had sustained influence: common sense philosophy, ordinary language philosophy, and speech act theory.

(1) Common sense philosophy. Common sense philosophy originated as a reaction against forms of Idealism and Skepticism that were prevalent in England at the turn of the century. The first major work of this approach was G.E. Moore's paper "A Defence of Common Sense" (1925). Against Skepticism Moore argued that he and other human beings have known many propositions about the world to be true with certainty. Among these propositions are: "The Earth has existed for many years" and "Many human beings have existed in the past and some still exist." Because Skepticism maintains that nobody knows any proposition to be true, Skepticism can be dismissed. Furthermore, since these propositions entail the existence of material objects, Idealism, which holds that the world is wholly mental, can also be rejected. Moore called this outlook "The Common Sense View of the World" and insisted that any philosophical system whose propositions contravene it can be rejected out of hand without further analysis.

Moore and Skepticism

(2) Ordinary language philosophy. The two major proponents of ordinary language philosophy were Gilbert Ryle (1900–76) and J.L. Austin (1911–60). Both held, though for different reasons, that philosophical problems frequently arise through a misuse or misunderstanding of ordinary speech. In *The Concept of Mind* (1949), Ryle argued that the traditional conception of the human mind—namely, that it is an invisible, ghostlike entity occupying a physical body—is based on what he called a category mistake. The mistake is to treat the terms *mind* and *body* as if

both denoted entities, the one invisible, the other visible. His diagnosis of this error involved an elaborate description of how mental epithets actually work in ordinary speech. To speak of intelligence, for example, is to describe how human beings respond to certain kinds of problematic situations. Ryle insisted that he was not a behaviourist and claimed that he was instead "charting the logical geography" of the mental concepts we all use in everyday life.

Austin's emphasis was somewhat different. In a celebrated paper, "A Plea for Excuses" (1956), he explained that the appeal to ordinary language in philosophy should be regarded as the first word but not the last word. In other words, he asserted that one should be sensitive to the nuances of everyday speech in approaching conceptual problems, but that in certain circumstances everyday speech could be and should be augmented by technical concepts.

The "first-word" principle

According to the "first-word" principle, because ordinary language has been used for eons to draw various distinctions—males from females, friends from enemies, and so forth—the drawing of such distinctions is essential to everyday life; but of equal importance is the fact that such distinctions are more than verbal. They pick out actual features of the world. We can thus assume that if a distinction has existed in ordinary language for a long time, it must have a correlate in the real world. Starting from this principle, Austin dealt with major philosophical difficulties such as the problem of other minds, the nature of truth, and the nature of responsibility.

(3) Speech act theory. Austin was also the creator of one of the most original philosophical theories of the century: speech act theory. A speech act is an utterance that is grammatically similar to a statement but is neither true nor false, though it is perfectly meaningful. Austin called such utterances "performatives." For example, in saying "I do" in the normal circumstances of marrying, one is not saying anything that is either true or false; rather, one is doing something—in this case, getting married. According to Austin, a speech act is a case in which in saying something one is doing something. The theory of speech acts was thus a profound criticism of the Positivist thesis that every meaningful sentence is either true or false. This view, according to Austin, was overly influenced by the tendency to think of ordinary language as a logical calculus or as part of a scientific theory. Both of these forms of discourse emphasize propositions, statements, and assertions and more or less ignore other functions of language.

After Austin's death in 1960, speech act theory was deepened and refined by his American student John R. Searle. In *The Construction of Social Reality* (1995), Searle showed that many social and political institutions are created by speech acts. A government, for example, prints designs on pieces of paper and declares those items to be money. Many institutions, such as banks, universities, and police departments, are social entities created by similar speech acts. Searle's development of speech act theory was thus an unexpected extension of the philosophy of language into social and political theory. (Av.S./Ed.)

Continental philosophy. Analytic philosophy had comparatively little influence in the European Continent, where the speculative and historical tradition remained strong. Dominated by Phenomenology and Existentialism during the first half of the century, after World War II Continental philosophy came to embrace increasingly far-reaching structuralist and "poststructuralist" critiques of metaphysics and philosophical rationality.

The Phenomenology of Husserl and Heidegger. Considered the father of Phenomenology, Edmund Husserl (1859–1939), a German mathematician turned philosopher, was an extremely complicated and technical thinker whose views changed considerably over the years. His chief contributions were the phenomenological method, which he developed early, and the concept of the "life-world," which appeared only in his later writings. As a technique of phenomenological analysis, the phenomenological method was to make possible "a descriptive account of the essential structures of the directly given." It was to isolate and lay bare the intrinsic structure of conscious experience by focusing the philosopher's attention on the pure data of consciousness, uncontaminated by metaphysical theories or scientific or empirical assumptions of any kind. Husserl's concept of the life-world was similarly concerned with immediate experience. It is the individual's personal world as he directly experiences it, with the ego at the centre and with all of its vital and emotional colourings.

Husserl's most notable student was Martin Heidegger (1889–1976), whose famous *Sein und Zeit* (*Being and Time*) appeared in 1927. In Heidegger's work the influence of the phenomenological method is clear: throughout his startlingly original investigations of human existence, with its unique dimensions of being-in-the-world, dread, care, and being toward death, Heidegger adheres to the phenomenological principle that philosophy is not empirical but is the strictly self-evident insight into the structure of experience. Later, the French philosophical psychologist Maurice Merleau-Ponty, building on the concept of the life-world, used the notions of the lived body and its facticity to build up a hierarchical order of man's lived experience.

Sein und Zeit

The Existentialism of Jaspers and Sartre. Existentialism, true to its roots in Kierkegaard and Nietzsche, was oriented toward two major themes: the analysis of human existence, or Being, and the centrality of human choice. Thus its chief theoretical energies were devoted to ontology and decision.

Existentialism as a philosophy of human existence found its best spokesman in the German Karl Jaspers (1883–1969), who came to philosophy from medicine and psychology. For Jaspers, as for Dewey, the aim of philosophy is practical. But whereas for Dewey philosophy is to guide human action, for Jaspers its purpose is the revelation of Being, "the illumination of existence," the answering of the question of what man is and what he can become. This illumination is achieved, and Being is revealed most profoundly, through the experience of those "extreme" situations that define the human condition—conflict, guilt, suffering, and death.

Existentialism as a philosophy of human decision is best exemplified in the work of the French philosopher and man of letters Jean-Paul Sartre (1905–80). Sartre, too, is concerned with Being and with the dread experienced before the threat of Nothingness. But he finds the essence of this Being in liberty—in freedom of choice and the duty of self-determination—and therefore devotes much effort to describing the human tendency toward "bad faith," reflected in perverse attempts to deny one's own responsibility and to flee from the truth of one's inescapable freedom. Sartre does not overlook the legitimate obstacles to freedom presented by the facts of place, past, environment, society, and death; but he demands that one surmount these limitations through acts of conscious decision. For only in acts of freedom does human existence achieve authenticity. In *The Second Sex* (1949), Sartre's fellow philosopher and life-long companion, Simone de Beauvoir, attempted to mobilize the Existentialist concept of freedom for the ends of modern feminism.

After World War II, Sartre came to believe that his philosophy of freedom had wrongly ignored problems of social justice, and in his later work, especially the *Critique of Dialectical Reason* (1960), he sought to reconcile Existentialism with Marxism.

Postwar trends. The main theme of postwar Continental philosophy, however, was the enthusiastic reception in France of Nietzsche and Heidegger and the consequent rejection of metaphysics and the Cartesian rationalism inherited by Sartre and his fellow Existentialists. For millennia, the goal of metaphysics, or "first philosophy," had been to discern the ultimate nature of reality. Postwar Continental philosophy, recoiling from omnipresent images of mass annihilation, increasingly held metaphysical holism itself responsible for the catastrophes of 20th-century history. The critics of metaphysics argued that only a relentless castigation of such excesses could produce a philosophy that was genuinely open toward Being, thinghood, and world.

Rejection of metaphysics

In the 1950s, French philosophy faced a series of major challenges arising from structuralism, the new movement in anthropology that analyzed cultures as systems of structurally related elements and attempted to discern the uni-

versal patterns underlying all such systems. In his *A World on the Wane* (1955), for example, the anthropologist Claude Lévi-Strauss issued a pointed indictment of philosophical method, claiming that it lacked empirical grounding and was so arbitrary as to be capable of proving or disproving anything. Sartre's political missteps during the early 1950s, when he had been an enthusiastic fellow-traveler of the French Communist Party, did little to enhance the credibility of his philosophical rationalism.

In his influential book *The Order of Things* (1966), the French philosopher and intellectual historian Michel Foucault paradoxically employed structuralist methods to criticize the scientistic pretensions of natural history, linguistics, and political economy, the disciplines known in France as the "human sciences." But the main target of his critique was the anthropocentric, or "man-centred," orientation of the humanities, notably including philosophy. Foucault provocatively argued that "man" was an artificial notion, an invention of the 19th century, and that its obsolescence had become apparent in the postwar era.

In later books, such as *Discipline and Punish* (1975) and *The History of Sexuality* (1976), Foucault's gaze shifted to systems of power. In a Nietzschean spirit, he coined the term "power-knowledge" to indicate that knowledge is always implicated in the maintenance of power relations. As he argued in the essay "Nietzsche, Genealogy, History," an examination of "truth" reveals that "all knowledge rests upon injustice, that there is no right, not even in the act of knowing, to truth or a foundation for truth, and that the instinct for knowledge is *malicious* (something murderous, opposed to the happiness of mankind)."

Deconstruction

The movement known as "deconstruction," derived mainly from work begun in the 1960s by the French philosopher Jacques Derrida, displayed a similar hostility to metaphysics and its quest for totality and absolute truth. Under the sway of Heidegger's call for "a destruction of the history of ontology," Derrida endorsed the deconstruction of Western philosophy—*i.e.*, the uncovering and undoing of the false conceptual dichotomies, or "oppositions," inherent in philosophical thinking since the time of the ancient Greeks. In Derrida's view, these oppositions ultimately derive from a "metaphysics of presence," which is the tendency to conceive fundamental philosophical notions such as truth, reality, and being in terms of ideas such as identity, presence, and essence and to limit or ignore the equally valid notions of otherness, absence, and difference. Because of this tendency, Derrida concludes, there is a necessary relationship between the metaphysical quest for "totality" and political "totalitarianism." As he expressed this insight in an early essay, "Violence and Metaphysics": "Incapable of respecting the Being and meaning of the other, phenomenology and ontology would be *philosophies of violence*. Through them, the entire philosophical tradition ... would make common cause with oppression and technico-political possession." In *Philosophy and the Mirror of Nature* (1979), the American Pragmatist philosopher Richard Rorty, agreeing with Foucault and Derrida, claimed that there was no such thing as objective truth. He urged that philosophy abandon its search for the true nature of reality and instead become more like literature.

The French philosopher Emmanuel Levinas attributed the misguided quest for totality to a defect in reason itself. In his major work, *Totality and Infinity* (1961), he contended that, as it is used in Western philosophy, reason enforces "domination" and "sameness" and destroys plurality and otherness. He called for the transcendence of reason in a first philosophy based not on logic but on ethics—and in particular on the Sixth Commandment, "Thou shalt not kill."

It is no small irony, then, that Continental philosophy, whose roots lay in the attempt by Kant, Hegel, and their successors to defend reason against the twin excesses of dogmatism and epistemological skepticism, should come to equate reason with domination and to insist that reason's hegemony be rejected or overthrown.

A powerful alternative to this view appeared in work from the 1970s by the German philosopher Jürgen Habermas. Although agreeing with the French Nietzscheans that traditional metaphysics was obsolete, and in particular that it did not provide a path to absolute truth, Habermas did not reject the notion of truth entirely, nor did he accept the Nietzscheans' call for a "farewell to reason." While acknowledging that the notion of truth is often used to mask unjust power relations and partisan class interests, he insisted that the very possibility of such an insight presupposes that one can conceive of social relations that are just and interests that are held in common by all members of society.

Habermas' theory of communicative action

Habermas' *Theory of Communicative Action* (1985; 1987) was devoted in part to developing an account of truth that did not imply that there existed an "absolute" truth of the kind traditionally posited by metaphysics. Following the doctrines of Pragmatism and reinterpreting Austin's earlier work on speech acts, Habermas contended that ordinary communication differs from other forms of human action in that it is oriented toward mutual agreement rather than "success"—that is, it aims at reaching intersubjective understanding rather than at mastering the world through instrumental action. The process of constructing such an understanding, however, requires that each individual assume that the utterances of the other are for the most part "true," and that the other can provide reasons to support the truth or validity of his utterances should he be called upon to do so. Specifically, individuals must interpret each other's utterances as true assertions about objects and events in an "external world"; as descriptions of morally "right" actions in a social world of shared norms; or as "sincere" expressions of thoughts and feelings in the speaker's "inner world." In this "discourse theory of truth," the notion of truth, far from being a misguided fiction of metaphysics, is a regulative ideal that makes communication possible.

(A.W.L./R.B.W./Ed.)

BIBLIOGRAPHY

General histories: RICHARD H. POPKIN, ed., *The Columbia History of Western Philosophy* (1999), is the first single-volume general history of Western philosophy since BERTRAND RUSSELL, *A History of Western Philosophy* (1945). ROBERT L. ARRINGTON, *A Companion to the Philosophers* (1999), is a collection of essays on particular philosophers and movements. WALLACE I. MATSON, *A New History of Philosophy*, 2nd ed. (2000), is comprehensive and elegantly written. FREDERICK C. COPLESTON, *A History of Philosophy*, 9 vol. (1946–74), remains a classic.

Ancient Greek and Roman philosophy: Detailed histories on the whole course of Greek and Roman philosophy can be found in W.K.C. GUTHRIE, *A History of Greek Philosophy*, 6 vol. (1962–81). For short introductions to Greek philosophy in English, see REX WARNER, *The Greek Philosophers* (1958); and W.K.C. GUTHRIE, *The Greek Philosophers: From Thales to Aristotle* (1950, reissued 1975).

Medieval philosophy: ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955, reissued 1980), is the best account of medieval philosophy. ARMAND A. MAURER, *Medieval Philosophy*, 2nd ed. (1982), sketches medieval philosophy from Augustine to the Renaissance; while NORMAN KRETZMANN, ANTHONY KENNY, and JAN PINBORG (eds.), *The Cambridge History of Later Medieval Philosophy* (1982), covers the period from the rediscovery of Aristotle to the decline of Scholasticism (1100–1600).

Modern and contemporary philosophy. (*Analytic*): P.M.S. HACKER, *Wittgenstein's Place in Twentieth-Century Analytic Philosophy* (1996), focuses on Wittgenstein but is essentially a history Analytic philosophy from Frege through Quine. MICHAEL DUMMETT, *Origins of Analytical Philosophy* (1993, reissued 1996), emphasizes Austrian philosophy. The most general study of the 20th century is AVRUM STROLL, *Twentieth Century Analytic Philosophy* (2000). (*Continental*): VINCENT DESCOMBES, *Modern French Philosophy* (1980; originally published in French, 1979), is indispensable. JÜRGEN HABERMAS, *The Philosophical Discourse of Modernity: Twelve Lectures* (1987, reissued 1990; originally published in German, 1985), is a critique of poststructuralism's "retreat from reason." LUC FERRY and ALAIN RENAUT (eds.), *Why We Are Not Nietzscheans* (1997; originally published in French, 1991), offers a critical appreciation of Nietzsche's reception in France. RICHARD WOLIN, *The Terms of Cultural Criticism: The Frankfurt School, Existentialism, Poststructuralism* (1992), is strong on postwar trends and developments in Continental thought.

(K.v.Fr./A.A.Ma./A.W.L./Av.S./R.B.W./Ed.)

Photography

Photography (from the Greek *photos* ["light"] and *graphos* ["writing"]) is the recording of visible images by light action on light-sensitive materials. The term usually refers to the formation of optical images projected by a lens in a camera onto a film or other material carrying a layer of light-sensitive silver salts and the duplication and reproduction of such images by light action (printing); in an extended sense it also includes the formation of images by certain invisible radiations (ultraviolet and infrared rays) and images recorded in other sensitive materials not containing silver by means of chemical or physical processes or both. Related processes include the recording of images by X rays, electron beams, and nuclear radiations (radiography) and the recording and transmission of light images in the form of electromagnetic signals (television and videotape).

This article treats the historical, technical, and aesthetic aspects of still photography. For a similar treatment of motion-picture photography, or cinematography, see MOTION PICTURES.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 628 and 735, and the *Index*.

The article is divided into the following sections:

The history and art of photography 761
- The pioneers 762
  - Contributions of Niepce and Daguerre
  - Contributions of Wedgwood and Talbot
  - First criticism
- Pre-World War I history 764
  - Development and use of the calotype process
  - Development and use of the collodion process
  - Portraiture
  - Influence of painting
  - Combat photography
  - Landscape photography
  - Stereoscopic photography
  - Development of the dry plate
  - Photography of movement
  - Naturalistic photography
- Post-World War I history 772
  - Influence of abstract art
  - Photomontage
  - Documentary photography
  - Photojournalism
  - Colour photography
  - Later trends

The technology of photography 782
- Cameras and lenses 782
  - Basic camera functions
  - Principal camera types
  - Shutter and diaphragm systems
  - Methods of focusing and framing
  - Exposure-metering systems
  - Flash systems
  - System cameras
  - Characteristics and parameters of lenses
  - Optical performance
  - Special lens types
  - Lens-changing systems
  - Lens coating
- Black-and-white films 787
  - The latent image
  - Sensitometry and speed
  - Colour sensitivity
  - Filters
  - Other film characteristics
  - Film structure and forms
- Picture-taking technique 788
  - Sharpness control
  - Exposure technique
  - Lighting technique
- Black-and-white processing and printing 789
  - Negative development
  - Printing
- Colour photography 790
  - Colour reproduction
  - Colour films
  - Colour-film processing
  - Colour printing
  - Transparency projection
- Instant-picture photography 792
  - History and evolution
  - Black-and-white diffusion transfer
  - The Polacolor process
  - Applications
- Special photosensitive systems 793
  - Electrophotography
  - Colloid and photopolymer processes
  - Diazonium processes
  - Photochromic systems
  - Electronic photography
- Special techniques and applied photography 794
  - High-speed and stroboscopic photography
  - Aerial photography
  - Satellite and space photography
  - Underwater photography
  - Close-range and large-scale photography
  - Stereoscopic and three-dimensional photography
  - Infrared photography
  - Ultraviolet photography
  - Radiography and other radiation recording techniques
  - Nuclear-track recording
  - Astronomical photography
  - Microfilming and microreproduction
- The photography industry 797

Bibliography 797

The history and art of photography

As a means of visual communication and expression, photography has marked aesthetic capabilities. In order to understand them, the characteristics of the process itself must first be understood. Of these the first is immediacy. Usually, but not necessarily, the image that is recorded is formed by a lens in a camera. Upon exposure to the light forming the image, the sensitive material undergoes changes in its structure; a latent image is formed, which becomes visible by development and permanent by fixing. With modern materials, the processing may take place immediately or may be delayed for weeks or months. But, either way, the elements of the final image are determined at the time of exposure. This characteristic is unique to photography and sets it apart from other ways of picture making. Although the photographer can control the character of the original image he captured upon film by the way he develops the negative and prints it, he cannot alter it except by manual interference.

Immediacy of the photograph

A second characteristic of the photograph is that it can contain more than the photographer intended it to. The first daguerreotypes, shown to an astounded public in Paris in the winter of 1838–39 by the inventor Louis-Jacques-Mandé Daguerre, were praised because of the amount of detail recorded by them; looking at one with a magnifying glass, it was said, was like looking at nature with a telescope. The rival inventor of photography, William Henry Fox Talbot, after noting this characteristic, commented:

> It frequently happens, moreover—and this is one of the charms of photography—that the operator himself discovers on examination, perhaps long afterwards, that he has de-

picted many things he had no notion of at the time. Sometimes inscriptions and dates are found upon the buildings, or printed placards most irrelevant, are discovered upon their walls: sometimes a distant dial-plate is seen, and upon it—unconsciously recorded—the hour of the day at which the view was taken.

As technological advances have improved photographic equipment, materials, and techniques, the scope of photography has expanded enormously. High-speed photography has made visible certain aspects of motion never before seen; with material sensitive to invisible radiation, hidden aspects of nature can be revealed; and, by a combination of photographic, electronic, and space technology, even the planets can be observed in new ways. Photography pervades every sphere of activity in modern civilization. Its thousandfold applications have made it indispensable in daily life. Photography disseminates information about humanity and nature, records the visible world, and extends human knowledge into areas the eye cannot penetrate. Next to the printed word the image drawn by light is the most important means of communication, and for this reason photography has been aptly called the most important invention since the printing press.

Authenticity of the photograph

The seemingly automatic recording of an image by photography has given the process a sense of authenticity shared by no other picture-making technique. The fact that the photograph can show more than the eye can see and that the image is not filtered through the brain of a man and put down by the skill of his hand has given it value as evidence. The photograph has become, in the popular mind, so much a substitute for reality and of such apparent accuracy that the adage "The camera does not lie" has become a cliché.

This intrinsic characteristic is of such strength that it has dominated the evaluation of photography's role in the arts. In the past photography was sometimes belittled as a mechanical art because of its dependence on technology. It has also been used over and over again as a foil by art critics to denounce paintings that rely heavily upon exact representation of subject matter. Indeed, after reviewing the daguerreotype process, the painter and art expert Paul Delaroche, who served on the committee that advised the French government to purchase the rights to the new process, declared: "From today painting is dead."

In truth, photography is not the automatic process that is implied by the use of a camera. A fully automatic camera can produce a correctly exposed and sharp negative, but it cannot distinguish between a banal snapshot and a well-composed picture. The ability to make such a distinction rests solely with the person behind the camera. The creative photographer perceives the essential qualities of the subject and interprets it according to his judgment, taste, and involvement. The mechanical photographer merely reproduces what he sees.

Technical controls

Although the camera does limit the photographer to depicting existing objects rather than imaginary or interpretive views, the skilled photographer has at his command a wide variety of controls that can be used to overcome the constraints of literalness and to introduce creativity into the mechanical reproduction process. The image can be modified by different lenses and filters. The type of sensitive material used to record the image is a further control, and the contrast between highlight and shadow can be changed by variations in development. In printing the negative, the photographer has a wide choice in the physical surface of the paper, the tonal contrast, and the image colour.

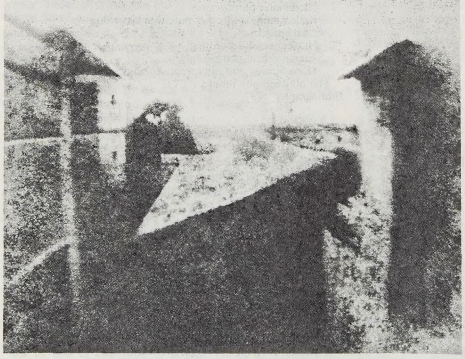
The most important control is, of course, the photographer's vision. He chooses the vantage point and the exact moment of exposure. Through experience he knows how the camera will record what he sees. He learns to previsualize the final print. If he has visual imagination and perception, he can make more than a passive record. He can express universal qualities. He can extend the vision of the viewer.

So facile a medium is photography that it is difficult to grasp its aesthetic capabilities and accomplishments. Of the billions of photographs that are taken every year, only a relatively small number can be considered art. Few camera users are deliberately concerned with the production of photographs to be judged as art. A far greater number look upon photography as a means of communication. While the aim of the commercial photographer, the photojournalist, and the scientist may not primarily be aesthetic, it is significant and remarkably characteristic of the medium that often in their work can be found memorable pictures that reach beyond the particular to the universal. Recognition plays an overwhelming role in photography: recognition by the creative photographer of the picture possibilities presented to him and recognition by the viewer of aesthetic qualities in photographs that he sees.

THE PIONEERS

The forerunner of the camera was the camera obscura, a dark chamber or room with a hole (later a lens) in one wall through which images of objects outside the room were projected on the opposite wall. The principle was probably known to Aristotle more than 2,000 years ago. The Italian scientist and writer Giambattista della Porta, late in the 16th century, demonstrated and described in detail the use of a camera obscura with a lens. By the 18th

By courtesy of the Gernsheim Collection, The University of Texas at Austin



First successful photograph from nature, heliograph on pewter by Nicéphore Niepce, 1826/27.

century artists commonly used various types of camera obscura to trace accurate images from nature. These devices still depended on the artist's drawing skills, however, and the search for a method to reproduce images completely mechanically continued.

In 1727 the German professor of anatomy Johann Heinrich Schulze proved that the darkening of silver salts, known since the 16th century, and possibly earlier, was caused by light and not heat. He demonstrated the fact by using sunlight to record words on the salts, but he made no attempt to preserve the images permanently. His discovery, in combination with the camera obscura, provided the basic technology necessary for photography. It was not until the early 19th century, however, that photography actually came into being, largely through the artistic aspirations of two Frenchmen, Nicéphore Niepce and Louis-Jacques-Mandé Daguerre, and two Englishmen, Thomas Wedgwood and William Henry Fox Talbot.

Contributions of Niepce and Daguerre. Niepce, an amateur inventor living near Chalon-sur-Saône, a city 189 miles southeast of Paris, came to photography through his interest in lithography. In this process drawings were copied by hand onto the lithographic stone. To make the drawings Niepce relied upon his son's artistic skill, but, when his son entered military service, he was left without a draftsman. Not artistically trained, he devised a method by which light drew the pictures he needed. He oiled an engraving to make it transparent, then placed it on a plate coated with a light-sensitive solution and exposed the setup to sunlight. After a few hours the solution under the light areas of the engraving hardened, while that under the dark areas remained soft and could be washed away, leaving a permanent, accurate copy of the engraving. Using a type of asphalt, bitumen of Judea, which changes its solubility in oil of lavender according to its exposure to light, he succeeded from 1822 onward in copying oiled engravings onto lithographic stone, glass, and zinc and from 1826 onto pewter plates. In 1826/27, using a camera obscura fitted with a pewter plate, Niepce produced the first successful photograph from nature, a view of the courtyard of his country estate, Gras, from an upper window of the house. The exposure time was about eight hours, during which the sun moved from east to west so that it appears to shine on both sides of the building. The photograph was rediscovered in 1952 by the historian Helmut Gernsheim and is now preserved in the Gernsheim Collection at the University of Texas.

Niepce's heliographic process

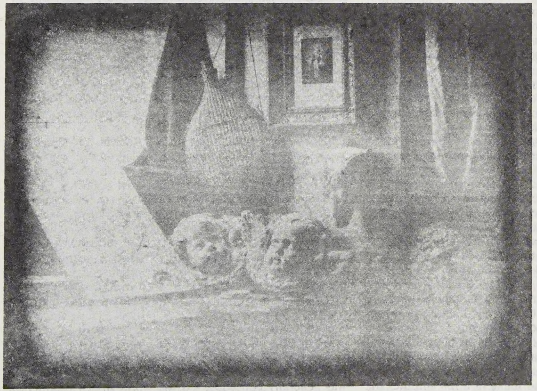
Niepce produced his most successful copy of an engraving, a portrait of Cardinal d'Amboise, in 1826. It was correctly exposed in about three hours, and in February 1827 he had the pewter plate etched to form a printing plate and had two prints pulled. The plate and prints are the oldest photomechanical reproductions still in existence. (The plate and one print are in the Science Museum, London; the other print is in the Gernsheim Collection.) Paper prints were the final aim of Niepce's heliographic (*i.e.*, sun-drawn) process, yet all his other attempts, whether made using a camera or engravings, were underexposed and too weak to be etched. Nevertheless, Niepce's discoveries showed the path that Daguerre and others were to follow with more success.

Daguerre was a professional scene painter. Between 1822 and 1839 he was co-proprietor of the Diorama in Paris, an auditorium in which he and his partner Charles-Marie Bouton displayed immense paintings, 45½ by 71½ feet (14 by 22 metres) in size, of famous places and historical events. The partners painted the scenes on translucent paper or muslin and by the careful use of changing lighting effects were able to present vividly realistic tableaux. The views provided grand entertainment in the illusionistic style, and the amazing trompe l'oeil effect was purposely heightened by the accompaniment of appropriate music and the positioning of real objects, animals, or people in front of the painted scenery.

Like many other artists, Daguerre made his preliminary sketches by tracing the images produced by a camera obscura. About 1826 he began unsuccessful experiments in recording the camera image "by the spontaneous action of light." Learning of Niepce's work, he wrote to him, and on Dec. 14, 1829, the two men formed a partnership for the express purpose of improving Niepce's invention of heliography. From then on Daguerre worked using the improved materials Niepce had adopted—silvered copper plates and iodine—without achieving any improved results until 1835, two years after the death of his partner. By accident Daguerre discovered that a latent image forms on a plate of iodized silver and that it can be "developed" and made visible by exposure to mercury vapour, which settles on the exposed parts of the image. Exposure times could thus be reduced from eight hours to 30 minutes. The results were not permanent, however; when the developed picture was exposed to light, the unexposed areas of silver darkened until the image was no longer visible. By 1837, though, Daguerre was able to fix the image permanently by using a solution of table salt to dissolve the unexposed silver iodide. That year he produced a photograph of his studio on a silvered copper plate, a photograph that was remarkable for its fidelity and detail. Contrary to his contract with Niepce, Daguerre now called the improved process after himself: daguerreotype.

The daguerreotype process

Collection de la Societe Francaise de Photographie, Paris



"Still Life," daguerreotype by Louis-Jacques-Mandé Daguerre, 1837.

In 1839 Daguerre and Niepce's son sold full rights to the daguerreotype and the heliograph to the French government, in return for annuities for life. On August 19 full working details were published. Daguerre wrote a booklet describing the process, *An Historical and Descriptive Account of the Various Processes of the Daguerreotype and the Diorama,* which at once became a best-seller: 29 editions and translations appeared before the end of 1839.

Contributions of Wedgwood and Talbot. In 1802 Thomas Wedgwood, son of the famous potter Josiah Wedgwood, reported his experiments in recording images on paper or leather sensitized with silver nitrate. Although he could record silhouettes of objects placed on the paper, he was not able to make them permanent and, to his disappointment, he failed to record a camera image. Nonetheless, the paper published by Sir Humphry Davy in the *Journal of the Royal Institution,* London, in June 1802, on the experiments of his friend Wedgwood, is the first account of an attempt to produce photographs.

Unaware of the work of Wedgwood and the French pioneers, Talbot, trained as a scientist at Cambridge University, was led to invent a photographic process because of his inability to draw landscapes. On a holiday trip to Italy in 1833, the idea came to him of recording by chemical means the images he observed in his camera obscura. By 1835 he had a workable technique: he made paper light-sensitive by soaking it alternately in solutions of common salt (sodium chloride) and silver nitrate. Silver chloride was thus produced in the fibres of the paper. On exposure to light the silver chloride became finely divided silver, dark in tone. Theoretically, the resulting negative could be used to make any number of positives simply by putting fresh sensitized paper in contact with the negative and exposing it to light. Talbot's method of fixing the print by washing it in a strong solution of sodium chloride was inadequate, however, and the process was not successful until February 1839, when the scientist Sir John F.W. Herschel suggested fixing the negatives with sodium hyposulphite (now called sodium thiosulfate) and waxing them before printing, which reduced the grain of the paper.

Talbot's photogenic drawing process

When news of Daguerre's process reached England in January 1839, Talbot rushed publication of his "photogenic drawing" process and subsequently explained his technique in complete detail to the members of the Royal Society—six months before the French government divulged working directions for the daguerreotype. There were many others who had similar techniques and who were to claim priority, but to Talbot and Daguerre are owed the two basic processes that were to establish photography as the most facile and convincing way to produce pictures.

First criticism. The two pioneer processes were different in several ways. Daguerreotypes were on metal; photogenic drawings were on paper. Each daguerreotype was unique; photogenic drawings could be duplicated. The aesthetic as well as physical character differed markedly. The daguerreotype rendered detail to a degree that was remarkable; the photogenic drawing, because of the fibrous structure of the paper supporting the silver image, gave a broader, somewhat diffused effect.

Differences between the two pioneer processes

The first criticism of photography was necessarily based on a comparison with painting or drawing, since no other standards of picture making existed. Photography's remarkable ability to record a seemingly inexhaustible amount of detail was marveled at again and again. The critics regretted that, because of the great length of exposure, moving objects were not recorded or were rendered blurry and indistinct. The inability of the first processes to record colours was disappointing, but since the critics were already conditioned to black-and-white prints and drawings, this was not as serious a drawback as the harshness of the tonal scale. The technique of photography was at once recognized as a shortcut to art. No longer was it necessary to spend years in art school drawing from sculpture and from life, mastering the laws of linear perspective and chiaroscuro. As Daguerre boasted in a broadsheet in 1838, "with this technique, without any knowledge of chemistry or physics, one will be able to make in a few minutes the most detailed views."

PRE-WORLD WAR I HISTORY

Daguerre's process rapidly spread throughout the world. Before the end of 1839, travelers were bringing back to Paris daguerreotypes of famous monuments in Egypt, Israel, Greece, and Spain; from them engravings were made that were published in two volumes as *Excursions daguerriennes* between 1841 and 1843. Although his process was published "free to the world" by the French government, Daguerre took out a patent for it in England; the first licensee was Antoine-François-Jean Claudet. The first daguerreotypes in America were made on Sept. 16, 1839, just four weeks after the announcement of the process. Exposures were at first of excessive length—a daguerreotype of King's Chapel, Boston, in the International Museum of Photography, Rochester, N.Y., bears a label stating that it was made between 4:40 and 5:30 PM on April 19, 1840. At such exposures moving objects could not be recorded, and portraiture was impractical. Even in blazing sunlight and with the face whitened by flour, a person had to sit immobile for several minutes.

Experiments were started in Europe and the United States to improve the optical, chemical, and practical aspects of the daguerreotype process to make it more feasible for portraiture, the most desired application. Using a camera with a mirror substituted for the lens, Alexander Wolcott opened in New York in March 1840, a "Daguerrean Parlor" for tiny portraits. This was the earliest known photography studio anywhere; the first studio in Europe was opened by Richard Beard in a glasshouse on the roof of the Royal Polytechnic Institution in London on March 23, 1841. Unlike the many daguerreotypists who were originally scientists or miniature painters, Beard had been a coal merchant and patent speculator. Having acquired the exclusive British license for the American mirror camera (he later also purchased the exclusive rights to Daguerre's invention in England, Wales, and the colonies), Beard employed the chemist John Frederick Goddard to try to improve and accelerate the exposure process. Among the techniques Goddard studied were two that Wolcott had tried: increasing the light sensitivity of the silver iodide with bromine vapours and filtering the blindingly bright daylight necessary for exposure through blue glass to ease the portrait sitter's eye strain. By December 1840 Goddard had succeeded well enough to produce tiny portraits ranging in size from one centimetre in diameter to 1.5 by 2.5 inches (four by six centimetres). By the time Beard opened his studio exposure times were said to vary between one and three minutes according to weather and time of day. Daguerreotype portraits were immensely popular, and the studio made considerable profits the first few years, but competition soon appeared and Beard lost his fortune in several lawsuits against infringers of his licenses.

Early daguerreotype studios

The finest daguerreotypes in Britain were produced by Claudet, who opened a studio on the roof of the Royal Adelaide Gallery in June 1841. He was responsible for numerous improvements in photography, for the discovery that red light did not affect sensitive plates and could therefore be used safely in the darkroom, and for the practical introduction of stereoscopic daguerreotypes in 1851.

The most important advances in photographic lens and camera design came from József Petzval and Friedrich Voigtländer, both of Vienna. Petzval produced an achromatic portrait lens that was about 20 times faster than the simple meniscus lens the Parisian opticians Charles Chevalier and N.M.P. Lerebours had made for Daguerre's cameras. Voigtländer reduced Daguerre's clumsy wooden box to easily transportable proportions for the traveler. These valuable improvements were introduced by Voigtländer in January 1841. That same month another Viennese, Franz Kratochwila, freely published a chemical acceleration process in which the combined vapours of chlorine and bromine increased the sensitivity of the plate five times.

The improvements that had been made in lenses and sensitizing techniques reduced exposure times to approximately 20 to 40 seconds. Daguerreotyping became a flourishing industry, especially in the United States, which, it was generally conceded, led the world in the production of daguerreotypes. In the late 1840s every city had its

Early daguerreotypes in the United States

"daguerrean artist," and villages and towns were served by traveling photographers who had fitted up wagons as studios. In New York City alone there were 77 galleries in 1850. Of these, the most celebrated was that of Mathew B. Brady, who began in 1844 to form a "Gallery of Illustrious Americans," and to that end collected portraits of notables taken by his own and other cameramen. Twelve of the portraits were published by lithography in a folio volume. In Boston a studio operated by Albert Sands Southworth and Josiah Johnson Hawes that was advertised as "The Artists' Daguerreotype Rooms" produced the finest portraits ever made by the daguerreotype process. The partners avoided the stereotyped lighting and posing formulas of the average daguerreotypist and did not hesitate to portray their sitters unprettified and "as they were." Lemuel Shaw, a judge of the Supreme Court of Massachusetts, stands with crumpled coat and unruly locks of hair under a glare of sunshine; Lola Montez, adventuress, dancer, actress, lolls over the back of a chair, a cigarette between her gloved fingers. Cities and towns, as well as their inhabitants, were photographed by American daguerreotypists: the rapid growth of San Francisco was documented month by month, and the first history of the city, published in 1855, was illustrated by engravings made from daguerreotypes.

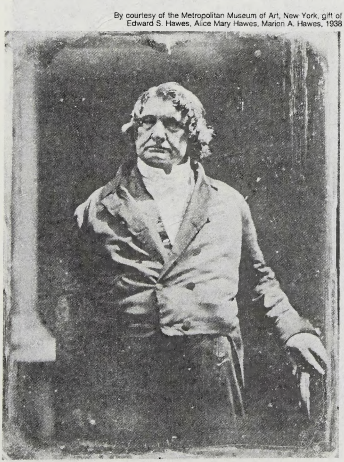
Development and use of the calotype process. The popularity of the daguerreotype surpassed that of the photogenic drawing, but Talbot continued work to improve his process. On Sept. 21–23, 1840, while experimenting with gallic acid, a chemical he was informed would increase the sensitivity of his prepared paper, Talbot discovered that the acid could be used to develop a latent image. This procedure revolutionized photography on paper as it had photography on metal in 1835. Whereas previously Talbot had needed a camera exposure of one hour to produce a 6.5- by 8.5-inch negative, he now found that one minute was sufficient. Developing the latent image had put photography on paper on a par with the daguerreotype. Talbot named his improved negative process the calotype, from the Greek meaning "beautiful picture," and protected his rights by patent.

The first and most aesthetically satisfying use made of this improved process was in the work of David Octavius Hill, a Scottish landscape painter, and his partner, Robert Adamson, an Edinburgh photographer. In 1843 Hill decided to paint a group portrait of the ministers who in that year formed the Free Church of Scotland. There were more than 400 figures to be painted. Sir David Brewster, who knew of Talbot's process from the inventor himself, suggested to Hill that he make use of this new technique. Hill then enlisted the aid of Adamson, and together they made hundreds of photographs, not only of the members of the church meeting but also of people from all walks of life. Although their sitters were posed outdoors in glaring sunlight and had to endure exposures of upward of a minute, Hill and Adamson managed to retain spontaneity. Hill's vision was dominated by the painting style of the period in lighting and posing, particularly in the placement of the hands. Many of the calotypes are strikingly reminiscent of canvases by Sir Henry Raeburn and other contemporary artists. Indeed, William Etty, a Royal Academician, copied in oils the calotype Hill and Adamson made of him in 1844 and exhibited it as a self-portrait. In addition to their formal portraiture, the partners made a series of photographs of fishermen and their wives at Newhaven, in Edinburgh, and architectural studies.



"John Henning and Alexander Handyside Ritchie," calotype by David Octavius Hill and Robert Adamson, *c.* 1845.
By courtesy of the Art Institute of Chicago

By courtesy of the Metropolitan Museum of Art, New York, gift of Edward S. Hawes, Alice Mary Hawes, Marion A. Hawes, 1938



"Chief Justice Lemuel Shaw," daguerreotype by Albert Sands Southworth and Josiah Johnson Hawes, 1851.

Calotypes of architectural monuments

The potential of the calotype for recording great monuments of architecture was shown by a number of Frenchmen, many of whom were trained as painters. In the 1850s they began to photograph historical buildings for the government. Working with cameras making photographs as large as 20 by 29 inches, Henri Le Secq, Charles Marville, and Charles Nègre produced remarkable calotypes of the cathedrals of Notre-Dame in Paris, Chartres, and Amiens, as well as other structures that were being restored after centuries of neglect. An establishment was set up in Lille, Fr., by Louis-Désiré Blanquart-Evrard for bulk printing these paper negatives. Among the products of this firm was a superb volume of photographs by the Parisian writer Maxime Du Camp taken during his travels with the writer Gustave Flaubert in Egypt, Palestine, and Syria, from 1849 to 1851.

Development and use of the collodion process. Photography was revolutionized in 1851 by the introduction of the collodion process for making glass negatives. This new technique, invented by the English sculptor Frederick Scott Archer, was 20 times faster than all previous methods and was, moreover, free from patent restrictions. The

glass plate negatives recorded detail in a way that rivaled the daguerreotype, and from them paper prints could be made. The process had one serious drawback: the photographer had to sensitize his plate almost immediately before exposure and expose it and process it while the coating was moist. Collodion is a solution of nitrocellulose (guncotton) in alcohol and ether; when the solvents evaporate, a clear plasticlike film is formed. Since it is then impervious to water, the chemicals used for developing the exposed silver halides and removing the unexposed salts cannot penetrate to them. Despite the drawback that the photographer had to have a complete darkroom outfit with him always, the collodion process was almost at once universally adopted. It reigned supreme for more than 30 years and greatly increased the popularity of photography. Some of the most remarkable photographs of all time were produced by this wet-plate process.

At first the positive prints made from the glass plate negatives were produced by Talbot's salt paper method, but from the mid-1850s on they were made on albumen paper, a slow printing-out paper (*i.e.,* paper that produces a visible image on direct exposure, without chemical development) that had been coated with egg white before being sensitized. The egg white gave the paper a glossy surface that improved the picture. Albumen paper was introduced in 1850 by Blanquart-Evrard and remained in general use until World War I.

The new collodion process was also used to produce imitation daguerreotypes called positives on glass or ambrotypes. They were simply underexposed or bleached negatives that appeared positive with a dark coating or backing. In posing and lighting, these popular portraits were identical to daguerreotypes; they were of the same standard sizes, and they were enclosed in the same type of case. They did not approach the brilliancy of the daguerreotype, however. Tintypes, first known as ferrotypes or melainotypes, were cheap variations of the ambrotype. Instead of glass the collodion emulsion was coated on thin iron sheets enameled black. At first they were presented in cases, surrounded by narrow gilt frames, but by the 1860s this elaborate presentation had been abandoned, and the metal sheets were simply inserted in paper envelopes, each with a cutout window the size of the image. Easy to make, inexpensive to purchase, tintypes remained a kind of folk art through the 19th century. Poses were often informal, if not humorous.

Development of the tintype

Portraiture. A new style of portrait, introduced in Paris by André-Adolphe-Eugène Disdéri in 1854, was universally popular from 1859 onward. It came to be called the carte de visite because the size of the mounted photograph (four by 2½ inches) corresponded to that of a calling card. Disdéri used a four-lens camera to produce eight negatives on a single glass plate. Each picture could be separately posed, or several exposures could be made at once. The principal advantage of the system was its economy: to make eight portraits the photographer needed to sensitize only a single sheet of glass and make one print, which he then cut up into separate pictures. At first cartes de visite almost invariably showed the subjects standing. Backgrounds, which had usually been plain in the days of the daguerreotype, became ornate: furniture and such architectural fragments as papier-mâché columns and arches were introduced, and heavy-fringed velvet drapes were hung within range of the camera. With the advent of the cabinet-size (6½ by four inches) picture in 1866, the baroque tendencies of the photographer became yet more audacious, so that in 1871 a photographer wrote: "One good, plain background, disrobed of castles, piazzas, columns, curtains and what not, well worked, will suit every condition of life." It was at this period that retouching, the use of handwork on the negative, was introduced, as was the practice of painting over the photograph in oil colours.

Contributions of Nadar, Carjat, and Cameron

In contrast to the excessive reliance on accessories and retouching shown by the popular portrait photographers of Europe and America, the work of two Frenchmen and one Englishwoman stands apart. In their portraiture they reached a level unsurpassed since the daguerreotypes of Southworth and Hawes and the calotypes of Hill and Adamson. These photographers were Gaspard-Félix Tournachon, a Parisian writer, editor, and caricaturist who used the pseudonym of Nadar; Étienne Carjat, likewise a Parisian caricaturist; and Julia Margaret Cameron, wife of an eminent British jurist.

Nadar took up photography in 1853 as a means of making studies of the features of prominent Frenchmen for inclusion in a large caricature lithograph, the "Panthéon Nadar." He posed his sitters against plain backgrounds and bathed them with diffused daylight, which brought out every detail of face and dress. He knew most of them, and the powers of observation he had developed as a caricaturist led him to recognize their salient features, which he recorded directly, without the exaggeration that he put in his drawings. When Nadar's photographs were first exhibited, they won great praise in the *Gazette des Beaux Arts,* then the leading art magazine in France. Nadar was a colourful man who had a passion for balloons. He combined his interests by taking a series of aerial photographs, which inspired the French artist Honoré Daumier to produce a cartoon bearing the mocking title "Nadar Élevant la Photographie à la Hauteur de l'Art" ("Nadar Raising Photography to the Height of Art").

Carjat's portraits, more intense perhaps than Nadar's, have the dignity and distinction of those of his contemporary and rival.

Cameron took up photography as a pastime in 1864. Awkward as it was, she used the wet-plate process and began to take portraits of such celebrated Victorians of her acquaintance as Alfred Tennyson, George Frederick Watts, Thomas Carlyle, Charles Darwin, and Sir John F.W. Herschel. A number of her portraits were shown at the Paris International Exhibition of 1867. Cameron used a lens of the extreme focal length of 30 inches to obtain large close-ups. This lens required such long exposures that the subjects frequently moved. The lack of optical definition plus this accidental blurring was universally criticized, yet the very power of her work won her international praise. This can only be explained by the intensity of her vision. "When I have had these men before my camera," she wrote about her portraits of great men,

> my whole soul has endeavoured to do its duty toward them in recording faithfully the greatness of the inner man as well as

George Eastman House Collection



"The Mountain Nymph, Sweet Liberty" by Julia Margaret Cameron, 1867.

the features of the outer man. The photograph thus obtained has almost been the embodiment of a prayer.

Besides these memorable portraits, Cameron produced a large number of allegorical studies, of children and young women in costume, acting out biblical scenes or themes based on the poetry of her hero, Tennyson. In making these pictures—which today seem weak and sentimental—she was much influenced by the Pre-Raphaelite painters and the photographic work of Oscar G. Rejlander and Henry Peach Robinson (see below), leaders in the production of photographs that emulated paintings.

Influence of painting. Consideration of photography as an aesthetic medium was given impetus by the formation of photographic societies, made up of both professionals and amateurs, who had been attracted to the camera by the popularity of the collodion process. In 1853 the Photographic Society, parent of the present Royal Photographic Society, was formed in London, and in the following year the Société Française de Photographie was founded in Paris.

At the first meeting of the Photographic Society the president, Sir Charles Eastlake (who was then also president of the Royal Academy), invited the miniature painter Sir William Newton to read a paper "Upon Photography in an Artistic View" (*Journal of the Photographic Society*, i, 1853). His argument was that photographs could be useful to the painter so long as they were taken "in accordance [as far as it is possible] with the acknowledged principles of Fine Art." One way by which the photographer could make his results more like works of art, Newton suggested, was to throw the subject slightly out of focus. He also recommended liberal retouching.

Combination printing

An outcome of the urge to create photographs that would fit a priori concepts of what "art" should be was the practice of combining several negatives to make one print in order to achieve painterly compositions of subjects too complicated to be photographed in a straightforward manner. A famous example was by Oscar G. Rejlander, a Swede who had studied art in Rome and was practicing photography in England. He used 30 negatives to produce a 31- by 16-inch print titled "The Two Ways of Life," showing, in allegory as obvious as it was sentimental, that the way of the blessed led through good works and the way of the damned through vice. Rejlander, who described the technique in detail in photographic journals, stated that his purpose was to prove to artists the aesthetic possibilities of photography, which they had generally denied. The photograph was shown in the Manchester Art Treasures Exhibition of 1857 and was purchased by Queen Victoria for Prince Albert.

Rejlander's technique stimulated Henry Peach Robinson, a professional photographer who had been trained as an artist, to produce similar combination prints. He achieved fame with a five-negative print, "Fading Away," produced in 1858. The subject, a dying girl, was considered by critics as too painful a subject to be represented by photography. Perhaps the implied authenticity of the camera bothered them, for painters had long presented subjects of a far more painful nature.

Robinson became a most articulate member of the Photographic Society, and his teaching was even more influential than his photography. In 1869 appeared the first of many editions and translations of his book, *Pictorial Effect in Photography*. From an outmoded handbook on painting, Robinson borrowed compositional formulas the use of which, he claimed, would bring artistic success. The importance of balance was stressed, and the opposition of light against dark was made clear. The fault of the book lay not only in the assumption that rules set up for one art form could be applied to another but also in its intellectual and academic approach to art.

Robinson's rules for photography

Robinson's work is weak and artificial by present standards of taste. Not only did he practice combination printing when it was not technically necessary, but he preferred to work in the studio, against painted backdrops and with props of natural objects, even foliage, mounted on casters. When he did photograph the real world, he took models with him, dressing them up to play the part of country girls.

So long as photographers maintained that the way to art was by the emulation of painting, critics were reluctant to admit the new medium to an independent aesthetic position. Portraits, when done as sensitively and as directly as those produced by Hill and Adamson, Nadar, and Cameron, won the praise of art critics. But sentimental genre scenes, posed and arranged for the camera, lacked the sharp objective truth that is a characteristic of photography. Other photographers, not concerned about producing art for exhibition, were making photographs of the world and man's activities with such extraordinary perception and understanding of the medium that often their work surpassed more consciously artistic works. These men took their cameras to battlefields and to far-away places, often at the risk of their lives.

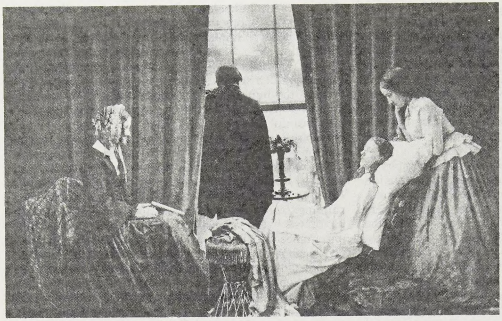
Contributions of Fenton and Brady

Combat photography. In 1855 Roger Fenton sailed from London to the Crimea to photograph the war. He was sent to provide visual evidence countering the caustic written reports dispatched by William Russell, war correspondent for *The Times* of London, criticizing military mismanagement and the inadequate, unsanitary living

George Eastman House Collection



"The Two Ways of Life," a composite photograph made from 30 negatives taken and assembled by Oscar G. Rejlander, 1857. The tableau is allegorical, showing a young man torn between a life of (left) debauchery and (right) virtuous labour.



"Fading Away," a composite photograph made from five negatives by Henry Peach Robinson, 1858.
George Eastman House Collection

conditions of the soldiers. Fenton had to develop his wet plates in a horse-drawn van that had been converted into a darkroom. It was visible for miles in the bare landscape and a few times attracted enemy fire. Despite the difficulty, during his four-month stay Fenton produced 360 photographs, the first large-scale camera documentation of a war.

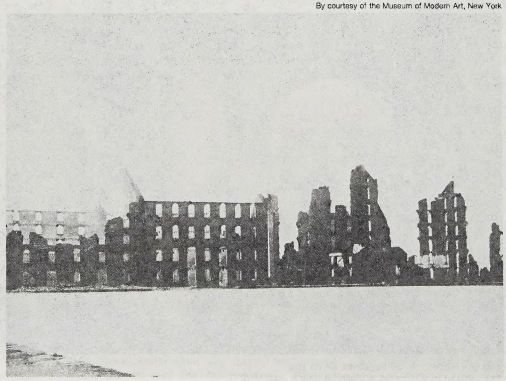
When the Civil War broke out in the United States, Mathew Brady, the New York daguerreotypist and portraitist, who had been among the first to adopt the wet-plate process, conceived the bold plan of making a photographic record of the hostilities. When President Lincoln told him the government could not finance such an undertaking, he invested his own savings in the project, expecting to recover his outlay by selling thousands of prints. Brady and his photographers—notably Alexander Gardner and Timothy H. O'Sullivan, who left his employ in the midst of hostilities—produced an amazing record of the battlefield. At his New York gallery, Brady showed pictures of the dead at Antietam. The *New York Times* reported on Oct. 20, 1862:

> Mr. Brady has done something to bring home to us the terrible reality and earnestness of war. If he has not brought bodies and laid them on our dooryards and along the streets, he has done something very like it. . . . It seems somewhat singular that the same sun that looked down on the faces of the slain blistering them, blotting out from the bodies all semblance to humanity, and hastening corruption, should have thus caught their features upon canvas, and given them perpetuity for ever. But it is so.

Long prized for their value as historical documents, the Civil War photographs are now valued for their aesthetic qualities as well. Unfortunately for Brady, immediately after the war they were seen as unnecessary reminders of hardship and conflict. Unable to sell the prints as he had planned, Brady died embittered in a charity hospital in New York City. Fenton's Crimean War photographs had similarly lost their audience as soon as the peace treaty was signed. Nevertheless, entrepreneurs hoping to sell prints or commemorative albums continued to finance the photographic documentation of the more important conflicts of the late 19th century. The Boer War and the Russo-Japanese War were also covered by photographers engaged by newspapers and by three American mass producers of stereographs.

Landscape photography. During the collodion period scores of photographers journeyed to the far corners of the world, producing memorable travel views despite the trying conditions of the wet-plate process. Among the most successful was the Englishman Francis Frith. The

By courtesy of the Museum of Modern Art, New York



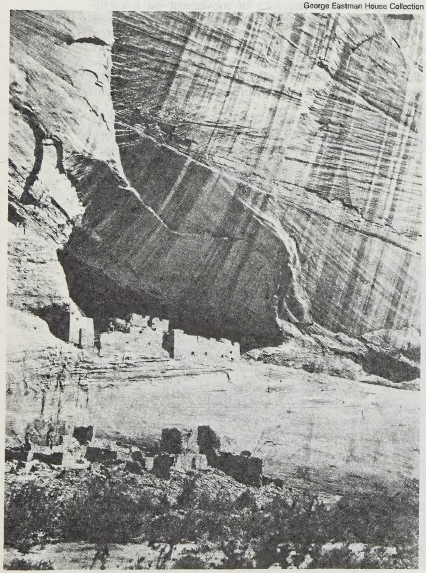
"Ruins of the Gallego Flour Mills, Richmond, Virginia" by Mathew B. Brady, 1865.

most active of several European photographers working in the Middle East in the late 1850s, he took hundreds of fine pictures of monuments along the Nile from Cairo to Abu Simbel, as well as in Syria and Palestine. Samuel Bourne, Felice Beato, John Thomson, and other British amateurs traveled to Asia, bringing back to England lively images of the nature, people, and customs of India, China, and Japan. Other British photographers concentrated on Europe: Charles Clifford recorded the landscape and architecture of Spain, Robert MacPherson that of Rome, and Thomas Annan of Glasgow and George Washington Wilson of Aberdeen the wildness, castles, and abbeys of Scotland. The Bisson brothers (Louis-Auguste and Auguste-Rosalie), Gustave Le Gray, and Edouard-Denis Baldus depicted the landscape and architecture of France. In the United States Carleton E. Watkins and the English-born Eadweard Muybridge both won recognition for their scenic views of Yosemite, the Columbia River, Alaska, and other wilderness regions of North America.

Landscape photography was usually intended for publication in books or as portfolios of prints to be sold to collectors, but in the United States photographers were often important members of government surveys and were also commissioned by railroad companies to make publicity pictures of track laying, bridge building, and spectacular scenery through which the new lines ran. Of the photographers of the American frontier, two stand out: Timothy H. O'Sullivan, of Civil War fame, and William Henry Jackson. O'Sullivan's photographs of the Southwest are of great beauty, particularly his views of Indian cliff dwellings in the Canyon de Chelly in Arizona, made in 1873. Jackson, self-trained as a painter in Vermont, crossed the plains as a wagon driver. In 1868 he opened a photographic gallery in Omaha, Neb. The Union Pacific Railroad was under construction, and he received an order to produce 10,000 stereographs. The excellence of his work led F.V. Hayden, a geologist, to hire him to photograph the Yellowstone as part of Hayden's government-financed expedition there in 1871. The photographs Jackson took were influential in the decision by Congress to create Yellowstone National Park. Later, in 1875, he recorded the immensity of the western landscape, using large glass plates.

Photography on the American frontier

George Eastman House Collection



"Canyon de Chelly, Arizona" by Timothy H. O'Sullivan, 1873.

Stereoscopic photography. Many of the landscape photographers also took stereographs. These double pictures, taken after 1856 with twin-lens cameras, produce a remarkable effect of three dimensions when viewed through a stereoscope. Stereography, first described in 1832 by the English physicist Charles Wheatstone, is uniquely photographic, since no artist could draw two scenes in exact perspective from viewpoints separated only 2½ inches—the normal distance between human eyes. Wheatstone's mirror stereoscope, however, was not practical for use with photographs, and the invention languished until the Scottish scientist Sir David Brewster designed a simplified viewing instrument, which was exhibited at the 1851 Great Exhibition in the Crystal Palace, London. Queen Victoria was entranced by the stereo daguerreotypes she saw there, and with the introduction of the collodion process, which simplified exposure and printing techniques, three-dimensional photography became a popular craze.

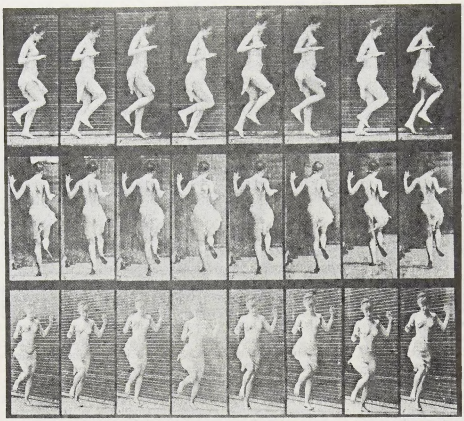
In 1854 the London Stereoscopic Company was formed. Their chief photographer was William England, whose lively street scenes of New York City in rainy weather and views of Niagara Falls taken in 1859 were the wonders of the day. The instantaneous street scenes, which showed pedestrians and vehicles stopped in their tracks, were made possible because the small size of the stereo-camera reduced exposure times to less than half a second. To minimize movement street views were usually taken from a first-floor window with the camera focused directly down the street. (Such views later inspired several Impressionists to paint similar street scenes.) Between 1860 and about 1920 a stereo viewer was as ubiquitous in British and American homes (where a simplified and cheap hand viewer was introduced by Oliver Wendell Holmes [the American physician was a great lover of photography]) as the television set is today. Millions of stereographs were circulated in the years before newspaper reproduction of photographs, and their impact was enormous.

Development of the dry plate. In the 1870s many attempts were made to find a dry substitute for wet collodion so that plates could be prepared well in advance and developed long after exposure. The suggestion casually made in 1871 by Richard Leach Maddox, an English physician, to suspend silver bromide in a gelatin emulsion led, in 1878, to the introduction of factory-produced dry plates coated with gelatin containing silver salts, an event that marked the beginning of the modern era of photography.

Gelatin plates were about 60 times more sensitive than collodion plates. The increased speed freed the camera from the tripod, and a great variety of small hand cameras that allowed photographers to take instantaneous snapshots became available at relatively low cost. Of these, the most popular was the Kodak camera, introduced by George Eastman in 1888. Its simplicity greatly speeded the growth of amateur photography. In place of glass plates, it contained a roll of negative material sufficient for taking 100 circular pictures, each roughly 2½ inches in diameter. After exposing the last negative, the entire camera was sent to one of the Eastman factories (Rochester, N.Y., or Harrow, Middlesex), where the roll was processed and printed. "You Press the Button, We Do the Rest" was Eastman's description of the Kodak system. At first Eastman's so-called "American film" was used in the camera. This film was paper based, and the gelatin layer containing the image was stripped away after development and fixing and transferred to a transparent support. In 1889 it was replaced by film on a transparent plastic base of nitrocellulose that had been developed by the Reverend Hannibal Goodwin of Newark, N.J., in 1887.

Invention of the Kodak camera

Photography of movement. A few years before the introduction of the dry plate, the world was amazed at the photographs of horses taken by Eadweard Muybridge in California. Using a series of 12 to 24 cameras ranged side by side opposite a reflecting screen, with their shutters released by the breaking of threads as the horse dashed by, Muybridge secured sets of sequence photographs of succes-



"Figure Hopping" by Eadweard Muybridge, 1887. Sequence of eight stages of movement, simultaneously photographed by multiple cameras at three different positions.
By courtesy of the Cooper-Hewitt Museum of Decorative Arts and Design, Smithsonian Institution

sive phases of the walk, the trot, and the gallop. When the pictures were published internationally in the popular and scientific press, they were so different from the traditional hand-drawn representation of a horse's steps that it was difficult to believe that they were accurate. To prove that his photographs were correct, Muybridge threw them upon a screen one after the other with a lantern-slide projector he had built for the purpose; the result was the world's first motion-picture presentation. This memorable event took place at the San Francisco Art Association in 1880.

Muybridge's early studies were taken with wet plates. With the new gelatin plates, he was able to improve his technique greatly, and in 1884–85, at the invitation of the University of Pennsylvania, he produced 781 sequence photographs of many kinds of animals as well as men and women engaged in a wide variety of activities.

Muybridge's photographic analysis of movement led the French physiologist Étienne-Jules Marey to develop chronophotography. Whereas Muybridge had employed a battery of cameras to record detailed, separate images of successive stages of movement, Marey used only one, recording an entire sequence of movement on a single plate. With Marey's method, the images of various phases of motion sometimes overlapped, but it was easier to see and understand the flow of movement. Marey was also able to record higher speeds at shorter intervals than Muybridge. Both his and Muybridge's work greatly contributed to the field of motion study and to the development of the motion picture.

Naturalistic photography. *Contributions of Emerson.* In the late 19th century the growing number of amateur photographers used the camera to capture daily occurrences

George Eastman House Collection



"Pond in Winter" by Peter Henry Emerson, 1888.

and important moments in their lives, but the members of the societies and clubs concerned with photography as an art became more and more divorced from matters of ordinary life. Subjects in the so-called art photographs were artificially composed in the studio in imitation of 17th-century Dutch paintings. Photographers strove to master complicated printing methods allowing manual interference. Photographing everyday life was considered mere record-making or documentation. Landscape pictures, the strength of British photographers in past decades, found little favour. When similar beliefs prevailed in French academic painting 35 years previously, the French realist painter Gustave Courbet was prompted to call for a "return to nature." So now Peter Henry Emerson, physician by profession and an ardent amateur photographer, attacked the artificiality of the photographs generally accepted as outstanding examples of the artistic use of the camera. Emerson's passionate plea for the return to natural subjects was indeed salutary, but of greater importance was his advice that photographers should respect the photographic process and limit their controls to those that were inherent.

Emerson's photographs of natural subjects

In his book *Naturalistic Photography* (1889) Emerson further developed his theories (some of which he later disclaimed). Although his writings were influential, his photographs of the life of simple country folk presented a far more convincing argument for his beliefs. Emerson's photographs were far removed from the usual artificial genre studies and close to the graphic work of the French painter J.-F. Millet, which Emerson greatly admired. They were published in limited editions in handsome folio volumes and motivated other amateurs to seek inspiration in nature.

The photographs in Emerson's first and finest album, *Life and Landscape on the Norfolk Broads* (1886), were printed on the newly invented platinotype paper. In this printing paper, salts of iron and platinum replaced those of silver as the light-sensitive material. Platinotypes had a long and delicate tonal scale, and they did not fade, unlike the more common silver prints. Emerson helped to popularize the paper, which remained in use until about 1920, when the rising price of platinum made it impractical.

The Linked Ring and the Photo-Secession. The recognition of photography as an art rather than a mechanical process and its evaluation on its own terms rather than according to the traditional rules governing painting were further advanced by the formation in London in 1892 of the Brotherhood of the Linked Ring. The group, which was founded by the prominent pictorial photographer H.P. Robinson, George Davison, a leader of the Art Nouveau movement, and others dissatisfied with the scientific bias of the London Photographic Society, held annual exhibitions, which they called salons. By 1901 it was their proud boast that "through the *Salon* the Linked Ring has clearly demonstrated that pictorial photography is able to stand alone and that it has a future entirely apart from that which is purely mechanical."

Similar groups formed in other countries. One of the most influential was the Photo-Secession, founded in the United States in 1902 by Alfred Stieglitz. Stieglitz, who had previously organized the Camera Club of New York City and served as editor of the club journal, *Camera Notes,* was a strong proponent of "straight" photography. He did not believe in retouching or manipulating in any way his negatives or prints. He had, as early as 1892–93, demonstrated the pictorial possibilities of the hand camera with his photographs of New York under all weather conditions.

Stieglitz's contributions to photography

A few of the Photo-Secession members, including Clarence H. White and Harry C. Rubincam, favoured naturalistic photography like Stieglitz. Many, however, notably Edward Steichen and Alvin Langdon Coburn, were adherents of the impressionistic soft-focus school and of the newly introduced gum print process. This technique gave the photographer the utmost manual control. He coated paper with watercolour pigment of any desired tint mixed with gum arabic and potassium bichromate. On exposure to light beneath a negative, the pigment became insoluble according to the amount of light received. The print was "developed" simply by bathing it in water. If desired, areas could be eliminated by brushing them with hot water or by drawing on them.

Steichen's gum prints

Despite their stylistic differences, the members of the Photo-Secession were united in their disdain for the lack of standards and the general conduct of photographic exhibitions in the United States. They chose the name "secession" to dramatize their rebellion against the status quo, just as avant-garde German and Austrian painters had used the same word to make manifest their independence of officialdom. The record of the Photo-Secession

Collection, The Museum of Modern Art, New York City; gift of William A. Grigsby



"Steeplechase Day, Paris: Grand Stand," duogravure by Edward Steichen, 1905; from *Camera Work,* 1913. 15.72 cm × 20.16 cm.



"The Steerage" by Alfred Stieglitz, 1907.
By courtesy of the Art Institute of Chicago, Alfred Stieglitz Collection

is contained in 50 issues of the much praised *Camera Work,* published by Stieglitz between 1903 and 1917; this quarterly publication contained superb reproductions of photographs but was of uneven quality, partly because of the members' stylistic differences. The photographs were occasionally overly sentimental, artificial, or banal, and indeed, far better work was being produced in New York City at this time by documentary photographers such as Jacob Riis and Lewis W. Hine, whose photographs were not then considered art.

In addition to *Camera Work,* the Photo-Secession had a gallery, which came to be known as 291 from the street number on Fifth Avenue, New York City, where it was located. There Stieglitz showed not only pictorial photographs but also, from 1906 on, avant-garde modern art, selected at first by Steichen in Paris. At 291 Americans saw, long before the Armory Show of 1913 made them popular, paintings and sculpture by Rodin, Marin, Toulouse-Lautrec, Cézanne, Van Gogh, Picasso, Brancusi, and Matisse. Stieglitz also organized there the first exhibitions in the world of drawings by children (1912) and of African art (1914).

Other important photographers of the period were Paul Strand and Coburn. The last two issues of *Camera Work* contained photographs by Strand only. They showed an entirely new approach. There were views looking down from unusual angles; there were bowls in quasi-abstract arrangements; there was a group of powerful open-air portraits taken in the streets of the lower East Side of New York City. These portraits had been taken with a 45° prism fitted over the lens, so that the subjects were unaware that they were being photographed. Coburn, one of the first of the Photo-Secessionists, made a series of photographs in 1912 looking down from tall buildings, which he exhibited as "New York from Its Pinnacles"; they were remarkable for the way in which emphasis was placed on form. He pushed this interest in abstraction to the total elimination of recognizable subject matter in his "Vortographs," some of which were published in *Photograms of the Year* in 1917.

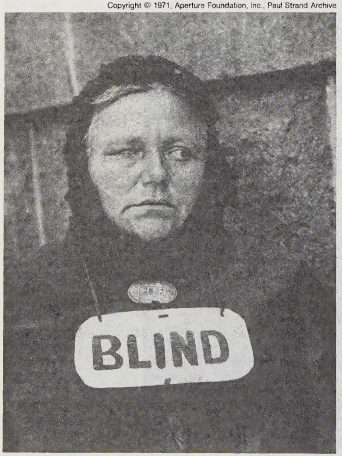
POST-WORLD WAR I HISTORY

During the 1920s and '30s, a new, more realistic style of photography gained prominence, a reflection, perhaps, of post-World War I disillusionment. At the close of the war, Steichen, who had been in command of aerial photography for the American Expeditionary Forces, abandoned the broad, impressionistic style he had earlier practiced for sharply focused portraits of celebrities, taken for *Vanity Fair* and *Vogue* magazines. The painter Charles Sheeler used the camera to record the stark beauty of Pennsylvania barns and the forms of industrial structures.

Realistic photography in the United States

In California, Edward Weston, who had been working as a photographer for more than 15 years, turned in 1923 to a more direct and realistic use of the camera, with a certain emphasis upon abstract form that he never allowed to detract from the recognition of subject matter. "The camera should be used for a recording of *life,*" he wrote in 1924, "for rendering the very substance and quintessence of the *thing itself,* whether it be polished steel or palpitating flesh." He formulated a working method based on the use of the classic view camera for eight- by 10-inch negatives, a high-quality lens, well stopped down to ensure great depth of field, and the complete avoidance of artificial lighting. He printed the negatives by contact on glossy bromide paper to secure maximum detail. Enlarging and retouching of any kind was inadmissable.

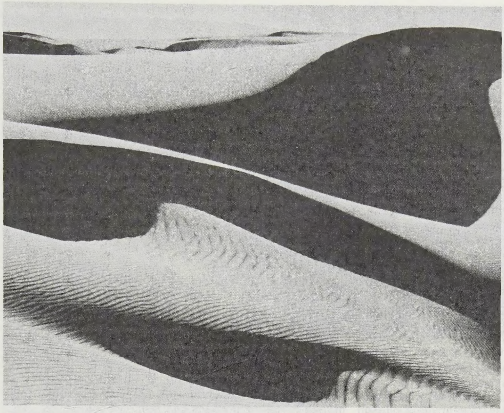
Copyright © 1971, Aperture Foundation, Inc., Paul Strand Archive



"Blind Woman, New York" by Paul Strand, 1916; from *Camera Work,* 1917.

To Stieglitz this revaluation of photographic aesthetics was in fact a strengthening of the beliefs he had held since his student days. He produced some of his most powerful work in the 1920s—especially his photographs of cloud forms that he called "Equivalents" because they were equivalent to his thoughts and emotions. Ill health forced him to abandon the camera in 1936, but he continued to maintain an art gallery until his death in 1946.

Paul Strand, whose striking close-ups and semiabstract photographs Stieglitz had first exhibited, developed a style equally rigorous and self-disciplined during the same period. He produced powerful landscapes, direct portraits, and minute details of driftwood and plant life. In 1940 a portfolio of superb photographs Strand had taken in Mexico was issued. After Strand settled in France in 1950, he traveled extensively, producing similar publications on France, Italy, the Hebrides, Egypt, and Ghana, each providing insight into the life in small communities. Asked to define his sphere of interest, Strand once replied that he considered himself a photographer of people.



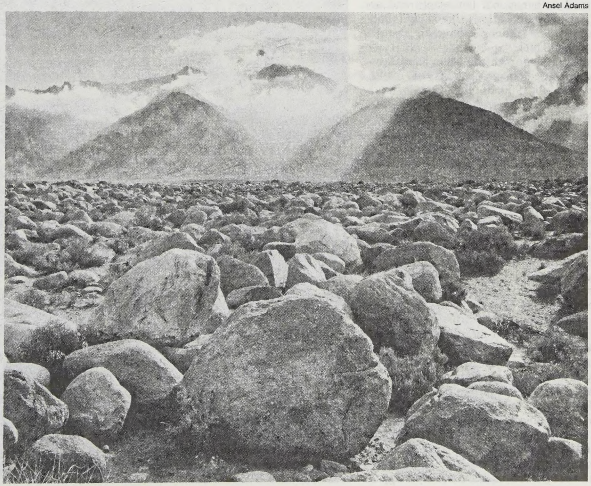
"Dunes, Oceano" by Edward Weston, 1936.
Edward Weston

Viewing the negatives of Strand led Ansel Adams in 1930 to make photography his career. He was then studying the piano and photographing for his own satisfaction. Long interested in nature—his first photographs were of the Sierra Nevada Range and Yosemite National Park—he refined and sharpened his technique. In 1932, with Willard Van Dyke, Edward and Brett Weston, Imogen Cunningham, and Sonya Noskowiak, Adams founded an informal society, Group f.64, so named for the smallest setting of the aperture of the lens that coupled maximum depth of field with maximum sharpness. Adams' great contribution was in what he called "the interpretation of the natural scene." His photographs "Mount Williamson—Clearing Storm" (1944) and "Sierra Nevada from Lone Pine" (1944) are classics.

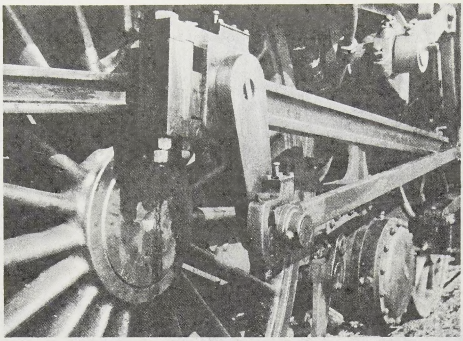
New Objectivity in Germany

A major representative of the postwar realistic style, known as the Neue Sachlichkeit ("New Objectivity") in Germany, was Albert Renger-Patzsch, a professional photographer who spent most of his life in Essen. Beginning in 1922, he introduced a completely new approach to photography in Germany. He abhorred the vagueness and falsification of photography of the art nouveau period (1890–1914). A firm believer in straight photography, he was fascinated by the beauty of everyday things. Like Strand before him, Renger-Patzsch considered a purely objective photography to be the true, if unattainable, goal.

Ansel Adams



"Mount Williamson—Clearing Storm" by Ansel Adams, 1944.



"Driving Shaft of a Locomotive" by Albert Renger-Patzsch, 1923.
By courtesy of the Gernsheim Collection, The University of Texas at Austin

His photographs are characterized by strong design, factual documentation, and stark realism stressing materials. Like Weston he insisted that the final image should exist, in all its completeness, before the exposure was made and that the print should directly record this image in full detail.

Renger-Patzsch's work was first exhibited in 1925, and three years later his most famous book appeared, *Die Welt ist schön* ("The World Is Beautiful"). The very title set the work apart in a period when artists were more concerned with creating abstractions than interpreting the environment. Renger-Patzsch looked hard at the world. The art historian Heinrich Schwarz wrote in 1929: "If today the photographs of Renger-Patzsch create more pure pleasure than many paintings, it is not an accident, but evidence that the time has found in the photographer a more sensitive instrument for the expression of its artistic needs than in the painter." Renger-Patzsch's subsequent books were largely concerned with regional landscapes and architecture, but his last two publications, *Im Wald* (1958; "Trees") and *Gestein* (1966; "Stones"), have an abstract beauty inspired by their themes.

Neue Sachlichkeit gathered momentum when Karl Blossfeldt's breathtaking, detailed magnifications of plants, which he had taken around 1900 to assist him in modeling plants, were published in *Original Forms of Art* (1928), followed by the *Magic Garden of Nature* (1932). About that time, August Sander, sick of the sweet-looking, posed studio portraits by which he had made his living for nearly 30 years, vowed "From now on I only want the honest truth about our time and people." In *Portrait of an Epoch* (1929) Sander presented such an unflattering portrait of the German middle class that in 1934 the Nazis impounded all unsold copies. Sander's idea of photographing tradespeople in his studio in Cologne inspired the American fashion photographer Irving Penn to shoot a similar series 30 years later. In both cases the portraits are unconvincing because the workers are divorced from their usual surroundings. The finest portraits of the German intelligentsia of the 1920s were taken by Hugo Erfurth of Dresden. They are imbued with strong artistic conception and a sympathetic understanding of the sitter.

Influence of abstract art. In 1919 Christian Schad, a member of the Dada group of modern artists in Geneva, amused himself by arranging small, flat objects directly on photographic paper. Upon exposure to light, the paper darkened more or less or not at all, according to the opacity or transparency of the objects. These "schadographs," as they came to be called, were minor contributions to the Dada movement and would be forgotten except that they inspired the American Surrealist painter Man Ray and the Hungarian constructivist painter László Moholy-Nagy to produce similar, though larger, abstract photographs.

Cameraless photographs

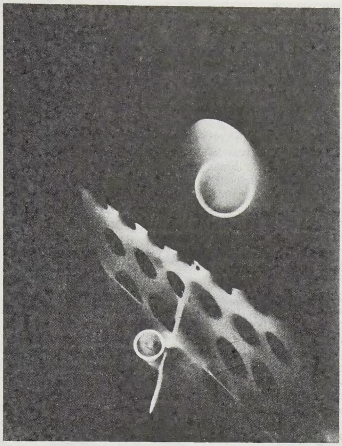
Man Ray had settled in Paris in 1921 and was supporting himself by taking portraits and fashion photographs. One day he accidentally set a glass funnel, a graduate, and a thermometer on a piece of photographic paper, thus producing "rayographs."

> I turned on the light; before my eyes an image began to form, not quite a simple silhouette of the objects as in a straight photograph, but distorted and refracted by the glass more or less in contact with the paper.

© AUGUST SANDER ARCHIVE, courtesy Sander Gallery, Inc., New York City



"Bricklayer's Mate" by August Sander, Cologne, 1929.



"Photogram," photogram by László Moholy-Nagy, 1926.
George Eastman House Collection

About the same time, in Berlin, Moholy-Nagy (or perhaps his wife, Lucia, a trained photographer) began to use three-dimensional objects to make similar cameraless photographs, which he called "photograms." Although identical in technique, the work of each was quite different. Man Ray emphasized the distorted but recognizable object; Moholy-Nagy the play of light, no matter how abstract. The former brought to photography the vision of the Surrealists; the latter, that of the Constructivists.

As a teacher in the influential Bauhaus art school, Moholy-Nagy explored the potential of the unconventional use of the camera as a means of discovering form. He delighted in the worm's-eye and the bird's-eye view. He considered the negative an end as well as a means. Amazed at the power of the medium to make visible the invisible, he collected X rays, photomicrographs, and astronomical and ultrahigh-speed photographs. When in 1925 he put together a book of these pictures titled *Painting, Photography, Film,* the eyes of the world were opened to the scope and breadth of photography as a tool for vision.

Photomontage. The art of combining photographs with watercolour paintings was a popular pastime in the 1870s. In the first decades of the 20th century the idea of freely combining mediums was revived when Cubist painters began to glue on their abstract canvases words clipped from newspapers, labels from bottles, and even actual objects. The extension of this collage technique (from the French *coller,* meaning "to glue") to photography was logical. The German artist John Heartfield, the greatest master of photomontage, claimed that the origin of the technique lay in postcards that he and his friend the German artist George Grosz sent to friends at the front during World War I. These were

Photomontage technique

> a mischmasch of advertisements for hernia belts, student song books and dog food, labels from schnaps and wine bottles, and photographs from picture papers, cut up at will in such a way as to say, in pictures, what would have been banned by the censors if we had said it in words.

Heartfield's photomontages, which were published weekly, first in Berlin and later in Prague, between 1929 and 1938, have a savage quality. Violent contrasts of the scale and perspective of the image elements, the ruthless cropping of heads and bodies, the substitution of machine parts for vital organs, and other seeming illogical juxtapositions, were carefully calculated to have a shock effect. Heartfield's anti-Fascist montages were among the strongest protests made by any visual artist.

Excellent montages were also produced in the 1920s by the German artists Hannah Höch, Herbert Bayer, Otto Umbehr, and the Surrealist painter Max Ernst. Most of these combined photographs cut from newspapers with other ephemera to express a specific idea. The montages of the Constructivists are more architectural: space is created with the purely photographic self-portrait of the Russian artist El Lissitzky (1924) and a fantasy world is built by Moholy-Nagy in his "Leda and the Swan." The power of montage lies in the tensions set up by the juxtaposition of disparate visual elements.

Documentary photography. At this same period, the documentary photographs of Eugène Atget first became known to the public. Beginning around 1898, this French photographer produced approximately 10,000 photographs of Paris and its environs that were direct, straightforward, and poetic in their sympathetic rendering of the very fabric of the city. He photographed shop fronts, buildings, wheeled vehicles of all kinds, decorative details, and the

By courtesy of Mrs. Gertrud Heartfield



"Twenty Years Later," photomontage by John Heartfield, 1934.

people who earned their living in the streets. Unknown to the photographic world, Atget worked alone, supporting himself by selling prints to architects, painters, and, above all, museums. The beauty of his photographs attracted the attention of Man Ray, who published a few of them in the periodical *La Revolution Surréaliste* in 1926. Upon Atget's death in 1927 his entire collection of prints and negatives was saved for posterity by the U.S. photographer Berenice Abbott (with the help of the New York art dealer Julien Levy); they are now in the collection of the Museum of Modern Art, New York City.

Although Atget was one of the most prominent photographers in the documentary field, he was by no means the first. The value of the photograph as a record of the world and man's conditions and achievements has been evident since the inception of the medium in 1839.

Early uses of documentary photography

Early documentary photographs were used to relay information about important events (*e.g.,* the Brady staff's record of the Civil War) as well as the scenery and people of distant or unexplored lands. They were also used to record the successive stages of significant or complex projects. The English artist Philip Henry Delamotte, for example, was hired to document in weekly photographs the progress of the construction of the Crystal Palace in London, from the laying of its foundation in 1852 to its opening by Queen Victoria two years later. Shortly afterward, by order of the French government, Édouard-Denis Baldus photographed to scale the sculpture, capitals, scrollwork, and other architectural details of the new wing of the Louvre Museum. Valuable work of a similar timely nature was undertaken by the English photographers Alfred and John Bool and Henry Dixon. Between 1875 and 1886 they worked for the Society for Photographing Old London, recording the historic buildings and relics that were gradually disappearing as a result of modernization.

The recognition of the power of photography to persuade as well as to inform came somewhat later. The classic sociological study *London Labour and the London Poor,* by Henry Mayhew (1851–62), was illustrated with drawings partly copied from daguerreotypes taken by Richard Beard; a sequel, patterned upon it, appeared in 1877—*Street Life in London,* by Adolphe Smith and John Thomson. The photographs, taken by Thomson, were moving, straightforward pictures of chimney sweeps, flower sellers, bargemen, and other tradesmen. They were reproduced by the woodburytype process, which gave exact, permanent facsimiles of the original prints. The intent of the publication was to show—as Charles Dickens had shown in his novels—the hardships and problems faced by the ever-growing working-class population of London. Each of the photographs was accompanied by a detailed explicative text by Smith. Oscar G. Rejlander photographed orphan children in the streets of London performing such humble tasks as cleaning boots and sweeping streets.

To Jacob A. Riis, a police reporter in New York City in the 1880s, the camera became the ally of his pen in the personal crusade he was waging to better the lot of the immigrants who then lived and worked in wretched conditions in tenements in Manhattan's Lower East Side. Himself an immigrant—he arrived in New York from Denmark in 1870—he knew at first hand the conditions he was seeking to eradicate. To supplement his written descriptions, he turned to photography. When the cameraman he employed proved unsatisfactory, he learned the process himself and was one of the first to use flash powder, a German invention recently introduced to the United States. His photographs, published in crude facsimile in newspapers and in his now-famous books, *How the Other Half Lives* (1890) and *Children of the Poor* (1892), were instrumental in stimulating legislative reform.

Beginning in 1905, Lewis W. Hine, a teacher at the Ethical Culture School in New York and a trained sociologist, began to make a pictorial record of immigrants passing through Ellis Island. He followed them in New York City, photographing them in their living quarters, at work, and in the streets. Later Hine traveled to textile mills, mines, and other places where young children were employed; largely through the evidence of his photographs, legislation was eventually passed against child-labour abuses.

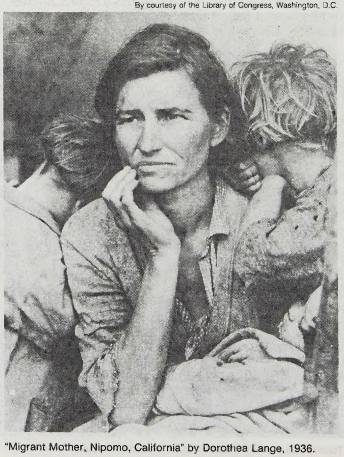
In England Sir Benjamin Stone, a member of Parliament for Birmingham, was obsessed by the wish to document old English customs and pageants that he rightly feared would gradually disappear. He was the most active member of the National Photographic Record Association, which he had founded for this purpose in 1895, leaving to the city of Birmingham a collection of 22,000 photographs.

Documenting the Depression in the United States

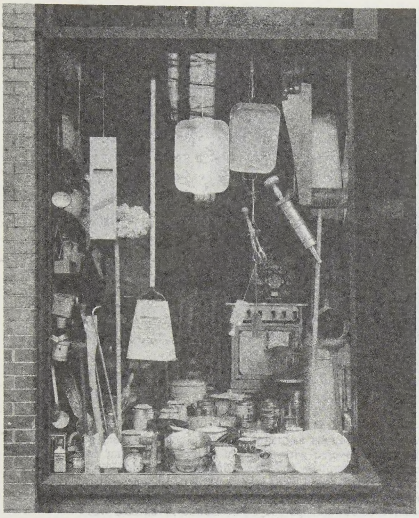
The body of photographs produced for the Farm Security Administration during the Great Depression in the United States has been preserved for future generations by the Library of Congress. The pictures cover the period from 1935 until the outbreak of World War II and were taken by a group of dedicated photographers, under the direction of Roy E. Stryker. Stryker, a professor of economics at Columbia University, was invited by the Secretary of Agriculture to form a historical section in the department to document the plight of farmers driven from their land in the dust bowl and who were migrating to the West. This at once took the form of a photographic project. The first photographer to be hired was Arthur Rothstein, a student of Stryker's. From California came Dorothea Lange, who had photographed migratory workers there. Her photographs are notable for their compassionate attitude toward people. In her "Migratory Pea Picker," a destitute young mother, surrounded by her children, peers at the camera with determination and courage. Walker Evans, with a direct uncompromising sense of environment and the beauty of everyday architecture, contributed a notable series. With the writer James Agee he documented the lives of sharecroppers in the book *Let Us Now Praise Famous Men* (1941). Others were documenting America at this time: Berenice Abbott produced a notable series of photographs of New York City, and Margaret Bourke-White with her husband, Erskine Caldwell, did a passionate survey of the South, which appeared in book form as *You Have Seen Their Faces* (1937).

Photojournalism. From the outset, photography has served the press. Within weeks after the French government's announcement of the process in 1839, magazines were publishing woodcuts or lithographs with the byline "from a daguerreotype." In fact, the two earliest illustrated weeklies—*The Illustrated London News,* which started in May 1842, and *L'Illustration,* based in Paris from its first issue in March 1843—owe their origin to the invention of photography. Early reproductions were generally crude, however, and carried little of the conviction of the original photograph. Regular use of photographs in magazines be-

By courtesy of the Library of Congress, Washington, D.C.



"Migrant Mother, Nipomo, California" by Dorothea Lange, 1936.



"Window Display, Bethlehem, Pennsylvania," gelatin silver print by Walker Evans, 1935.
Library of Congress, Washington, D.C.

gan with the perfection of the halftone process for facsimile reproduction in the 1890s. By 1915 newspapers had also turned to photography for reporting topical events, and the profession of newspaper illustrator gradually became obsolete. Although technical advances improved reproduction quality, the subjects and styles of early journalistic photography were generally unimaginative and dull.

Influence of the Ermanox and Leica cameras

It was not until the appearance of the Ermanox in 1924 and the Leica in 1925 that a new approach to pictorial journalism began to emerge. These two German-made miniature cameras, fitted with wide-aperture lenses, required extremely short exposure times for outdoor work and were even able to photograph indoor scenes by available light. This capability led to photographs whose informality of pose and sense of presence were remarkable. In 1928–29 two of the largest picture magazines in Europe, the *Münchner Illustrierte Presse* and the *Berliner Illustrirte Zeitung,* began to print the new style of photographs. Erich Salomon captured revealing candid portraits of politicians and other personalities by sneaking his camera into places and meetings officially closed to photographers. Felix H. Man was encouraged by Stefan Lorant, editor of the *Münchner Illustrierte,* to take sequences of photographs at interviews and cultural and social events. Lorant then laid out the photographs in imaginative picture essays.

The example of the German picture magazines was followed in other parts of Europe and in the United States. One was the short-lived *Vu,* established in Paris in 1928. An issue of *Vu* devoted entirely to the Spanish Civil War contained memorable photographs by Robert Capa. In 1936 both *Life* and *Look* were conceived in America, and a formula was evolved in which the picture editor, photographer, researcher, and writer constituted a team. The result was the creation of a definite photographic style.

Life's first photographers

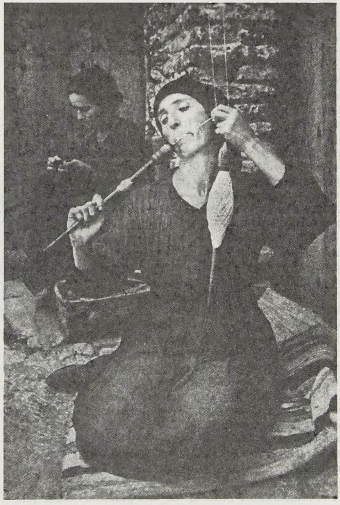
Life's first photographers were Margaret Bourke-White, already famous for her industrial photographs made largely for the magazine *Fortune;* Alfred Eisenstaedt, an experienced photo reporter for the Keystone Picture Agency in Germany; Peter Stackpole, whose photographs of the Golden Gate Bridge in San Francisco attracted much attention; and Thomas D. McAvoy, a news photographer who had pioneered in photographing interior scenes by available light. The concept of *Life* from the start, according to its founder, Henry Luce, was to replace haphazard picture taking and editing with the "mind-guided camera." Photographers were briefed for their assignments and encouraged to take great quantities of photographs, in order that the editors might have a large selection. The visual organization of the picture story was carefully planned for maximum reader impact. The opening photograph of the picture essay established the situation, and like written narration there was a visual climax and a definite conclusion. Usually the photographs were chosen and arranged on the pages before the accompanying text was written. Unlike the illustrated article, the picture essay quite logically is based upon the photographs, and the text is devoted to information that cannot be expressed visually: names, dates, places.

Life and *Look* preferred to use pictures of great sharpness and depth. Thus, instead of unobtrusive miniature cameras, American photographers used large-format cameras requiring slow lenses, large plates, and additional flash light. At first, the photographers made great use of so-called synchroflash; *i.e.,* flash that was synchronized with the camera shutter. The next step was the multiple flash, which made possible more sophisticated and pleasing lighting effects. By duplicating the existing illumination with the flash lighting, photographers could await the moment when people were at their most natural and then make the exposure without the need for posing.

This way of photographing was soon challenged. Lorant, who had left the *Münchner Illustrierte Presse,* moved in 1934 from Germany to London. There, he established the magazines *Weekly Illustrated* (1934) and *Picture Post* (1938). Staff photographers on both magazines included old colleagues from Germany, such as Man and Kurt Hutton. They, as well as other contributors, were encouraged to develop the technique and pictorial style of available-light photographs so brilliantly begun in the 1920s. Their pictures had a remarkable naturalness that brought great reader appeal—so much so that *Life* began to publish sim-

Available-light photography

W. Eugene Smith



"Spanish Village" by W. Eugene Smith, 1951.



"Children in Seville, Spain" by Henri Cartier-Bresson, 1933.
Henri Cartier-Bresson—Magnum

ilar photographs and in 1945 hired a former *Picture Post* photographer, Leonard McCombe, with an extraordinary clause in his contract: he was forbidden to use flash.

Memorable groups of photographs have been taken for the picture magazines. Examples are Man's "A Day with Mussolini," first published in the *Münchner Illustrierte Presse* (1931), and then, with a brilliant new layout, in *Picture Post;* W. Eugene Smith's "Spanish Village" (1951) and "Nurse Midwife" (1951) in *Life;* and Eisenstaedt's informal, penetrating portraits of famous Britons, also in *Life.*

The photojournalist's ability to train himself to perceive the significant in the fraction of a second and to use the camera with such speed and precision that the instantaneous perception is preserved forever is a great creative gift. The gift is evident in the work of the Hungarian André Kertész as early as 1915 and in his later work in Paris during the 1920s. The Frenchman Henri Cartier-Bresson began around 1930 to develop the style that he later called the search for the "decisive moment." To him the camera was an "extension of the eye." With extraordinary precision he perceived a fully composed picture of the most fleeting scenes. Unlike many other photographers, he did not crop or recompose his pictures after they had been taken: every detail of a Cartier-Bresson photograph is present in the negative. He preferred the miniature 35-mm-film camera. When he found a picture possibility he stalked his prey unobtrusively, working with his camera to a visual climax.

Cartier-Bresson's instantaneous image

Colour photography. Photography's transmutation of nature's colours into various shades of black and white had been considered a drawback of the process from its inception. Hence, at the request of a client, many portrait photographers collaborated with artists who hand-tinted daguerreotypes and calotypes or painted over albumen prints in oils. Some artists also copied the photograph onto canvas; others, such as Franz von Lenbach in Munich, had the image projected onto canvas that had been made light-sensitive, whereupon they painted freely over it. In Japan, where hand-coloured woodcuts had a great tradition and labour was cheap, some firms from the 1870s on sold photographs of scenic views and daily life that had been delicately hand-tinted. In the 1880s photochromes, colour prints made from hand-coloured photographs, became fashionable and remained popular until they were gradually replaced in the first decades of the 20th century by Autochrome plates.

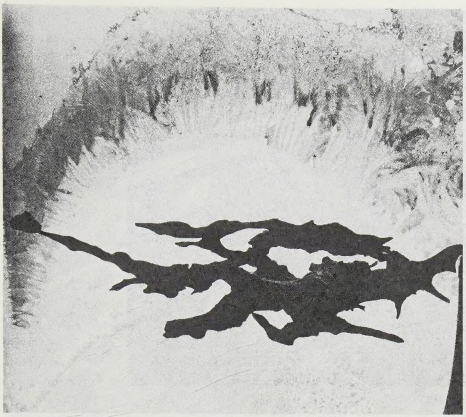
The Autochrome process, the first practical colour photography process, was introduced in France in 1907 by Auguste and Louis Lumière. It used a colour screen (a glass plate covered with grains of starch dyed to act as primary-colour filters and black dust that blocked all unfiltered light) coated with a thin film of panchromatic (*i.e.,* sensitive to all colours) emulsion and resulted in a positive colour transparency. The Lumières' success was due in part to the introduction of panchromatic emulsion the previous year by a London firm of photographic plate manufacturers. All previous experimenters trying to solve the problem of colour photography had been seriously impeded by the comparative insensitivity of the earlier negative material to all colours except blue and violet.

Researchers continued to look for improvements and alternative colour processes, and in 1935 Leopold Godowsky, Jr., and Leopold Mannes, two American musicians working with the Kodak Research Laboratories, initiated the modern era of colour photography with their invention of Kodachrome film. With this reversal (slide) film, colour transparencies could be obtained that were suitable both for projection and for reproduction. A year later the Agfa Company of Germany developed the Agfacolor negative–positive process, but due to World War II the film did not become available until 1949. Meanwhile, Kodak had introduced in 1942 the Kodacolor negative–positive film that, 20 years later—after many improvements in quality and speed and a great reduction in price—became the most popular film used for amateur photography. Today about 80 percent of all photographs are shot in colour.

The invention of Kodachrome film

Later trends. Throughout its history, there have been two complementary yet distinct aesthetic approaches to photography. On the one hand, there has been the recognition of the basic qualities of photography and the desire to make use of them in a functional way. On the other hand, there have been those who believe that the most aesthetic use of photography is to relate it to other mediums. Since 1950 both these trends have been pursued with vigour, and to them has been added a third approach, the expressive, emotional use of photography pioneered by Stieglitz with his "equivalents" series.

In the United States, Minor White, through his long career, his writing, his teaching, and his founding and editing of the influential magazine *Aperture,* developed the Stieglitz approach to a highly sophisticated level. For him, the photograph must be transformed in such a manner that the viewer can read an inner message, which is not visible upon the surface, but which is carried by it. White's book *Mirrors, Messages, Manifestations* (1970) is a collection of superb photographs that present his spiritual biography.



"Ritual Branch" by Minor White, 1958.
George Eastman House Collection

To Aaron Siskind, who worked with wall scrawls, weathered wood and plaster, torn billboards, and what he called "the detritus of our world," the photograph must communicate more than the subject itself. The photographs of Harry Callahan express his highly developed sense of linear form, often by means of sharp contrasts of black and white and multiple images.

One of the finest photographers working after World War II was former *Life* photographer Andreas Feininger. His dramatic close-ups of architecture and nature reveal a thorough understanding of design, composition, and structure, which can perhaps be attributed to his early training as an architect.

Colour photography has become increasingly popular within the ranks of the amateur. Although many professionals have explored the artistic possibilities of colour, which can add intensity and realism to the picture and increase interest in the subject, some prefer black-and-white to colour film for aesthetic reasons. Among professional colour photographers, Eliot F. Porter and Marie Cosindas (one of the first to work with Polaroid instant films) were the leaders in America: both preferred a somewhat realistic approach.

A considerable impediment to a more widespread use of colour in monographs and other publications studying the art of photography is the often prohibitive price of colour reproduction.

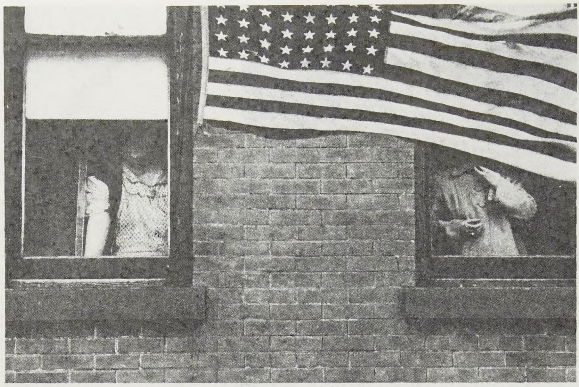
"Social landscape" photography

The urban social scene viewed objectively without sentiment or moralization—often called the "social landscape"—has been a subject of much interest to photographers in both the United States and Europe. The work of such photographers as Robert Frank, Bruce Davidson, and William Klein takes the form of penetrating sociological observations, somewhat reminiscent of the documentary

Aaron Siskind



"Chicago 1949" by Aaron Siskind.



"Parade—Hoboken, New Jersey" by Robert Frank, 1955.
Robert Frank

photography of the 1930s. The approach differs in that, rather than presenting problems faced by a certain level of society, emphasis is placed on the effects of the urban environment upon people.

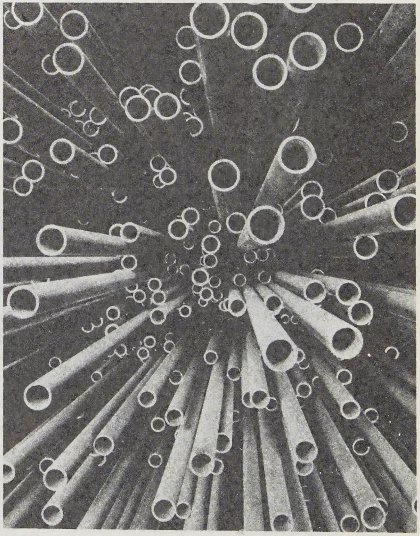
In England and Europe, trends closely parallel those in the United States. Bill Brandt, whose classic social reportage began with the book *The English at Home* (1936) and continued during World War II with trenchant photographs of life in the bomb shelters of London, changed his style during the 1950s. Using an extremely wide-angle lens he created startling abstract studies of nudes, some of which are reminiscent of Henry Moore sculptures. Cecil Beaton, fashion photographer for British *Vogue,* created pictures with exquisite taste. His moving photographs of London during the bombing of World War II and of other scenes of the war stress the human aspect rather than the military. The "social landscape" school is well represented in the work of Tony Ray-Jones and Raymond Moore. The strongest contribution probably lies in the field of photojournalism. George Rodger, one of the founders of the Magnum agency and a former *Life* photographer, and Bert Hardy, a former *Picture Post* photographer, provided a solid tradition for the work of Don McCullin, who—like Robert Capa before him—traveled from war to war, photographing with deep compassion the conflicts that appeared in his book *The Destruction Business* (1971).

The most imaginative photographs in Europe are mostly made for publication, rather than for exhibition or hanging in galleries and homes. Europeans, in general, do not consider the photographic print as an end in itself, but as a step toward reproduction in periodicals and books. Brassaï (the pseudonym of Gyulas Halász) made his name in photography with the publication of *Paris de nuit* (1933), intimate and sympathetic photographs of night life in the more humble quarters of Paris. In the 1950s, like Aaron Siskind, he became fascinated with wall scrawls and graffiti. Robert Doisneau was a master of humour and satire, catching moments of absurdity in everyday life. Lucien Clergue turned to the natural scene and to the nude in surf. The tradition of cameraless abstraction was enriched by the Belgian Pierre Cordier with the introduction in 1958 of his "chimigrammes"—colour images made, not by light, but chemical action on photographic paper.

Steinert's Subjective Photography

In Germany the greatest influence was the teaching of Otto Steinert at Saarbrücken and, since 1959, at the Folkwangschule in Essen. Almost single-handedly he brought back to Germany that spirit of experimentation and boldness of concept that had been suppressed during the Third Reich. An excellent photographer, Steinert was the founder in 1951, together with the art historian J.A. Schmoll gennant Eisenweth, of a movement they named "Subjective Photography." He led the group "Fotoform," which first exhibited its work in 1950, to explore the creative potential of any possible expressive technique. Peter Keetman is one of the strongest representatives of Fotoform's dedication to creating innovative, expressive graphic designs and abstract patterns. Robert Häusser, a student of Steinert, was strongly influenced by Fotoform in his early work. Later he introduced mystifying ele-

PPS, Galerie F.C. Gundlach, Hamburg



"Pipes" by Peter Keetman, 1958.



Salt-encrusted, sand-lined drainage patterns at the fringe of the Etosha Pan, Namibia, aerial photograph by Georg Gerster, 1980–81.
© COMSTOCK INC./Georg Gerster

ments into his landscapes, willfully distorting reality until it bordered on abstract expressionism. Chargesheimer, like Robert Frank in the United States, laid great stress on the unpleasant side of his themes in hard, almost brutal photographs of German cities. Floris M. Neusüss, who teaches at the University of Kassel, in Germany, is a great exponent of conceptual photography, which uses concepts as material and in which the preconceived idea is more important than the object. Erwin Fieger stands out with such books as *13 Photo-Essays* (1969), *Japan, Sunrise Island* (1971), and *Mexico* (1973) as one of the finest in the field of colour reportage. Horst Baumann specialized in illustration, particularly in colour.

The Austrian Ernst Haas was a master of colour photography, turning toward the abstract in his remarkable photographs of blurred action and bold compositions. A member of the Magnum group, a cooperative formed by Cartier-Bresson and others in 1947, Haas produced work that is international in scope.

Outstanding among Swiss photographers working after World War II was Werner Bischof, who, until his death in 1954, movingly photographed refugees in Europe, the famine in India, Japan, and the Incas of Peru. Superb colour work was produced by Emil Schulthess for his books *Africa* (1959), *The Amazon* (1962), *China* (1966), and others. Georg Gerster revealed in his aerial views, primarily taken in colour, a beauty of design that frequently comes close to modern art.

The Czech Josef Sudek (1896–1976) is best known abroad for his still lifes, Vilem Heckel for his photographs of industry and mountains, Karel Plicka for his views of Prague, and Josef Koudelka for his impressive work on Gypsies. In Sweden each member of the group of photographers known as TIO ("Ten") has produced outstanding work. Each works in a different field, but all are united by their modern style. The most gifted Russian photographer, Alexander M. Rodchenko, was too modern for Stalin's taste, and his work was banished until the dictator's death. During World War II Dimitri Baltermans produced fine reportage work on the front.

In Italy Franco Fontana, shooting in colour with telelenses, created amazing abstractions of landscapes, fields, and buildings; Mario de Biasi made a fine record of the uprising in Budapest in 1956; and Fulvio Roiter produced a series of travel books, the most sensitive and romantic

Copyright Hiroshi Hamaya –Magnum Photos



Children wearing straw capes to protect them from the cold, Niigata prefecture, Japan, from *The Snow Country* by Hiroshi Hamaya, 1956.

of which focused on his hometown, Venice. The chief concern of most photographers in Italy and elsewhere in Europe, however, is centred upon the recording of the social scene: Mario Giacomelli's series of photographs of nuns and village life is typical and outstanding.

In Japan the extraordinary documentary and landscape photographs by Hiroshi Hamaya parallel the straight photographic approach. Superb photographs were also produced by Takayuki Ogawa, the social realist, and by the imaginative Eikoh Hosoe.

Experimental approaches to photography

A lively interest in expanding the medium of photography beyond the straight approach was characteristic of much work in the late 20th century. The experimentation took many forms, including the revival of long-obsolete printing techniques such as the gum bichromate process and the platinotype. Perhaps the most startling change to come about after 1960 was an interchange of mediums between photographers and painters. Many photographers made liberal use of manual techniques, such as negative and print retouching and the addition of colour. Simultaneously, painters, who had long utilized photographs as tools for observation, boldly imitated the very quality of photographic vision and sometimes introduced unaltered photographs by collage techniques or silkscreen reproduction directly into their canvases.

Another late 20th-century trend was the photographer's increasing reliance on books for the presentation of his work. Several factors contributed to this practice, among them the demise of many of the major picture magazines, technical developments that provided better printing quality at lower costs, and the complete acceptance of photography as an art worthy of study and preservation.

(Be.N./H.E.R.G.)

The technology of photography

The most widely used photographic process is the black-and-white negative–positive system (Figure 1). In the camera the lens projects an image of the scene being photographed onto a film coated with light-sensitive silver salts, such as silver bromide. A shutter built into the lens admits light reflected from the scene for a given time to produce an invisible but developable image in the sensitized layer, thus exposing the film.

During development (in a darkroom) the silver salt crystals that have been struck by the light are converted into metallic silver, forming a visible deposit or density. The more light that reaches a given area of the film, the more silver salt is rendered developable and the denser the silver deposit that is formed there. An image of various brightness levels thus yields a picture in which these brightnesses are tonally reversed—a negative. Bright subject details record as dark or dense areas in the developed film; dark parts of the subject record as areas of low density; *i.e.,* they have little silver. After development the film is treated with a fixing bath that dissolves away all undeveloped silver salt and so prevents subsequent darkening of such unexposed areas. Finally, a wash removes all soluble salts from the film emulsion, leaving a permanent negative silver image within the gelatin layer.

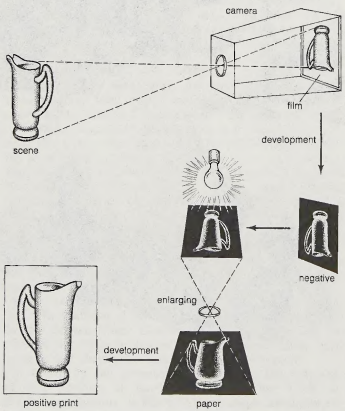


Figure 1: Sequence of negative–positive process, from the photographing of the original scene to enlarged print (see text).

Creation of a negative image

A positive picture is obtained by repeating this process. The usual procedure is enlargement: the negative is projected onto a sensitive paper carrying a silver halide emulsion similar to that used for the film. Exposure by the enlarger light source again yields a latent image of the negative. After a development and processing sequence the paper then bears a positive silver image. In contact printing the negative film and the paper are placed face to face in intimate contact and exposed by diffused light shining through the negative. The dense (black) portions of the negative image result in little exposure of the paper and, so, yield light image areas; thin portions of the negative let through more light and yield dark areas in the print, thus re-creating the light values of the original scene.

CAMERAS AND LENSES

Basic camera functions. In its simplest form, the camera is a light-tight container carrying a lens, a shutter, a diaphragm, a device for holding (and changing) the film in the correct image plane, and a viewfinder to allow the camera to be aimed at the desired scene.

The lens projects an inverted image of the scene in front of the camera onto the film in the image plane. The image is sharp only if the film is located at a specific distance behind the lens. This distance depends on the focal length of the lens (see below *Characteristics and parameters of lenses*) and the distance of the object in front of the lens. To photograph near and far subjects, all but the simplest cameras have a focusing adjustment that alters the distance between the lens and the film plane to make objects at the selected distance produce a sharp image on the film. In some cameras focusing adjustment is achieved by moving only the front element or internal elements of the lens, in effect modifying the focal length.

The two basic elements of exposure

The shutter consists of a set of metallic leaves mounted in or behind the lens or a system of blinds positioned in front of the film. It can be made to open for a predetermined time to expose the film to the image formed by the lens. The time of this exposure is one of the two factors controlling the amount of light reaching the film. The other factor is the lens diaphragm, or aperture, an opening with an adjustable diameter. The combination of the diaphragm opening and exposure time is the photographic exposure. To obtain a film image that faithfully records all the tone gradation of the object, this exposure must be matched to the brightness (luminance) of the subject and to the sensitivity or speed of the film. Light meters built into most modern cameras measure the subject luminance and set the shutter or the lens diaphragm to yield a correctly exposed image.

Principal camera types. The simplest camera type, much used by casual amateurs, has most of the features listed in the previous section—lens, shutter, viewfinder, and film-holding system. The light-tight container traditionally had a box shape. Present-day equivalents are pocket cameras taking easy-load film cartridges or film disks. Typically, a fixed shutter setting gives about $^1/_{50}$-second exposure; the lens is permanently set to record sharply all objects more than about five feet (1.5 metres) from the camera. Provision for a flash may be built in. Though simple to handle, such cameras are in daylight restricted to pictures of stationary or slow-moving subjects.

The 35-mm miniature camera. Perforated 35-millimetre (mm) film (originally standard motion-picture film) in cartridges holding 12 to 36 exposures with a nominal picture format of 24 × 36 mm is employed in miniature cameras.

Smaller image formats down to 18 × 24 mm (half frame) may be used. The 35-mm camera has a lens with a range of apertures and a shutter with exposure times typically from one second to $^1/_{1,000}$ second or shorter, and it can focus on subject distances from infinity down to five feet or less. A winding lever or built-in motor advances the film from one frame to the next and at the same time tensions (cocks) the shutter for each exposure. At the end of the film load the film is rewound into the cartridge for removal from the camera in daylight.

A 35-mm camera usually has a direct-vision viewfinder, often combined with a rangefinder or autofocus system for accurate distance settings. Most current versions incorporate a light meter coupled with the exposure settings on the camera. Advanced models may have interchangeable lenses and an extended accessory system. Many 35-mm cameras are single-lens reflex types (see below).

The ultraminiature or subminiature. This camera takes narrow roll film (16-mm or 9.5-mm) in special cartridges or film disks. The picture size ranges from 8 × 10 mm to 13 × 17 mm. These formats are used for making millions of snapshooting pocket-size cameras; special versions may be as small as a matchbox for unobtrusive use.

The view, or technical, camera. For studio and commercial photography the view, or technical, camera takes single exposures on sheet films (formerly plates) usually between 4 × 5 inches and 8 × 10 inches. A front standard carries interchangeable lenses and shutters; a rear standard takes a ground-glass screen (for viewing and focusing) and sheet-film holders. The standards move independently on a rail or set of rails and are connected by bellows. Both standards can also be displaced laterally and vertically relative to each other's centre and swung or tilted about horizontal and vertical axes. These features provide versatility in image control (sharpness distribution, subject distance, and perspective), though not speed in use. The view camera is nearly always mounted on a tripod.

The medium-size hand camera. This type of camera takes sheet film (typical formats of from 2½ × 3½ inches to 4 × 5 inches), roll film, or 70-mm film in interchangeable magazines; it has interchangeable lenses and may have a coupled rangefinder. Special types use wide-angle lenses and wide picture formats (*e.g.*, 2¼ × 4½ to 2¼ × 6¾ inches [6 × 12 to 6 × 17 centimetres]). The medium-size hand camera was popular with press photographers in the first half of the 20th century. Older versions had folding bellows and a lens standard on an extendable baseboard or strut system. Modern modular designs have a rigid body with interchangeable front and rear units.

The folding roll-film camera. The folding roll-film camera, now rare, resembles the 35-mm miniature camera in shutter and viewfinder equipment but has bellows and folds up to pocketable size when not in use. Generally it takes roll films holding eight to 16 exposures; typical picture sizes are 2¼ × 2¼, 2¼ × 3¼, or 1¾ × 2¼ inches. Some 35-mm cameras were also produced with bellows.

The single-lens reflex. The ground-glass screen at the back of the studio, or view, camera slows down picture taking because the screen must be replaced by the film for an exposure. The single-lens reflex camera (Figure 2) has a screen, but the film remains constantly in position. A 45° mirror reflects the image-forming rays from the lens onto a screen in the camera top. The mirror moves out of the way during the exposure and back again afterward for viewing and focusing the next picture. The image on the screen therefore temporarily disappears from view during the exposure. Present-day single-lens reflexes are either 35-mm cameras or advanced roll-film models. Most 35-mm reflexes have optical prism systems for eye-level screen viewing, built-in light-meter and electronic exposure-control systems, interchangeable lenses, and numerous other refinements. Often the camera is part of an extensive accessory system. Advanced roll-film reflexes are even more modular, with interchangeable viewfinders, focusing screens, and lenses.

Reflecting the image through the lens

The twin-lens reflex. The twin-lens reflex is a comparatively bulky dual camera (Figure 3) with a fixed-mirror reflex housing and top screen mounted above a roll-film box camera. Its two lenses focus in unison so that the top

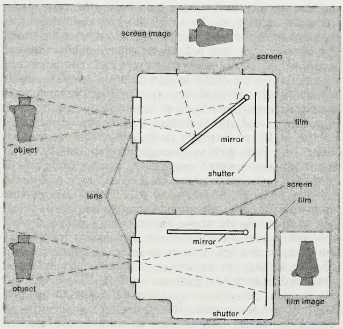


Figure 2: *Single-lens reflex principle.*
(Top) Before exposure, upright but laterally reversed image is visible on top screen. (Bottom) During exposure, with mirror up and shutter open, an upside down but laterally correct image is projected on the film.

screen shows the image sharpness and framing as recorded on the film in the lower section. The viewing image remains visible all the time, but the viewpoint difference (parallax) of the two lenses means that the framing on the top screen is not exactly identical with that on the film.

Shutter and diaphragm systems. Principal present-day shutters are the leaf shutter and the focal-plane shutter.

The leaf shutter. The leaf, or diaphragm, shutter consists of a series of blades or leaves fitted inside or just behind the lens. The shutter opens by swinging the leaves simultaneously outward to uncover the lens opening. The leaves stay open for a fixed time—the exposure time—and then close again. A combination of electromagnets or electromagnets and springs drives the mechanism, while an electronic circuit—often coupled with a light metering system—or an adjustable escapement in mechanical shutters controls the open time. This is typically between one second and $^1/_{500}$ second.

Focal-plane shutter. The focal-plane shutter consists of two light-tight fabric blinds or a combination of metal blinds moving in succession across the film immediately in front of the image plane. The first blind uncovers the film

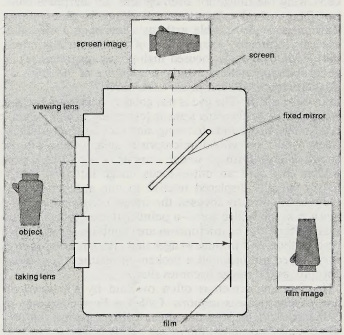


Figure 3: *Principles of the twin-lens reflex camera.*
The fixed mirror deflects light rays coming through the viewing lens to a top screen, which shows the image upright but laterally reversed. Light from the object also goes through the taking lens, which is mounted on a common panel with the viewing lens, and is projected on the film.

and the second blind covers it up again, the two blinds forming a traveling slit the width of which determines the exposure time: the narrower the slit, the shorter the time. The actual travel time is fairly constant for all exposure times. A mechanism or electromagnet and control circuit triggers the release of the second blind. Focal-plane shutters are usually adjustable for exposure times between one second (or longer) and $1/1{,}000$ to $1/4{,}000$ second.

Diaphragm and shutter settings. In the lens diaphragm a series of leaves increases or decreases the opening to control the light passing through the lens to the film. The diaphragm control ring carries a scale of so-called *f*-numbers, or stop numbers, in a series: such as 1.4, 2, 2.8, 4, 5.6, 8, 11, 16, 22, and 32. The squares of the *f*-numbers are inversely proportional to the amount of light admitted. In the above international standard series, each setting admits twice as much light as the next higher *f*-number, or stop (giving twice as much exposure).

The series of *f*-numbers

Shutter settings on present-day cameras also follow a standard double-or-half sequence—*e.g.*, 1, $1/2$, $1/4$, $1/8$, $1/15$, $1/30$, $1/60$, $1/125$, $1/250$, $1/500$, $1/1{,}000$ second, and so forth. The shorter the exposure time, the "faster" the shutter speed.

Exposure values. An attempt to simplify the mathematics of *f*-number and shutter speed-control functions led to the formulation of exposure values (EV). These run in a simple whole-number series, each step (EV interval) doubling or halving the effective exposure. The lower the EV number, the greater the exposure. Thus, EV 10 gives twice as much exposure as EV 11 or half as much as EV 9. Each EV value covers a range of aperture/speed combinations of the same equivalent exposure; for instance, *f*/2.8 with $1/250$ second, *f*/4 with $1/125$ second, and *f*/5.6 with $1/60$ second. For a time some cameras carried an EV scale and coupled the aperture and speed settings; at a given EV setting in such cameras selecting various speeds automatically adjusted the aperture to compensate and vice versa. Exposure-value setting scales became obsolete with exposure automation, but the notation remains in use to indicate either exposure levels or—at specified film speeds—lighting levels requiring a given exposure.

Automatic-diaphragm systems. On a camera with a viewing screen (view camera or single-lens reflex) viewing and focusing are carried out with the lens diaphragm fully open, but the exposure is often made at a smaller aperture. Reflex cameras (and increasingly also view cameras) therefore incorporate a mechanism that automatically or semiautomatically stops down (reduces) the lens to the working aperture immediately before the exposure.

Methods of focusing and framing. The ground-glass (now mostly grained plastic) screen is the most direct way of viewing the image for framing and for sharpness control. The screen localizes the image plane for observation. The image is also visible without a screen, but then the eye can locate the image plane of maximum sharpness only with a precisely focused high-power magnifier. This aerial focusing method avoids interference of the ground-glass structure with sharpness assessment.

Focusing aids. The eye is not good at recognizing slight unsharpness, so focusing screens (especially in reflex cameras) often incorporate focusing aids such as a split-image wedge alone or with a microprism area, in the screen centre. The split-image wedge consists of a pair of prism wedges that split an out-of-focus image into two sharp halves laterally displaced relative to one another. When the lens is correctly focused the image becomes continuous across the wedge area—a point that the eye can assess more precisely. The microprism area contains several hundred or thousand minute wedges that give a blurred image very ragged outlines and a broken-up texture; these clear abruptly as the image becomes sharp.

Micro-prism grid of multiple wedges

The focusing screen is often overlaid by a pattern of fine concentric lens sections. Called a Fresnel screen, it redirects the light from the screen corners toward the observer's eye and makes the image evenly bright.

Cameras without a screen generally are equipped with a distance scale, the lens being set to the estimated object distance. More advanced cameras have an optical rangefinder as a distance-measuring aid; it consists of a viewfinder (see below) and a swinging mirror a few inches to one side of the viewfinder axis. As the eye views an image of the object, the mirror superimposes a second image from a second viewpoint. On turning the mirror through the correct angle, which depends on the object distance, the two images are made to coincide. The mirror movement can be linked with a distance scale, or coupled with the lens focusing adjustment. When the lens is incorrectly focused, the rangefinder shows a double or split image. In place of a rotating mirror, the rangefinder may use swinging or rotating optical wedges (prisms).

Autofocus systems. Some cameras evaluate the coincidence (or lack thereof) between two rangefinder images by image analysis with a microchip system. This signals electronically when the lens is set to the correct distance and often carries out the distance setting by a servomotor built into the camera. Such focusing automation makes the camera even simpler to use. Alternative automatic ranging systems used in amateur cameras depend on triangulation with infrared rays or pulses sent out by a small light-emitting diode (LED), or on measurement of the time an ultrasonic signal takes to be reflected back from the subject (sonar).

While these devices measure distance automatically, single-lens reflex cameras may incorporate electronic image-analysis systems to measure sharpness. The signal output of such systems actuates red or green LEDs in the camera finder system to show whether the image is sharp or not. The same signal can control a servomotor in the lens for fully automatic focusing. These devices are limited at low lighting and contrast levels—where the human eye also finds sharpness assessment difficult.

Viewfinders. The sighting devices in cameras lacking screens are called viewfinders; they show how much of the scene will appear on the film. The simplest viewfinder is a wire frame above the camera front, with a second frame near the back to aid the eye in correct centring. Most present-day finders are built into the camera and are compact lens systems. Bright-frame finders show a white frame reflected into the view to outline the field recorded on the film. An alternative form is the reflecting viewfinder in which the photographer looks down into a field lens on top of the camera. The upper section of a twin-lens reflex camera is such a reflecting finder.

Parallax problems with close-at-hand subjects

As the viewfinder axis in a camera other than a single-lens reflex does not usually coincide with the lens axis, the finder's and the lens's views do not exactly match. This parallax error is insignificant with distant subjects; with near ones it is responsible for the familiar fault of a portrait shot of a head that appears partly cut off in the picture even though it was fully visible in the finder. Camera viewfinders may have parallax-compensating devices.

The optical finder gives a direct upright and right-reading view of the subject with the camera held at eye level. The traditional reflex camera, held at waist level, showed a laterally reversed view. Modern reflexes have a pentaprism arrangement that permits upright, right-reading, eye-level viewing by redirecting the image from the horizontal screen on top of the camera.

Exposure-metering systems. Exposure meters, or light meters, measure the light in a scene to establish optimum camera settings for correct exposures. A light-sensitive cell generates or controls an electric current according to the amount of light reaching the cell. The current may energize a microammeter or circuit controlling LEDs to indicate exposure settings. In most modern cameras the current or signal acts on a microprocessor or other circuit that directly sets the shutter speed or lens aperture. The cell usually is a silicon or other photodiode generating a current that is then amplified. In older cadmium sulfide cells the light falling on the cell changed the latter's resistance to a current passing through it. Selenium cells, still used in some cameras, also generate a current but are larger and less sensitive.

Through-the-lens (TTL) meter systems

Single-lens reflex cameras have one or more photocells fitted in the pentaprism housing to measure the brightness of the screen image. The exposure reading depends on the light coming through the lens (TTL metering) and so allows for the lens's angle of view, close-up exposure corrections, stray light, and other factors. Some TTL sys-

tems divert the light from the lens to a photocell before it reaches the screen (*e.g.*, by beam-splitting arrangements or the use of photocells behind a partly reflecting mirror), or they measure the light reflected from the film or from a specially structured first shutter blind at the beginning of, or during, the exposure. Such off-the-film (OTF) measurement is also used for electronic flash control (see below).

View cameras may use a photocell on a probe that can be moved to any point just in front of the focusing screen, thus measuring image brightness at selected points of the image plane. This takes place before the exposure, and the probe is then moved out of the way. Professional photographers also use hand-held separate exposure meters and transfer the readings manually to the camera.

Flash systems. Flash is a widely used artificial light source for photography, providing a reproducible light of high intensity and short duration. It can be synchronized with an instantaneous exposure. Being battery powered, small flash units are self-contained.

Electronic flash. The most common flash system depends on a high-voltage discharge through a gas-filled tube. A capacitor charged to several hundred volts (by a step-up circuit from low-voltage batteries or from the line voltage supply) provides the discharge energy. A low-voltage circuit generating a high-voltage pulse triggers the flash, which lasts typically $^1/_{1,000}$ second or less. Small electronic flash units may be built into or clipped onto the camera. Larger units are attached with brackets. Large professional units with floodlight and spotlight fittings are used in studio photography. Even small flashes often have adjustable reflectors, for example, to illuminate an indoor subject by the flash reflected from the ceiling or walls.

Automatic and dedicated flash. Electronic flash units often incorporate a fast-responding photodiode that cumulatively measures the light reflected from the subject and switches off the flash when that light has reached a preselected amount (computer flash). This flash-duration control thus adjusts the flash exposure automatically as long as the subject is within a certain distance range (typically from two to 20 feet). At lower power or near subject distances the duration of a computer flash may drop to $^1/_{50,000}$ second.

With certain camera–flash combinations OTF metering inside the camera can control the flash duration by suitable contacts made when the flash is attached to the camera. These "dedicated" flashes (so named because their control circuitry has to match that of specific cameras) may also signal in the camera finder when the flash is ready to operate and to set the camera automatically to its synchronizing shutter speed (see below).

Flashbulbs. An older type of flash is an oxygen-filled glass envelope containing a specific amount of aluminum or zirconium wire and means for igniting the wire in the bulb. The wire burns away with a brilliant flash lasting typically about $^1/_{100}$ to $^1/_{50}$ second. Each flashbulb can, however, yield only one flash. Current flashbulb systems use four to 10 tiny bulbs, each in its own reflector, arranged in cube or bar carriers that plug into cameras designed for them. The individual flashes are fired in turn by a battery and circuit in the camera through mechanically generated current pulses or other means. In view of the greater convenience of electronic flash, flashbulbs in their various forms are largely obsolescent.

Firing and synchronization. Flash units are usually fired with a switch in the camera shutter to synchronize the flash with the shutter opening. A contact in the camera's flash shoe (hot shoe) or a flash lead connects the unit with this shutter switch. The shutter contact usually closes the instant the shutter is opened. A focal plane shutter must fully uncover the film (generally at a shutter speed of $^1/_{60}$ second or slower) for flash synchronization. With flashbulbs the shutter must also stay open while the flash reaches its peak brightness—about $^1/_{50}$ second.

System cameras. From the development of the 35-mm miniature camera in the 1930s evolved the concept of the system camera that could be adapted to numerous jobs with a range of interchangeable components and specialized accessories. Today, most moderately advanced 35-mm miniatures take interchangeable lenses, close-up and photomicrographic attachments, filters, flash units, and other accessories. The most elaborate camera systems also include such accessories as alternative finder systems; interchangeable reflex screens, film backs, and magazines; and remote-control and motor-drive systems. Modular professional roll-film and view cameras are built up from a selection of alternative camera bodies, film backs, bellows units, lenses, and shutters. This is the nearest approach to the universal camera, assembled as required to deal with practically every type of photography.

Modular systems

Characteristics and parameters of lenses. The lens forming an image in the camera is a converging lens, the simplest form of which is a single biconvex (lentil-shaped) element. In theory such a lens makes a light beam of parallel rays converge to a point (the focus) behind the lens. The distance of this focus from the lens itself is the focal length, which depends on the curvature of the lens surfaces and the optical properties of the lens glass. An object at a very long distance (optically regarded as at "infinity") in front of the lens forms an inverted image in a plane (the focal plane) going through the focus. Light rays from nearer objects form an image in a plane behind the focal plane. The nearer the object, the farther behind the lens the corresponding image plane is located—which is why a lens has to be focused to get sharp images of objects at different distances.

Focal length and image scale. The image scale, or scale of reproduction, is the ratio of the image size to the object size; it is often quoted as a magnification. When the image is smaller than the object, the magnification of the object is less than 1.0. If the image is $^1/_{20}$ the size of the object, for example, the magnification may be expressed either as 0.05 or as 1:20. For an object at a given distance, the scale of the image depends on the focal length of the lens (Figure 4). A normal camera lens usually has a focal length approximately equal to the diagonal of the picture format covered. A lens of longer focal length gives a larger scale image but necessarily covers less of the scene in front of the camera. Conversely, a lens of shorter focal length yields an image on a smaller scale but—provided the angle of coverage is sufficient (see below)—takes in more of the scene. Many cameras, therefore, can be fitted with interchangeable lenses of different focal lengths to allow varying the image scale and field covered. The focal length of a lens in millimetres (sometimes in inches) is generally engraved on the lens mount.

Aperture. The aperture, or *f*-number, is the ratio of the focal length to the diameter of an incident light beam as it reaches the lens. For instance, if the focal length is 50 millimetres and the diameter of the incident light beam is 25 millimetres, the *f*-number is 2. This incident-beam diameter is often roughly the lens-diaphragm diameter, but it may be appreciably larger or smaller. The maximum

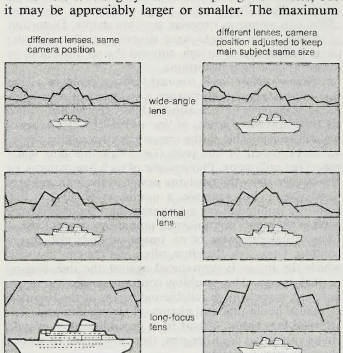


Figure 4: Effects of using lenses of different focal lengths.

aperture (*f*-number at the largest diaphragm opening) is also marked on the lens, usually in the form *f*:2, *ƒ*/2, or 1:2.

Angle of coverage. A lens must cover the area of a camera's film format to yield an image adequately sharp and with reasonably even brightness from the centre to the corners of the film. A normal lens should cover an angle of at least 60°. A wide-angle lens covers a greater angle—about 70° to 90° or more for an ultrawide-angle lens. A long-focus lens covers a smaller angle.

The angle of coverage depends on the lens design. Designations like "wide angle" or "narrow angle" are not necessarily synonymous with "short focus" and "long focus," as the latter terms refer to the focal length of the lens relative to the picture format.

Optical performance. A simple lens produces a very imperfect image, which is usually blurred away from the centre. The image may have colour fringes around object outlines, and straight lines may be distorted. Such defects, called aberrations, can be eliminated—and even then not completely—only by replacing the single lens element by a group of elements of appropriate shape and separation. Aberrations arising from some of the lens elements then counteract opposite aberrations produced by other elements. The larger the maximum aperture, the greater the angle of coverage, and the higher the degree of correction aimed at, the more complex camera lenses become. Lens design for relative freedom from aberrations involves advanced computer programming to calculate the geometric parameters of every lens element. Some aberrations can also be corrected by making one or more of the surfaces of a lens system aspheric; *i.e.,* with the variable curvature of a paraboloid or other surface rather than the constant curvature of a spherical one.

Groups of lens elements for correcting aberrations

Lenses usually consist of optical glass. Transparent plastics also have come into use, especially as they can be molded into elements with aspheric surfaces. They are, however, more sensitive to mechanical damage.

Aberrations. There are a number of lens aberrations, each with its own characteristics. Chromatic aberration is present when the lens forms images by different-coloured light in different planes and at different scales. Colour-corrected lenses largely eliminate these faults. Spherical aberration is present when the outer parts of a lens do not bring light rays into the same focus as the central part. Images formed by the lens at large apertures are therefore unsharp but get sharper at smaller apertures. Curvature of field is present when the sharpest image is formed not on a flat plane but on a curved surface. Astigmatism occurs when the lens fails to focus image lines running in different directions in the same plane; in a picture of a rail fence, for instance, the vertical posts are sharp at a focus setting different from the horizontal rails. Another aberration, called coma, makes image points near the edges of the film appear as irregular, unsharp shapes. Distortion is present when straight lines running parallel with the picture edges appear to bow outward (barrel distortion) or inward (pincushion distortion).

Resolving power and contrast-transfer function. One way of testing lens performance is to observe the image it forms of patterns of increasingly closely spaced black lines separated by white spaces of line width. The closest spacing still recognizable in the image gives a resolving power value, expressed in line pairs (*i.e.,* black line plus white space) per millimetre. Photographs of such line patterns, or test targets, show the resolving power of the lens and film combination. For example, a resolution of 80–100 line pairs per millimetre on a fine-grain film represents very good performance for a normal miniature camera lens.

The visual sharpness of an image depends also on its contrast. Opticians, therefore, often plot the contrast with which the image is reproduced against the line spacing of that image. The resulting contrast-transfer curve, or function, gives a more reliable indication of the lens performance under practical picture-taking conditions.

Special lens types. Apart from general-purpose camera lenses of various focal lengths, there are lenses of special characteristics or design.

Telephoto lenses. Long-focus lenses are bulky, because they comprise not only the lens itself but also a mount or tube to hold it at the appropriate focal distance from the film. Telephoto lenses are more compact; their combinations of lens groups make the back focus (the distance from the rear lens element to the film) as well as the length of the whole lens appreciably shorter than the focal length. Strictly, the term telephoto applies only to a lens of this optically reduced length; in practice long-focus lenses of all types tend to be called indiscriminately telephoto or "tele" lenses.

If a camera lens is interchangeable, an accessory teleconverter lens group can be positioned between the prime lens and the camera. This turns a normal lens into an even more compact telephoto system, which is less costly than a telephoto lens but which reduces the speed of the prime lens and usually impairs sharpness performance.

Tele-converter lenses

Wide-angle and retrofocus lenses. Short-focus, wide-angle lenses are usually mounted near the film. Single-lens reflex cameras need a certain minimum lens-to-film distance to accommodate the swinging mirror. Wide-angle (and sometimes normal-focus) lenses for such cameras therefore use retrofocus designs. In these the back focus is appreciably longer than the focal length. Both a telephoto and a retrofocus lens must be specially designed for its particular use to ensure optimum image performance.

Fish-eye lenses. For image angles greater than 110°, it becomes difficult to bring the lens close enough to the film to allow the rays between the lens and film to diverge sufficiently. The fish-eye lens overcomes this difficulty by making the rays diverge less behind the lens than they do in front. The resulting image shows appreciable distortion, with image details near the edges and corners progressively compressed. Fish-eye lenses usually cover angles between 140° and 210° and are used for unusual wide-angle effects where the distortion becomes a deliberate pictorial element. They also have certain scientific applications, for instance, to cover a horizon-to-horizon view of the sky in recording cloud formations.

Mirror lenses. Images can also be formed by light reflected from curved mirrors. This method, long used in astronomical telescopes, is applied to long-focus lens systems of short overall length by folding the light path back onto itself. A mirror lens or catadioptric system has no chromatic aberrations. Other aberrations are corrected by incorporating one or more appropriate lens elements. The arrangement of the system, with a central opening in the primary mirror, makes stopping down with a customary diaphragm difficult, and neutral-density filters are used to control light transmission.

Variable-focus lenses. In variable-focus lenses the focal length can be varied by movement of some of the elements or groups within the lens system. One lens can thus replace a range of interchangeable lenses.

The variable-focus, or zoom, lens was originally developed for motion-picture photography, in which adjustment of the focal length during a shot produced a zooming-in or zooming-out effect (hence the name). It is now widely used in single-lens reflex cameras where the reflex finder permits accurate continuous assessment of image coverage. In a true zoom lens the image changes in scale but not in sharpness during zooming; some varifocal lenses, however, need refocusing at different focal lengths. Due to correction requirements over a range of focal lengths, zoom lenses are complex systems containing from 12 to 20 elements. Zoom lenses for still cameras have focal-length ratios from 2:1 to 4:1 or more (*e.g.,* 35–135 mm for a 35-mm reflex).

Lens-changing systems. Miniature and roll-film cameras hold interchangeable lenses in screw or quick-change bayonet mounts. In a focal-plane shutter camera the usable range of focal lengths is practically unlimited. In cameras with leaf shutters, either the lens is mounted in front of the shutter or the lens is changed with the shutter. Some designs use convertible lenses with the rear components built into the camera together with the shutter; interchangeable front groups then provide different focal lengths in combination with the fixed rear group. View-camera lenses—usually with their own shutters—are mounted on lens boards that clip into and out of the front camera standard.

Convertible lens with fixed rear component

Afocal attachments provide the effect of alternative focal lengths with a fixed camera lens. They are magnifying or reducing telescopes without a focal length (hence afocal), yielding a virtual image that the camera lens projects onto the film. Their designated magnification factor indicates the effect on the image scale; *e.g.*, a 1.5× tele attachment magnifies the image on the film 1½ times, while a 0.7× wide-angle attachment reduces the image scale to 0.7 times that of the prime camera lens.

Lens coating. When light passes from one optical medium to another (especially from air to glass and vice versa in a lens), about 4 to 8 percent of it is lost by reflection at the interface. This light loss builds up appreciably in complex multielement lenses. Some of the reflected light still reaches the film as ghost images or light spots or as general contrast-reducing scattered light.

To reduce such losses, the air-to-glass surfaces of modern lenses typically carry a microscopically thin coating of metallic fluorides. The coating eliminates most reflected rays. Complete elimination can occur only for light of one wavelength if the coating thickness and refractive index are exactly right. In practice a coated lens surface reflects about 0.5 percent of incident white light—1/10 of the light lost by an uncoated lens. Multiple coatings can reduce reflections over a wider wavelength range.

BLACK-AND-WHITE FILMS

The latent image. The sensitive surface of ordinary film is a layer of gelatin carrying minute suspended silver halide crystals or grains (the emulsion)—typically silver bromide with some silver iodide. Exposure to light in a camera produces an invisible change yielding a latent image, distinguishable from unexposed silver halide only by its ability to be reduced to metallic silver by certain developing agents.

Formation of the latent image

Current theories postulate that silver halide crystals carry minute specks of metallic silver—so-called sensitivity specks—which amount in mass to about 1/100,000,000 part of the silver halide crystal. A silver halide is a compound of silver with fluorine, chlorine, bromine, or iodine, but only the last three are light-sensitive. When light action releases electrons from the silver halide crystal, they migrate to the sensitivity specks. The resulting electric charge on the specks attracts silver ions from the neighbouring silver halide; and as the silver ions accumulate, they become metallic silver, causing the speck to grow. Halogen (*e.g.*, bromine) atoms at the same time migrate to the surface of the silver halide crystal and are there absorbed by the gelatin of the emulsion. When the sensitivity speck is large enough, it provides a point of attack for the developer, which can then reduce the whole silver halide crystal to silver. Developers are selective organic reducing agents that attack only silver halide crystals that have sufficiently large sensitivity specks. The halide grains carrying a developable sensitivity speck make up the latent image.

Sensitometry and speed. The sensitivity or speed of a film determines how much light it needs to produce a given amount of silver on development. Sensitometry is the science of measuring this sensitivity, which is determined by giving the material a series of graduated exposures in an appropriate instrument (the sensitometer). After development under specified conditions, the density of the silver deposit produced by each exposure is measured and the densities are plotted on a graph against the logarithm of the exposure. The resulting characteristic curve, or *D*/log *E* curve (see below *Contrast*), shows how the film reacts to exposure changes. A specified point on the curve also serves as a criterion for calculating film speed by methods laid down in various national and international standards.

The internationally adopted scale is ISO speed, written, for example, 200/24°. The first half of this (200) is arithmetic with the value directly proportional to the sensitivity (and also identical with the still widely used ASA speed). The second half (24°) is logarithmic, increasing by 3° for every doubling of the speed (and matching the DIN speeds still used in parts of Europe). A film of 200/24° ISO is twice as fast (and for a given subject requires half as much exposure) as a film of 100/21° ISO, or half as fast as a film of 400/27° ISO.

All-around films for outdoor and some indoor photography have speeds between 80/20° and 200/24° ISO; fine-grain films for maximum image definition between 25/15° and 64/19° ISO; and high-speed and ultraspeed films for poor light from 400/27° ISO up.

Colour sensitivity. Initially, the silver halide emulsion is sensitive to ultraviolet radiation and to violet and blue light. Most films contain sensitizing dyes to extend their colour sensitivity through the whole visible spectrum. Such films, called panchromatic films, were introduced in 1904. They record subject colour values as gray tones largely corresponding to the visual brightness of the colours.

Non-colour-sensitized or blue-sensitive emulsions (without sensitizing dyes) are used for copying monochrome originals and similar applications needing no extended colour sensitivity. At one time orthochromatic films—sensitive to violet, blue, green, and yellow but not to red—were also used for general photography; now they are employed mainly for photographing of phosphor screens, such as cathode-ray tubes, and for other purposes requiring green but not red sensitivity.

Use of orthochromatic films

Infrared films, developed in 1919, are sensitized to invisible infrared wavelengths. They are used in aerial photography to cut through atmospheric haze (which scatters blue light but not infrared rays) and for special purposes in scientific and forensic photography.

Filters. Filters can modify the way in which a film records colours as monochrome tone values. They are disks of coloured glass or gelatin with controlled transmission characteristics. Placed in front of the camera lens, they preferentially transmit light of their own colour and hold back light of other colours. A yellow or yellow-green filter is often used in landscape photography to prevent overexposure of the blue sky and to bring out detail in cloud formations. Orange and red filters make the sky still darker and cut through haze by absorbing scattered blue light.

Contrast filters differentiate between the gray values of objects of different colour but of similar brightness. For instance, a red flower and green foliage record in similar shades of mid-gray. A red filter holds back green light to darken the green foliage, making the flower lighter; a green filter absorbs red light, thus darkening the flower. Such deliberate tone distortion is widely used in photomicrography and other fields.

Other filter types used in photography include ultraviolet, infrared, and polarizing filters. Ultraviolet-absorbing filters screen out ultraviolet rays at high altitudes (*e.g.*, in mountain photography). Because camera lenses are not normally corrected for such rays, the rays can reduce image sharpness, even though the lenses allow only a small amount of ultraviolet to be transmitted. Infrared filters are used with infrared film to hold back visible light. Polarizing filters polarize light and can absorb polarized light if suitably oriented. Light reflected at certain angles from shiny surfaces of nonmetallic media (glass, water, varnish) is polarized; a properly oriented polarizing filter subdues such reflections in a picture.

Because a filter screens out part of the light, its use calls for extra exposure, the amount of which is indicated by a filter factor—*e.g.*, 2×, which means the exposure time must be multiplied by 2. For cameras with an exposure-value scale, a filter may specify an exposure value reduction (such as −1 or −1½; *i.e.*, the indicated exposure value must be reduced by this amount). The factor of a given filter depends on the spectral sensitivity of the film, the colour quality of the lighting, the type of subject, the effect aimed at, and other exposure conditions.

Filter factors for proper exposure

Other film characteristics. Of practical interest to the photographer are the graininess, resolving power, and contrast of a film. Although they are characteristics of the film itself, they are influenced by the conditions of development (see below *Black-and-white processing and printing*).

Grain. The image derived from minute silver halide crystals is discontinuous in structure. This gives an appearance of graininess in big enlargements. The effect is most prominent with fast films, which have comparatively large silver halide crystals.

Resolving power and acutance. The fineness of detail

that a film can resolve depends not only on its graininess but also on the light scatter or irradiation within the emulsion (which tends to spread image details) and on the contrast with which the film reproduces fine detail. These effects can be measured physically to give an acutance value, which is preferred to resolving power as a criterion of a film's sharpness performance. Fine-grain films with thin emulsions yield the highest acutance.

Contrast. High-contrast films reproduce tone differences in the subject as great density differences in the image; low-contrast films translate tone differences into small density differences. The characteristic curve of a film obtained by plotting the density against the logarithm of the exposure (mentioned earlier under *Sensitometry and speed*) can be used to express a film's contrast (see Figure 5). The slope of the straight-line section of the

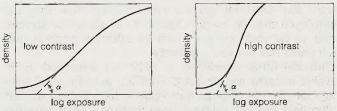


Figure 5: Characteristic curves of low-contrast and high-contrast film (see text).

curve (sometimes called the gamma, actually the tangent of the angle *a*) indicates contrast: the steeper the slope, the higher the contrast rendering. General-purpose films yield medium contrast (gamma 0.7 to 1). High-contrast films (gamma 1.5 to 10) are used for copying line originals and other specialized purposes; low-contrast films for continuous-tone reproduction. Gamma is also used to indicate degree of development, since increased development generally results in a higher gamma.

Film structure and forms. Film consists of a number of layers and components: (1) A supercoat of gelatin, a few micrometres (one micrometre is 0.001 millimetre) thick, protects the emulsion from scratches and abrasion marks. (Pressure and rubbing can produce developable silver densities.) (2) The emulsion layer (silver halide suspended in gelatin) is usually nine to 12 micrometres (up to $^1/_{2,000}$ inch) thick but may sometimes reach 25 micrometres. (3) A substrate or subbing layer promotes adhesion of the emulsion to the film base. (4) The film base, or support, is usually cellulose triacetate or a related polymer. The thickness may range from 0.08 to 0.2 millimetre (0.003 to 0.008 inch). Films for graphic arts and scientific purposes are often coated on a polyethylene terephthalate or other polyester support of high dimensional stability. Glass plates—once the most common support for negative materials—are now used only for applications requiring extreme emulsion flatness. (5) A backing layer on the rear of the film base counteracts curling. Usually it contains also a nearly opaque dye to suppress light reflection on the rear support surface. Such reflection (halation) reduces definition by causing halolike effects around very bright image points. Some film bases (especially in 35-mm films) are tinted gray to absorb light that has passed through the emulsion layer.

Use of anti-halation dye

Sheet film. View and studio cameras generally take sheet film—single sheets (typical sizes range between $2^1/_2 \times 3^1/_2$ and 8×10 inches) loaded in the darkroom into light-tight film holders for subsequent insertion in the camera.

Roll film. The term roll film is usually reserved for film wound up on a spool with an interleaving light-tight backing paper to protect the wound-up film. The spool is loaded into the camera in daylight, the backing paper leader threaded to a second spool, and the film wound from picture to picture once the camera is closed. This is the classical roll film of roll-film cameras. Common current film widths are 62 mm and 45 mm. The rear of the backing paper carries sets of consecutive numbers spaced at frame intervals for different image formats. In some roll-film cameras these numbers are visible through a viewing window in the camera and show how far the film must be wound to advance it from one picture to the next. Instant-loading cartridges also use paper-backed roll film.

Perforated film. Some film is perforated along its edges and rolled up on its own inside a light-tight cartridge, which can be loaded into the camera in daylight. Once the camera is closed, a transport sprocket engaging the edge perforations draws the film from the cartridge onto a spool and advances it from picture to picture. The most common film width is 35 mm (for 35-mm miniature cameras), and its cartridge typically holds enough film for up to 36 (sometimes 72) exposures. A 70-mm film for larger cameras and 16-mm strips for ultraminiatures are packed and used in a similar way.

In March 1983 the Eastman Kodak Company announced the development of a new coding system for 35-mm film and cartridges. The DX film system employs optical, electrical, and mechanical encoding to transmit to appropriately equipped cameras such information as film type, film speed, and number of exposures. The system also supplies data that enable automatic photofinishing equipment to identify and sort film quickly, simplifying processing and printing. In the interest of uniformity, Kodak freely offered the DX system to all film and camera manufacturers, and within two years it was generally adopted.

Disk film. Some compact mass-market cameras take circular disks of film, 65 millimetres in diameter, in light-tight cartridges and coated on a 0.18-mm polyester base. In the camera the disk rotates as up to 15 exposures (frame size 8×10 millimetres) are recorded around the disk circumference. The disk lies flatter in the camera than rolled-up film and is suitable for more automated photofinishing; the high printing magnification required, however, limits the image quality.

PICTURE-TAKING TECHNIQUE

The main areas of practical camera handling in photography concern sharpness control, exposure, and lighting.

Sharpness control. The image on the film is sharpest when the lens is focused to the exact object distance. Usually, however, a scene includes objects at varying distances from the camera. Various factors affect the sharpness distribution in a picture of such a scene.

Depth of field. The sharpness in the image of objects in front of and behind the focused distance falls off gradually. Within a certain range of object distances this sharpness loss is still comparatively unnoticeable. This range is the depth of field and depends on: (1) the amount of sharpness loss regarded as acceptable: miniature negatives requiring big enlargement must be sharper than larger format negatives, which are enlarged less; (2) the lens aperture used: stopping down the lens (higher *f*-numbers) increases the depth of field; (3) the object distance: the depth of field is smaller for near objects than for more distant ones; and (4) the focal length of the lens: depth of field is reduced with longer focus lenses (and with larger picture formats requiring lenses of longer focal length), and the depth increases with shorter focus lenses. A depth of field indicator, often included on the focusing mounts of lenses, shows on the distance scale how far in front of and behind the focused distance objects will be in focus at different diaphragm openings.

Factors influencing depth of field

Subject and camera movement. Movement of the subject while the camera shutter is open for the exposure leads to a blurred image. The exposure time must therefore be short enough to keep the blur within acceptable limits. The shutter speed required depends on the movement speed of the object, the scale of the image (movement blur becomes greater the nearer the subject or the longer the focal length of the lens used) and the movement direction; movement across the direction of view produces the most blurring.

Movement blur can be reduced, even with comparatively slow shutter speeds, by moving the camera (panning) to follow the subject during the exposure. This records the moving object comparatively sharply against a blurred background and emphasizes the impression of speed.

Camera shake through unsteady support during the exposure also creates image blur—over the whole picture in such cases. Hand-held shots generally demand shutter speeds of $^1/_{30}$ second or shorter. For longer times a firm camera support—such as a tripod—is essential.

Exposure technique. The correct exposure (aperture and shutter settings) can be derived from tables or calculators or by direct measurement of the subject luminance with a light meter.

Automatic meter control. Cameras with through-the-lens (TTL) exposure meters—and also hand-held meters pointed at the subject—measure the average reflected light intensity, yielding reliable exposures for subjects of average contrast and brightness distribution. Subjects of extreme contrast or very bright or dark dominant areas need overriding exposure corrections; automatic cameras often have provision for this. Such a TTL measurement is usually centre-weighted (predominantly based on the image centre). Some cameras (and meters) permit spot readings covering a small subject area only and give reliable exposures if this selected area is a medium subject tone.

The selection of an appropriate aperture and shutter speed among equivalent camera exposures depends on depth-of-field and subject-movement requirements. Some automatic cameras simplify this by selecting just one such combination at each exposure level (program automation).

Flash exposures. Most current electronic flash units incorporate a sensor cell that measures the light reflected from the subject and controls the flash duration (and hence the exposure) accordingly. In certain cameras in which photocells measure the light reflected from the film, the same cells can similarly control the flash duration of suitable dedicated flash units. Lacking these provisions, flash exposures may be determined by measurement or by guide-number calculation.

Special meters can measure flash light quantity on a scene during a test firing of flashes; these are used extensively with more elaborate studio setups.

Flash exposure calculations rely on the fact that the exposure depends only on the lens aperture. (The electronic flash is usually much shorter than the synchronizable shutter time.) The light intensity reaching the film is inversely proportional to the square of the diaphragm *f*-number. By basic illumination laws the light intensity on a scene is also inversely proportional to the square of the distance between the light source and subject. For a given flash source and film speed, the exposure is thus constant for a constant product of distance and *f*-number. Flash manufacturers quote this product as a guide number for various flash–film combinations. For rapid exposure calculation, dividing the guide number by the flash-to-subject distance gives the required *f*-number; dividing the guide number by the *f*-number gives the distance at which the flash must be arranged for correct exposure.

Lens aperture as the key to flash exposure

Some cameras use this principle for semiautomatic flash-exposure control: the aperture adjustment is coupled with the distance setting on the lens (or with an automatic rangefinding system) so that the lens aperture gets larger with increasing distance. This coupling is adjustable for different flash guide numbers.

Exposure latitude. The ideal negative exposure records the darkest subject shadows as a just visible density. More exposure yields a denser negative, which, however, can still give an acceptable print by appropriate print-exposure adjustment. This range of usable negative exposures, the exposure latitude, depends on the film and the subject. This latitude is greater the lower the subject contrast and the greater the film's exposure range (and, generally, the lower the film contrast). Because of exposure latitude, simple cameras with limited exposure adjustability can still yield acceptable pictures under differing light conditions.

Lighting technique. The kind of lighting on the scene governs the way in which the picture reproduces the subject. Orientation of the subject—as in taking a portrait—with respect to the light direction can often control the effect. Lighting from behind the camera gives flat effects, light from one side yields depth and modeling, while the principal light from behind the subject produces dramatic against-the-light effects of high contrast. Artificial light setups in the studio, with tungsten lamps or electronic flash, offer the greatest flexibility. Under such conditions the photographer can arrange two or more lamps for various lighting effects.

Flexibility with studio lighting

Directional lighting improves detail contrast and brilliance. Excessive subject contrast, however, makes accurate exposure settings difficult and may lead to loss of picture detail in the highlights or shadows. Fill-in lighting, by a flash or other light source on or near the camera, can illuminate heavy shadows facing the camera.

BLACK-AND-WHITE PROCESSING AND PRINTING

Negative development. Amateurs usually process films in developing tanks. In this type of development roll or miniature film is wound around a reel with a spiral groove, which keeps adjacent turns separated and allows access by the processing solutions. Once the tank is loaded (in the dark), processing takes place in normal light, the processing baths (developer, intermediate rinse, fixer) being poured into the tank at the appropriate intervals. Sheet films are similarly treated in small tanks or held in hangers and immersed sequentially in the different processing solutions. Large-scale commercial processing laboratories use machines that automatically feed the films through the solutions in proper sequence.

Developers and their characteristics. The developer consists typically of one or more developing agents, a preservative (such as sodium sulfite) to prevent oxidation by the air, an alkali (such as sodium carbonate) to activate the developer, and a restrainer or antifoggant to ensure that the developer acts only on exposed silver halide crystals. A developer's main characteristics are activity, development speed, and effect on film gradation, graininess, and sharpness. Developers may be prepared on the basis of published formulas or bought as ready-mixed powders or concentrates for dilution with water.

The developer is allowed to act for a specific time to build up the image to the required density and contrast. This time depends on the developer, the temperature, the degree of agitation, and the film—as indicated by recommendations from film and developer manufacturers.

Fixing. The fixing bath contains a chemical (sodium or ammonium thiosulfate) that converts the silver halide into soluble, complex silver salts that dissolve in the fixer. During this process the film loses its original silver halide milkiness overlaying the image and becomes clear. The fixer also contains a weak acid (to halt the development process) and a hardening agent to reduce gelatin swelling.

Washing and drying. Washing removes all residual soluble chemicals from the emulsion and must be thorough for image permanence. Films are hung up to dry after removal from the tank.

High-speed processing. Greatly reduced processing times are possible with high-activity developers at elevated temperatures and with fast-acting fixing agents, such as ammonium thiosulfate. Such processes can cut access time to the negative down to less than a minute. One-bath (monobath) processing in a solution containing both a fast-acting developing agent and fixing chemicals also reduces processing time. In special rapid-access processing equipment, films pass through chambers spraying the processing solutions onto the film surface or run in contact with monobath-soaked webs.

Monobath processing with developer and fixer

Printing. The simplest printing equipment is the contact printing frame in which the negative and printing paper are held together behind a glass plate during exposure to a suitable lamp. A printing box is essentially a printing frame with a built-in light source. Contact printing gives a positive of the same size as the negative.

Enlargers. Negatives usually are enlarged to prints of the desired final size. The enlarger is a projection system on a vertical column mounted on a horizontal baseboard. It has a lens, a film holder (negative carrier), and a lighting system (typically a lamp and condenser lens) for illuminating the negative. Raising or lowering the enlarger head on the column controls the image magnification; adjustment of the lens-to-negative distance focuses the image on the enlarging paper on the baseboard. In enlargers that focus automatically these two adjustments are linked mechanically to keep the image sharp all the time. Enlargers are made in various sizes to take different maximum negative formats.

Printing papers. Papers for enlarging and contact printing are produced in grades of differing exposure range—

i.e., ratios of shortest to longest exposure to produce the lightest tone and a full black, respectively. The various grades yield prints of a normal tone range from negatives of different contrasts: a soft paper grade for a high-contrast negative, a normal paper for a normal negative, a hard paper for soft negatives, and so on. Paper grades are also numbered—typically from 0 to 5—in ascending order of contrast. Variable-contrast papers use a mixture of two emulsions of a different contrast and colour sensitivity; the contribution of each is controlled by filters in the path of the exposing light.

Paper grades in order of contrast

Other characteristics of printing papers are the speed (slower for contact papers, faster for enlarging papers), image colour (blue-black to warm brown), surface texture (glossy, velvet, mat), and base thickness (single or double weight). Most printing materials use a resin-coated (plastic-laminated) paper base that absorbs no water during processing.

Printing exposures. Correct printing exposures are determined by trial and error or by test strips given a series of progressively increasing exposures. More sophisticated exposure control systems measure either the brightness of selected image portions projected on the enlarger baseboard or the average light intensity reaching the paper during the exposure. Printing papers are exposed and processed in a darkroom lit by an olive-green or orange safelight. Printing papers are sensitive to violet, blue, and sometimes green light.

Print processing. The processing of prints consists of development, an intermediate rinse or stop bath, fixing, and washing. The developer and fixer are similar in principle to those used for negative films. In the normal method, dish or tray processing, prints are immersed successively in the solutions in dishes laid out side by side. Development is checked visually, the print remaining in the developer until the image has reached its full density. For drying, the prints may be clipped to a line, placed in a heated print dryer, or squeegeed onto a mirror-finished plate for a high-gloss surface.

Stabilization processing. Certain rapid-processing papers incorporate developing agents in their emulsions and are processed on a roller processor. This processor runs the paper through an activating bath for instant development and then through a stabilizing bath, followed by a pair of squeegeeing rollers from which the print emerges merely damp. This process takes about 10 to 15 seconds; the prints, however, do not keep quite as well as conventional prints, since unexposed silver salts are not removed from the emulsion but only converted into moderately light-stable compounds. Such prints can be made more permanent by subsequent fixing and washing.

Disadvantage of prints finished by stabilization processing

Dry processing. Processing baths can be completely eliminated by incorporating in the emulsion of the paper development and stabilization chemicals that become active on heating. One method is to disperse the processing chemicals in the emulsion in microscopic capsules containing the solution and a blowing agent. On passing the exposed paper over a heated roller, the blowing agent bursts the capsules, and the liberated processing solutions act on the silver halide immediately around each capsule. The liquid solvent instantly evaporates, leaving a dry print. Encapsulation materials are used for such purposes as making proof prints of negatives and reenlarging microfilm images. Certain non-silver processes in photocopying systems also offer dry processing.

COLOUR PHOTOGRAPHY

Colour reproduction. Present-day colour photographic processes are tricolour systems, reproducing different colours that occur in nature by suitable combinations of three primary-coloured stimuli. Each of these primary colours—blue-violet, green, and red—covers roughly one-third of the visible spectrum. Tricolour impressions can be produced by combining coloured lights (additive synthesis) or by passing white light through combinations of complementary filters, each of which holds back one of the primary colours (subtractive synthesis).

In additive synthesis a combination of red and blue-violet light (*e.g.*, light beams of the two colours directed on the same spot of a white screen) gives a purplish pink (magenta); equal parts of red and green produce yellow, and equal parts of green and blue-violet produce bluish green (cyan). Superimposition of all three light beams on a screen yields white; combinations of varying proportions of two or three of the colours produce virtually all the other hues.

In subtractive synthesis yellow, magenta, and cyan filters or dye layers subtract varying proportions of the primary colours from white light. The yellow filter absorbs the blue component of white light and so controls the amount of blue present in a white-light beam that has passed through the filter. Similarly, the magenta filter controls the amount of green light left, and the cyan controls the amount of the red component. A cyan and a magenta filter superimposed in a white-light beam hold back both the red and the green component, making the emerging beam blue. Similarly, a cyan and a yellow filter together yield green, and a yellow and a magenta filter together yield red. Superimposing such filters or dye images of different densities in a white-light beam can therefore re-create any colour impression in the same way as superimposing light beams of the primary colours.

Subtraction of primary colours by filters or dye layers

The difference between additive and subtractive synthesis is the approach: in additive synthesis colours are built up by combining different intensities of primary-coloured light, and in subtractive synthesis colours are achieved by removing different proportions of primary-coloured light from white light. Most modern colour films are based on subtractive synthesis. Either method of colour synthesis should be capable of reproducing every existing colour in nature. In practice, the reproduction is imperfect; no filter dyes meet the required ideal specifications. Nevertheless, for most purposes reproduction is adequate.

Colour films. *Reversal (slide) films.* To reproduce colour by subtractive three-colour synthesis (Figure 6), colour films first break down the colours of an image into their primary components by means of three separate sensitized layers, each of which responds exclusively to blue, green, or red light. The image in each layer is reversal-processed to yield a positive dye image in a colour complementary to the layer's spectral sensitivity. Thus, the blue-sensitive layer first yields a negative image of everything blue in the original scene (*e.g.*, the blue sky) and then a positive image of everything that is not blue. This positive image is coloured yellow. Similarly, the green-recording layer yields a magenta positive image of everything that is not green, and the red-recording layer a positive cyan image of everything that is not red. Blue sky, for instance, does not figure in the yellow positive image but does figure in the magenta positive image (not being green) and in the cyan positive image (not being red). The magenta and cyan dyes in the areas that were blue sky are superimposed, and white light passing through the resulting transparency loses its green and red, but not its blue, component; thus, the sky appears blue. Similarly, green subject components end up as positive yellow image density in the blue-recording and positive cyan density in the red-recording layer, combining to green in the transparency. Yellow records as a negative image in the green-recording and red-recording layers, hence leaving a positive yellow image only in the blue-recording layer. All other colours are formed by similar combinations of different densities of the dye images.

Negative (print) films. Negative colour materials work

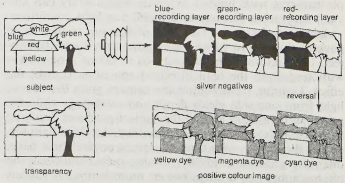


Figure 6: Colour reproduction sequence with subtractive reversal film (see text).

in a similar way but yield a negative dye image by direct development. Blue subject tones record in the blue-sensitive film layer to produce a yellow negative image. Green colour components yield a magenta dye image in the green-responding layer, and red components yield a cyan dye image in the red-recording layer. With respect to the subject, the colour negative therefore reverses the tones in brightness as well as in colour. Printing the colour negative on a colour paper with three differentially responding layers reverses the process once more, reconstituting the original subject colours in a positive print.

Colour-film structure. Reversal colour film has these components: (1) A top layer of plain gelatin, which protects the underlying layers against abrasion and damage. (2) The first emulsion layer, which contains blue-sensitive silver halide plus a yellow-forming colour coupler. This is a colourless substance that reacts with the decomposition products of the developing agent to generate dye in all areas where a silver image is produced and in proportion to the density of that silver image. (3) A yellow filter layer, which holds back blue light from the subsequent emulsion layers. It disappears during the bleaching stage of processing. (4) The second emulsion layer, which contains blue- and green-sensitive silver halide plus a magenta-forming colour coupler. The blue sensitivity is suppressed by the yellow filter layer. (5) The next emulsion, which is blue-and-red sensitive (blue again being suppressed) and contains a cyan-forming colour coupler. (6) A substrate, which ensures optimum adhesion of the emulsion layers to the film base and may also contain light-absorbing silver to prevent the scattering of light by reflection from the support surface (halation). (7) The film base, or support, of clear cellulose acetate derivative (or sometimes polyester), typically about 0.005 inch thick. (8) The back of the support, which carries a light-absorbing layer (an alternative to the antihalation layer in the substrate); on roll film this also acts as an anticurl layer.

Variations in reversal colour films

This scheme can vary. Often one or more of the selectively sensitized layers is duplicated. Some reversal colour films do not incorporate colour couplers in the emulsions but introduce them in the processing solutions. Processing such nonsubstantive colour films is more complex than processing substantive-coupler films containing couplers in the emulsions. Negative colour films are similar, but the couplers are often already coloured yellow or red or both. The unreacted couplers remain as positive images, which compensate for some deviations of the image dyes from the colour characteristics of ideal dyes. Such negatives have an apparently overall reddish or orange tint.

Some colour films come with different combinations of colour sensitivity and dye colour formed in the individual layers for deliberately falsified colour rendering (false-colour films) in special applications.

Colour-positive (print) materials have a paper or white, opaque film base instead of transparent film and have no antihalation layer. The emulsion sequence may be different from the above scheme; their spectral sensitivities may be keyed to the transmission characteristics of the negative dyes for better colour reproduction.

The dye images are reasonably lightfast but fade on prolonged exposure to ultraviolet-rich radiation. Colour transparencies and prints intended for continuous display may be protected by ultraviolet-absorbing coatings or filter layers.

Additive colour films. These are simpler in structure and consist, in addition to protective and other interlayers, of a film base, carrying a filter raster, and a black-and-white emulsion layer. The raster consists of sequences of very narrow red, green, and blue transparent filter lines (up to 1,800 lines per inch) through which the light from the lens passes before it reaches the emulsion layer. The emulsion layer is processed to a positive image. Red subject portions cause silver to be deposited behind non-red (*i.e.*, green and blue) filter elements, leaving the red filter lines transparent. Similarly, green subject details leave green filter lines transparent but block red and blue. Other colours affect areas behind two or even three filter lines—for example, yellow leaves red and green filter lines clear. In such areas the eye cannot resolve the separate filter elements but gets an additive impression of yellow. Other colours form corresponding additive effects, including white, where all three filter elements are transparent. Because of the presence of the filter elements everywhere in the image, additive colour transparencies are much denser than subtractive ones; at high magnification the filter raster pattern may also become visible. Additive colour transparencies are used only in rapid-access diffusion-transfer systems (see below *Instant-picture photography*).

Colour balance. Colour film reacts to all hue and tone differences, including the prevailing light colour. A film recording approximately natural colours in daylight reproduces scenes photographed by tungsten light with a reddish overall tint—because this lighting is richer in red rays than is daylight. This spectral balance of different "white" light sources may be rated numerically by the colour temperature—a concept of theoretical physics that, with tungsten lighting, corresponds roughly to the absolute lamp-filament temperature. Such absolute temperatures are expressed in kelvins (K). The higher the colour temperature the richer the light is in bluish and the poorer it is in reddish rays and vice versa. Average daylight is rated at about 5,500 K, the light from an overcast sky from 6,500 K up; the colour temperature of tungsten lamps ranges between 2,600 and 3,400 K.

To ensure correct "white-light" colour reproduction with different types of lighting, the sensitivities of the three film layers must be matched to the colour temperature of the light. Colour slide (reversal) films are therefore made in different versions balanced for faithful rendering either with 5,500 to 6,000 K light sources (such as daylight or electronic flash) or with specified tungsten lighting (3,200 to 3,400 K).

Such accurate film balance matching is less vital with negative colour films since the colour rendering of the print can be modified during colour printing. "Universal" amateur negative colour films are usable with any light, from tungsten to daylight. For high quality, professional negative colour films are still preferentially balanced to either daylight or tungsten sources.

"Universal" colour films

Strongly coloured filters are suitable only for special effects; they overlay the colour image with the filter colour. Pale correction filters can match a film to a light source other than that for which it is balanced—*e.g.*, pale blue, with a daylight-type film used in tungsten lighting, to raise in effect the colour temperature. Pale pink or amber filters similarly reduce the colour temperature for using artificial-light-balanced films in daylight. Colour-film manufacturers publish detailed recommendations of actual filters required for such conversion.

In outdoor photography, especially involving distant views, an ultraviolet-absorbing filter is often required, as ultraviolet radiation records in the blue-sensitive layer of the film, producing an overall blue cast in the transparency. A pale pink skylight filter for outdoor subjects lit only by skylight counteracts the cold, bluish colour rendering resulting from such illumination.

Colour-film processing. The processing sequence for colour materials is longer than for black-and-white films and requires more solutions. Development needs very precise timing and temperature control. Colour films can be processed in amateur developing tanks; professionals use sets of tanks in temperature-controlled water jackets with provision for standardized solution agitation.

Reversal colour-film processing. Most colour films use a standard processing sequence and chemistry (usually available in kits). For substantive films (incorporating couplers in the emulsion) the sequence comprises: (1) development to form a negative silver image in each emulsion layer; (2) a reversal bath that renders developable the remaining silver halide in each emulsion layer; (3) colour development to produce a positive silver image in the remaining silver halide plus a coincident dye image by reaction with the colour couplers; (4) bleaching and fixing to reconvert the negative and positive silver images into silver halide and to dissolve the latter out of the emulsion, leaving only the three dye images; (5) a final rinse and stabilizer to remove soluble chemicals and improve light-fastness of the dyes; and (6) drying. There are also intermediate rinse stages.

The complete sequence without drying takes a little longer than 30 minutes.

Processing of colour films lacking couplers

Processing of nonsubstantive colour films, in which the couplers are in the colour developer, is more complex because each emulsion layer is reexposed by appropriately coloured light and colour-developed separately. This operation requires automated processing machinery.

Negative colour processing. Negative colour films are practically all of the substantive-coupler type. Most again follow a standard processing sequence consisting of colour development (forming a negative silver image in each emulsion layer together with a corresponding dye image), a rinse, and a bleaching and fixing stage to convert the silver image into silver halide and dissolve that (plus residual halide) out of the emulsion. A final rinse and drying conclude the process, which, excluding drying, takes about 12 minutes. Substantially the same procedure is followed for processing positive colour papers.

Colour printing. Colour print processing may be done in dishes or trays or in light-tight drums that are rotated manually or mechanically, processing solutions being poured in and out in succession. Professional colour laboratories use more elaborate versions of such rotating drum systems or roller or other automated machines that transport prints through the different solutions in turn.

Methods of obtaining colour control

Positive prints from colour negatives. Positive prints may be obtained from colour negatives by enlarging the colour negative onto a positive colour paper. Colour control consists of modifying the colour of the printing light by yellow, magenta, and cyan filters (typically by inserting high-density filters of these colours to varying degrees in the light path) to obtain a print of the correct or desired colour balance. The light is thoroughly mixed in a diffusing box before reaching the negative. An alternative method is to have three light sources behind the yellow, magenta, and cyan filters and to adjust their relative intensities or switch them on for different exposure times.

This subtractive, or white-light, printing method depends on subtracting or holding back colour components of white light. Commercial photofinishing printers often use an additive system in which prints are given successive exposures through high-density red, green, and blue filters. Each of these exposures forms the image in one of the emulsion layers of the paper; colour balance depends on the proportions of the individual exposures. In automatic colour printing systems the exposures are controlled by photocells that evaluate the red, green, and blue components of light transmitted by the negative.

Reversal colour printing. Colour transparencies can be printed on a reversal colour paper similar to a reversal film and processed in an analogous way. The same kind of colour control with filters is again possible, but the colour effect of the subtractive filters or of the additive filter exposures is reversed.

Dye-destruction processes. Dye-destruction processes differ from chromogenic colour materials (where colour images are produced during development) in starting off with emulsion layers containing the final dyes. During processing these are bleached in proportion to the silver image formed. Straightforward processing of a dye-destruction or dye-bleach material yields a positive image from a positive original and consists of: (1) development to form a silver image; (2) stop-fixing to arrest development and remove unexposed silver halide; (3) dye bleaching to bleach the dye in the areas containing a silver image; (4) silver bleaching to convert the silver image into silver halide; and (5) fixing to remove residual silver halide. Washing is done between all the processing stages.

Obtaining a positive image from a negative requires a more elaborate processing sequence, analogous to reversal processing in a chromogenic system. Dye-bleach materials use far more light-stable dyes than those produced by colour-coupling development. The positive–positive procedure also yields duplicate transparencies on dye-bleach materials with a transparent film base.

Diffusion-transfer colour prints. Materials derived from instant-picture diffusion-transfer processes (see below *Instant-picture photography*) have been adapted to colour print production. They are more expensive than traditional colour print materials but considerably easier to process. In their simplest form they require only a single highly alkaline activator bath followed by a water rinse, the whole sequence lasting about five to 10 minutes, with considerable processing latitude. Such materials exist for prints from either negatives or transparencies. The colour printing and filter control principles are the same as with the traditional processes described above.

Assembly colour prints. The original method of producing colour prints was based on separation negatives obtained by photographing the original scene on separate black-and-white plates or films through a blue, green, and red filter, respectively. This method analyzes the subject in terms of its tricolour components in the same way as the initial negative images in a three-layer colour film. Positive prints from the separation negatives, converted into colour images (*e.g.,* by toning) and superimposed on top of each other, yield a subtractive tricolour print.

Prints by the dye transfer process

The main surviving assembly print process, dye transfer, uses a set of separation positives on a panchromatic matrix film made either from separation negatives of a colour transparency or by separation (three filtered exposures) from a colour negative. Appropriate processing converts the matrix film into a gelatin relief image whose depth is proportional to the positive silver density. Each matrix is soaked in a dye solution (yellow for the matrix derived from the blue-filter negative and so on), and the dyes from the matrices are transferred in succession to a single sheet of gelatin-coated paper. Elaborate care is required to ensure accurate superimposition (registration) of the dye images; the result is a positive colour print.

Transparency projection. Many amateur colour pictures are in the form of transparencies, particularly on 35-mm film. These are usually mounted in plastic or card frames or bound between glass for projection on a screen in a darkened room. The projector consists of a lens, a holder for the slide, and a lighting system (lamp, reflector, and condenser lenses to concentrate the light onto the slide). Modern slide projectors take the slides in magazines or trays holding 30 to 50 or more slides. An automatic slide transport feeds each slide from the tray into the light path of the projector and may be operated from a remote control unit or by pulses from a tape recorder, which can also record a commentary to the complete slide series. Some projectors feature remote-controlled or automatic focusing to keep each successive slide image sharp on the screen.

The standard miniature slide size is 2×2 inches for transparencies up to $1^{5}/_{8} \times 1^{5}/_{8}$ inches; the most usual transparency format in such slides is 24×36 millimetres. Projectors for larger slides (*e.g.,* $2^{3}/_{4} \times 2^{3}/_{4}$ inches for transparencies up to $2^{1}/_{4} \times 2^{1}/_{4}$ inches) and ultraminiature projectors (*e.g.,* 10×14-millimetre transparencies in 3×3-centimetre slide frames) are suitably scaled-up or scaled-down versions of the standard models.

The size of the projected image depends on the distance of the projector from the screen and the focal length of the projection lens. Projectors may also project from the back onto a translucent screen; such rear-projection setups are more compact, and the image is often bright enough for viewing in daylight. The rear-projection system is used in schools and for commercial displays. Elaborate slide shows are produced by linking two or more projectors aimed at the same or adjacent screen areas. With a suitably assembled slide set, the pictures can be made to change, overlap, and assemble, according to a predetermined program.

INSTANT-PICTURE PHOTOGRAPHY

History and evolution. Cameras with built-in processing facilities, to reduce the delay between exposure and the availability of the processed picture, were proposed from the 1850s onward. The ferrotype process later adapted for "while-you-wait" photography by itinerant street and beach photographers goes back almost as far. Because of the messiness of handling liquid chemicals in or just outside the camera, such systems remained largely impractical. In the 1940s Edwin H. Land, a U.S. scientist and inventor, designed a film configuration that included a sealed pod containing processing chemicals in a viscous jelly or paste form to permit virtually dry processing inside

the camera and yield a positive print within a minute or less of exposure. Land demonstrated (1947), and through his Polaroid Corporation marketed (1948), a camera and materials that realized this system. It used a positive sheet and negative emulsion, the latter being discarded after use. An instant-print colour film (Polacolor) was introduced in 1963 and an integral single-sheet colour film in 1972. After the mid-1970s other manufacturers offered similar instant-print processes. In 1977 Polaroid introduced an 8-mm colour movie film, and in 1982 it introduced still transparency films that permit rapid processing outside the camera.

The Polaroid process

Black-and-white diffusion transfer. The Polaroid process is based on negative paper carrying a silver halide emulsion and a nonsensitized, positive sheet containing development nuclei. After the exposure the two sheets are brought into intimate contact by being pulled between a pair of pressure rollers. These rupture a sealed pod (attached to the positive sheet) to spread processing chemicals—in the form of a viscous jelly—between the two sheets. This reagent develops a negative image and causes the silver salts from the unexposed areas to diffuse into the positive layer and deposit metallic silver on the development nuclei. After about 30 seconds to one minute the negative and positive sheets are peeled apart and the negative can be discarded. In special versions of the process the negative may be washed and treated to give a conventional negative for normal enlarging.

In the original Polaroid instant-picture process the material was a dual roll of negative and positive sheets. Later versions of this peel-apart process use film packs and sheet films. They require special cameras incorporating the pressure rollers that operate the spread of processing jelly while the peel-apart sandwich is fed out of the camera. Special camera backs with this mechanism allow the use of Polaroid materials in professional cameras taking interchangeable film holders or magazines. Peel-apart Polaroid systems include high-speed emulsions, high-contrast, process, transparency, and scientific materials.

Silver diffusion-transfer processes were invented in 1939 in Belgium and Germany and were used for a number of years in office copying systems until superseded by dry copying processes.

The Polacolor process. Polaroid colour film has a larger number of active layers, including a blue-sensitive silver halide emulsion backed by a layer consisting of a yellow dye–developer compound, a green-sensitive layer backed by a layer of magenta dye–developer, and a red-sensitive layer backed by a cyan dye–developer. The dye–developer in each case consists of dye molecules (not colour couplers) chemically linked to developing agent molecules.

After exposure and activation by the alkaline jelly, the dye–developer molecules in each layer migrate into the adjacent silver halide layer. Development of exposed silver halide to a negative image anchors the dye–developer molecule in position. Dye–developer molecules in unexposed image areas are not used up by development but migrate into the receiving layer of the positive material. There they are immobilized, remaining as dye images corresponding to a positive of each silver halide layer in the negative film. The dyes thus re-create a full-colour positive image. The process depends on the controlled diffusion of the dye–developer molecules, achieved by spacing layers and balanced exposure and development time. Developing takes about one minute. Polacolor films include an 8×10-inch material for regular studio and view cameras (with separate processing machinery) and giant formats of 20×24 inches or even larger for special cameras.

Single-sheet process. The Polaroid single-sheet, or integral, films contain all the negative and positive layers in a single preassembled film unit that is exposed through the transparent positive layer. The unit incorporates a viscous processing reagent that acts in principle similarly to the chemistry of the Polacolor process. It includes "opacifying" dyes and a highly opaque white pigment that together protect the negative layers against light during processing outside the camera. The pigment provides a background to the positive image after the dye–developer molecules from the negative layers have migrated into the receiving layer. Other constituents of the system neutralize residual active chemicals after processing, for all chemistry remains within the single-sheet print. The print size is about $3\frac{1}{2} \times 4\frac{1}{4}$ inches, the effective image size about $3\frac{1}{8} \times 3\frac{1}{8}$ inches. The Eastman Kodak and Fuji Photo Film companies also have marketed single-sheet films and cameras that accept each other's films. These materials and cameras are not compatible with the Polaroid products.

Autoprocess materials. Because it requires cameras or camera backs with integral processing facilities, the instant-picture process is not suitable for conditions precluding immediate processing of the picture (*e.g.*, in underwater or space photography), nor is it suitable for motion-picture or 35-mm cameras. Alternative procedures suggested to overcome this usually involve some form of semidry rapid-access processing. The Polaroid Autoprocess system uses 35-mm film in standard cartridges to fit any 35-mm camera. After exposure the film is driven through a table-top processor, which sandwiches the film with a stripping film carrying a thin layer of processing fluid. The latter processes the negative image, causes the formation of a positive image by a diffusion-transfer process, and then releases the negative layers, which are finally removed from the film (together with residual chemicals) by the stripping material. The transparencies remaining on the 35-mm film are immediately ready for viewing and projection. Black-and-white as well as colour systems (by an additive process) are available in this form.

Applications. Instant-picture processes have an advantage in applications that need quick access to a finished print. The initial field of the process was amateur snapshooting and instant portraits, from which evolved the taking of identification pictures for work and security passes. Such passes are made with special cameras that record a portrait together with personal details on a composite print that is then laminated to form a tamper-proof identity card. In studio photography instant prints provide a quick method of making exposure tests and checking the effect of lighting. Large- and giant-format Polacolor prints are used in studio portraiture; normal instant prints have numerous commercial applications. Instant pictures are also widely used in the laboratory to record experimental setups, for photomicrography and for infrared photography; for instant endoscopy and for clinical and forensic records; for rapid copying of normal colour transparencies; and for instant hard copy of oscilloscope, video, and computer graphic displays. Autoprocess transparencies are used for the rapid production of colour or black-and-white slides for lectures and publication and in various fields of scientific photography (including photomicrography) relying on the use of conventional 35-mm (usually single-lens reflex) cameras.

SPECIAL PHOTOSENSITIVE SYSTEMS

The high working speed (efficiency of converting light into permanent images) of silver halides makes them almost the only materials suitable for camera use. Numerous light-sensitive systems not using silver have been known since the beginning of photography. In view of silver's high price, a number of substitute systems have grown in importance, and new ones have appeared. Most of them are limited to office copying, microfilming, the graphic arts, and other applications in which flat copy is reproduced.

Electrophotography. Electrophotography covers a number of processes that rely on photoconductive substances whose electrical resistance decreases when light falls on them. A layer of such a substance with a grounded backing plate is given a uniform electrostatic charge in the dark. When a light image is projected onto the surface, the photoconductor allows the electrostatic charge to leak away in proportion to the exposure. This leaves an "image" charge that can be converted, in various ways, into a visible image.

Principles of xerography

In xerography the photoconductive layer is selenium, and the image is made visible by dusting the plate with an electrostatically charged powder (toner) having a charge that is the opposite of that of the electrostatic image. The powder adheres to the image portions only and is then transferred to a sheet of plain paper also under the in-

fluence of electrostatic fields. A final heat treatment fuses the powder into the paper for a permanent picture. The process usually makes a positive from a positive original. In office copying machines (the main application of xerography) the whole operating sequence is programmed and automated. A zinc oxide-coated paper may replace the selenium plate; if so, the pigment powder deposit is fused directly into the paper surface.

The process is used mainly for line images without intermediate tones between black and white. Modified procedures permit continuous-tone reproduction and—with coloured pigments—also colour printing.

In the electroplastic process a transparent thermoplastic serves as the photoconductive layer. After the plastic is charged and exposed, the residual electrostatic charge forms stresses in the thermoplastic. Controlled heating deforms the surface in the image areas into a grain pattern, which is frozen into the plastic on cooling. The resulting image is light-scattering and is viewed by reflection or in special projection systems.

Colloid and photopolymer processes. A comparatively early non-silver process depended on organic colloid (gum or gelatin) treated with a bichromate. Exposure to light hardened the gelatin, rendering it insoluble, while unexposed portions could be washed away with warm water, leaving a relief image.

Photopolymer systems substitute a plastic precursor in place of the gelatin. The plastic precursor polymerizes to an insoluble plastic when exposed to light, and the unexposed soluble material is washed out by a suitable solvent. Photopolymer processes have been adapted for forming resists (protective coatings) for etching, as, for instance, in the manufacture of printed circuits. In indirect photopolymer systems a light-sensitive substance is mixed with a plastic precursor and on exposure decomposes into compounds that initiate polymerization of the plastic. The polymerizable layer may include a pigment for a final coloured image. Superimposing colour images derived from separation negatives can yield positives; systems of this type are used for quick colour proofing in photomechanical reproduction.

Diazonium processes. A diazo, or dyeline, process depends on the decomposition by light of organic diazonium salts. These salts can also couple with certain other compounds to form dyes. After exposure only the exposed (and decomposed) diazonium salt forms dye, producing a positive image from a positive original.

The materials are usually papers or transparent supports impregnated with the required chemicals. They are mainly sensitive to ultraviolet rays and can therefore be handled by normal tungsten lighting.

The light-decomposition of diazonium compounds also produces gaseous nitrogen. This phenomenon is utilized in vesicular processes that incorporate the diazonium compound in a thermoplastic layer. The nitrogen slowly diffuses out of this layer, but, if heat is applied immediately after exposure, the expanding nitrogen gas forms minute light-scattering bubbles visible as an image. The scattering power corresponds to the exposure. Further general exposure, after the plastic has cooled, decomposes the residual diazonium compound with gradual diffusion of the nitrogen out of the layer, destroying the latter's light sensitivity. This process and thermal dyeline systems are dry-processing instant-access systems and are used for making microfilm duplicates.

Photochromic systems. Certain dyelike substances can exist in a colourless and a coloured state. They are called photochromic compounds. The coloured state is formed by exposure to radiations of a certain wavelength. The compound reverts to its colourless state either in the dark or on treatment with radiation of a different wavelength. This reversibility is a primary characteristic of photochromism, and it is an instant-image system involving no processing.

Reversibility of the photochromic systems

Photochromic systems are used in microrecording (see below *Microfilming and microreproduction*). As the change of state takes place on a molecular level, the images are practically grain-free, and resolution is limited only by the resolving power of the optical system being used. Photochromic materials can be negative- or positive-working. With some photochromic compounds the dye image can be rendered permanent by optical or other treatment.

Glasses containing certain metal compounds also act as photochromic materials. Exposure to light breaks down the compounds into metal that forms a visible (and permanent) image in the glass. Another type of photochromic glass contains silver halide crystals dispersed in the glass melt. The action of light decomposes the silver halide, forming a visible silver deposit. The halogen cannot escape from the glass, so it recombines with the silver in the dark and the image fades. Such photochromic glasses are incorporated in automatic light-control devices; light transmission decreases as the intensity of the light reaching the glass rises. Such glass has found use in certain types of sunglasses.

Electronic photography. As television cameras and recorders became more compact, home video recording began to replace home movies in the amateur field in the late 1970s. Video recording of still images was incidental to this; it became widely involved in the storage of computer-generated or computer-processed images on magnetic tape or discs, for instance, in satellite photography, radiography, image scanning in picture transmission, and photomechanical reproduction.

A still video camera resembling traditional photographic apparatus (the Sony Mavica single-lens reflex) was first demonstrated in 1981. It uses a fast-rotating magnetic disc, two inches in diameter, recording on it up to 50 separate video images formed in a solid-state device in the camera. The images can be played back through a television receiver or monitor, or converted to paper in a printer that uses the video signals to control a printout device. Apart from being a potential rival to instant-picture photography, electronic records of this type are capable of direct transmission via telephone lines. Thus the process is of interest to press photographers, who can transmit pictures from their cameras directly to newspaper editorial offices without intermediate processing. The magnetic record also is able to directly control halftone engraving machines to engrave printing plates or cylinders.

SPECIAL TECHNIQUES AND APPLIED PHOTOGRAPHY

High-speed and stroboscopic photography. High-speed photography is generally concerned with exposure times shorter than about $^1/_{1,000}$ second (one millisecond) and often exposures shorter than $^1/_{1,000,000}$ second (one microsecond). This field partly overlaps that of high-speed cinematography—sequences of very short exposures. Exposure times can be reduced by high-speed shutter systems or by short-duration flash sources.

High-speed photography, together with high-speed cinematography, aids in the study of missiles, explosions, nuclear reactions, and other phenomena of military and scientific interest. In industry high-speed pictures show up movement phases of machinery, relays, and switches; dynamic fractures of materials or insulation breakdown; and, in natural science studies, flight movement of birds and insects.

High-speed shutters. The shortest exposure with mechanical shutters is about $^1/_{4,000}$ second. Special high-speed shutter systems are magneto-optical, electro-optical, or electronic. A magneto-optical shutter (Faraday shutter) consists of a glass cylinder placed inside a magnetic coil between two crossed polarizing filters; so long as the filters remain crossed, virtually no light can pass through. A brief current pulse through the coil generates a magnetic field that rotates the light's plane of polarization in the cylinder so that during the pulse some light passes through the second polarizing filter. The electro-optical shutter (Kerr cell) is made up of a liquid cell of nitrobenzene fitted with electrodes and again placed between two crossed polarizers. An electric pulse applied to the electrodes changes the polarization properties of the nitrobenzene so that this arrangement again transmits light. Minimum exposure time is around five nanoseconds (5×10^{-9} second). Image converter tubes electronically transmit and amplify an optical image focused on one end of a tube onto a phosphorescent screen at the other end. Electrons flow in the tube only in

Principle of the Kerr cell

the presence of an electric field, which can be controlled by short-time pulses down to a few nanoseconds.

High-speed light sources. The shortest electronic-flash duration is around one microsecond. Spark discharges in air between electrodes yield still shorter exposures; discharge voltage may go up to tens or hundreds of thousands of volts. Short-duration pulses applied to X-ray tubes produce X-ray flashes for high-speed radiography. The shortest exposures are between 20 and 50 nanoseconds. Special switching modes turn lasers into high-speed sources with durations down to a fraction of a nanosecond.

Synchronization. Generally the event photographed is made to trigger the exposure (the current pulse to operate the shutter or flash or spark source) to ensure correct synchronization. Examples are bullets interrupting a light beam to a photocell or self-luminous phenomena (explosions) triggering the system via a photocell circuit. The event and the exposure may be also triggered together by a signal from a common source.

Stroboscopic photography. Electronic-flash units designed to flash in rapid succession (up to several hundred times a second) can photograph a moving subject in front of a stationary camera with its shutter open to yield multiple images of successive movement phases. The technique has been used in pictorial and sports photography (*e.g.,* recording the movement of dancers or golfers) and for analyzing movement cycles without a motion-picture camera. Stroboscopic flash can be synchronized with a selected movement phase of an object in rapid cyclic motion (*e.g.,* a rotating machine component); the moving component illuminated in this way then appears stationary.

Aerial photography. Photographs from airborne or spaceborne vehicles either provide information on ground features for military and other purposes (reconnaissance) or record the dimensional disposition of such features (surveying).

Reconnaissance photography

Reconnaissance photographs call for maximum sharpness and detail rendering. Infrared films are often used to bring out details not discernible visually. In nonmilitary applications such photographs may reveal ecological factors (tree diseases, crop variations) and traces of archaeological sites not visible from the ground. Such shots are generally taken with cameras using 5- or 9½-inch roll film in large magazines, built into the aircraft and operated electrically by the pilot or other crew member, or automatically at set intervals. Some systems incorporate a shutterless technique; the film runs continuously past a slit at a rate matched exactly to the image movement in the camera's focal plane as the aircraft flies over the ground (image motion compensation).

Aerial survey is a systematic procedure of photographing the ground for map production; exposures are made at intervals to partly overlap the view of successive pictures. The individual photographs are enlarged to the same degree and then assembled in a precise mosaic. Aerial photographs taken under precisely specified conditions can serve for accurate measurements of ground details by stereoscopic evaluation (see below *Stereoscopic and three-dimensional photography*).

Satellite and space photography. Satellites orbiting the Earth record changing meteorologic features (weather satellites) and broadcast the video images to ground stations where they may be recorded on magnetic tape or converted to hard-copy pictures by suitable printers. Video cameras in spacecraft sent to record surface details of other planets similarly scan electronically the view taken in by a lens and beam the scanning signals back to Earth, where they are recorded and reconverted to visible images. The signals are usually processed electronically to enhance image information and detail. Such enhancement often brings out more information than can be recorded by conventional photography. Similar techniques are used by military satellites monitoring ground features from high orbits above the Earth.

Underwater photography. Underwater photography requires either special watertight cameras or pressure-resistant housings for normal cameras. In both cases camera functions are controlled through pressure-tight glands. A flat glass or plastic window is usually in front of the camera lens. The red and yellow absorption of the water more than a few feet below the surface turns colour photographs taken by daylight into virtually monochrome shots; hence artificial light is essential to show up the full colour range of fish and other underwater subjects. Light sources are battery-powered tungsten or tungsten-halogen lamps or electronic flash units (again in self-contained pressure-proof housings). For comfortable handling the weight of the housing with camera is adjusted to slight negative buoyancy. Complete camera and lighting outfits may be built into sledgelike or torpedo-like units with an electric or compressed-air motor for self-propulsion through the water.

Since the refractive index ratio of glass to water is lower than for glass to air, the light-bending power of a glass lens is less in water than in air. This factor reduces the lens's angle of view and makes objects appear at about three-fourths of their actual distance. This difference must be allowed for in focusing—possibly by a suitably calibrated distance scale or by fitting the housing with a compensating porthole, which acts as a diverging lens.

Underwater cameras

Underwater cameras with lenses designed for direct contact with the water eliminate the air space between the lens and the porthole. Such lenses can cover wider angles of view without distortion, but they do not give sharp images outside the water.

Close-range and large-scale photography. Near photography to reveal fine texture and detail covers several ranges: (1) close-up photography at image scales between 0.1 and 1 (one-tenth to full natural size); (2) macrophotography between natural size and 10 to 20× magnification, using the camera lens on its own; (3) photomicrography at magnifications above about 20×, combining the camera with a microscope; and (4) electron micrography with an electron microscope at magnifications of 10,000 to 1,000,000×, which involves photography of the electron microscope's phosphor screen or placing a photographic emulsion inside the vacuum chamber of the electron microscope to record directly the image formed by the electron beams.

Close-up and macrophotography. Supplementary close-up lenses or extension tubes (placed between the lens and camera body) allow the camera to focus on near distances for large scales of reproduction. Special close-up rangefinders or distance gauges establish exactly the correct camera-to-subject distance and precise framing of the subject field. Special simple close-up cameras, as in fingerprint recording and certain fields of medical photography, are permanently set to a fixed near distance and have a distance gauge or similar device built in. Screen-focusing cameras (view and single-lens reflex) need no such aids, as the finder screen shows the precise focus and framing.

Extension tubes or extension bellows or both or "macro" lenses of extended focusing range are used for the macro range of distances. For optimum image quality macrophotographic lenses specially corrected for large image scales may be used or the camera lens reversed back to front.

Photomicrography. There are two principal methods of photographing through a microscope. In the first the camera, with its lens focused at infinity, is lined up in the optical axis of the microscope, which is also focused visually on infinity. In the other method the camera without lens is positioned behind the microscope eyepiece, which is focused to project the microscope image directly onto the film.

Special photomicrographic cameras generally employ the second method. Microscope adapters to provide a light-tight and rigid connection between the camera and microscope are available for both systems. Such microadapters may incorporate their own shutter and a beam splitter system for viewing and focusing of the microscope image through a focusing telescope. Photomicrographs are the essential adjunct to all microscopy to record biologic, bacteriologic, physical, and other observations in black-and-white or colour.

Stereoscopic and three-dimensional photography. Visual three-dimensional depth is perceived partly because of the fact that the human eyes see a scene from two viewpoints separated laterally by about 2½ inches. The two views show slightly different spatial relationships between

Basis of depth perception

near and distant objects (parallax); the visual process fuses these stereoscopic views into a three-dimensional impression. A similar impression is obtained by viewing a pair of stereoscopic photographs taken with two cameras or a twin camera with lenses 2½ inches apart, so that the left eye sees only the picture taken by the left-hand lens and the right eye only that of the right-hand lens. Binocular viewers or stereo-selective projection systems permit such viewing.

Stereo photographs can also be combined in a single picture by splitting up the images into narrow vertical strips and interlacing them. On superimposing a carefully aligned lenticular grid on the composite picture, an observer directly sees all the strips belonging to the left-eye picture with the left eye and all the strips belonging to the right-eye picture with the right eye. Such parallax stereograms are seen in display advertising in shop windows. They also can be reproduced in print, overlaid by a lenticular pattern embossed in a plastic covering layer.

Photogrammetry makes use of stereo photography in measuring dimensions and shapes of ground objects in depth, as from successive exposure pairs made during an aerial survey flight. If all exposure parameters, including flying height, ground separation between exposures, and focal length of the aerial camera lens are known, the height of each ground feature can be measured. Photogrammetric plotting instruments do this and draw height contour curves of all features for aerial maps. Similar photogrammetric evaluation of stereo photographs of nearby subjects can also be made. For instance, it is possible to reconstruct accurately the scene of a highway accident. In industry a photogrammetric plot of an automobile model can be fed into a computer to program the machine tools that will shape the full-scale motor body components.

Infrared photography. Images formed by infrared and heat radiations can be recorded directly, on films sensitive to them, or indirectly, by photographing the image produced by some other system registering infrared radiation.

Applications of direct infrared photography

Silver halide emulsions can be sensitized to infrared rays with wavelengths up to around 1,200 nanometres (one nanometre is $^{1}/_{1,000,000}$ of a millimetre). The usual sensitivity range is 800 to 1,000 nanometres. Direct infrared-recording aerial photography shows up ground features of differential infrared reflection but similar light reflection (*e.g.,* different types of foliage) and cuts through haze and mist. Special colour films with an infrared-sensitive layer and processed to colours different from the natural rendering (false-colour films) show up such differences still more clearly. In forensic photography infrared pictures reveal ink alterations in forgeries, differentiate stains, and help to identify specific textiles and other materials. In medicine infrared photographs show subcutaneous blood vessels, as the skin is transparent to infrared.

With suitable equipment it is possible to convert an infrared image into one visible on a fluorescent screen, where it can be photographed. In infrared scanner systems a moving mirror scans the object or scene and focuses the radiation onto an infrared-sensitive cell. The cell generates electric signals to modulate a light source, which, in turn, scans a photographic film or paper synchronously with the mirror. The resulting image records hotter and colder parts of the object as lighter and darker areas and can accurately establish actual temperatures of subject details. This system has been used to record temperature variations in the skin for the diagnosis of cancer.

Ultraviolet photography. Invisible shortwave ultraviolet radiations can be recorded directly or used in fluorescence photography. For direct ultraviolet recording, the photographically useful wavelength range lies between 400 nanometres (visible violet) and about 200 nanometres and needs special optical systems transparent to ultraviolet rays (quartz, silica, or fluoride elements or combinations thereof). Light sources rich in ultraviolet such as mercury vapour lamps—with an ultraviolet-transmitting, but visually opaque, filter in front of the camera lens—ensure that the photograph records only the ultraviolet-reflecting characteristics of the subject.

Fluorescence photography records the glow or visible light given off by certain substances when they are irradiated by ultraviolet rays. The object is illuminated by screening out the visible light with a filter that transmits only ultraviolet radiation, and another filter that absorbs the ultraviolet rays is placed over the camera lens, permitting only the visible light (fluorescence) to be recorded on the film. Normal lenses and panchromatic or colour materials are used.

Applications of ultraviolet photography

Ultraviolet photography can identify or separate pigments and fabrics and can detect forgeries of documents. Fluorescence photography can identify dyes, stains, specific chemical substances, and fluorescent components in microscope specimens. Ultraviolet microscopy offers increased resolution through the shorter-wavelength radiations employed. Aerial and satellite photography by ultraviolet can show up ultraviolet-reflective ground features.

Radiography and other radiation recording techniques. Silver halide emulsions are sensitive to X rays, gamma rays, and charged particles emitted by radioactive substances. Some of these rays penetrate visually opaque materials to varying degrees to show up internal structures. Radiography covers techniques of recording the subsurface features of objects.

X-ray radiography. X rays (wavelengths between $^{1}/_{100}$ and $^{1}/_{100,000}$ that of visible light) are produced by high-voltage electron streams bombarding an electrode in a vacuum tube. For radiography the object to be recorded is placed between an X-ray tube and the film; the film registers the differential absorption of the X rays by the object's internal structure as a projection shadowgraph.

The most familiar application is in medicine for diagnosis and recording, including dental radiography. Industrial radiography permits nondestructive inspection of castings, welds, and engineering structures.

Gamma radiography. The technique of gamma-ray radiography is similar to that of X-ray radiography except that it relies on rays emitted by radioactive substances. Gamma rays have wavelengths from 100 to 1,000 times shorter than X rays and correspondingly greater penetrating power. Small gamma-ray sources are placed in areas inaccessible to X-ray tubes, such as inside pipelines. In all radiographic applications the exposure occurs under conditions of normal light, from which the radiographic film is protected by a light-tight (but radiation-transparent) wrapping.

Autoradiography. Autoradiography records the distribution of radioactive materials in botanical and histological specimens placed in contact with a photographic emulsion. This technique has been applied to the study of metabolism of plants and animals; it records the activity of organic compounds of radioactive isotopes introduced into the system of the plant or animal. In engineering studies autoradiography can be used to follow the transfer of radioactive substances from one surface to another in lubrication. The technique also has applications in machining and other metal-treatment processes.

Nuclear-track recording. Tracks of subatomic particles, such as protons, electrons, and mesons, produced by nuclear reactions can be recorded by photographic means. The most common technique is to photograph the visible traces of such tracks in bubble or spark chambers with special camera and lens arrangements. Different arrangements can provide for coverage of large fields or the recording of tracks simultaneously from several directions for three-dimensional reconstruction.

Particle tracks can be recorded directly in thick (up to one millimetre) emulsion layers or in emulsion stacks (up to 20 inches) carried in high-altitude balloons and in spacecraft and satellites. Special processing procedures are required to deal with these emulsion thicknesses.

Astronomical photography. By the cumulative effect of light received over a long period, a photographic emulsion can record celestial objects too faint to be visible. Before radio telescopes (see TELESCOPES: *Radio telescopes*), photography was the only way of detecting many such objects.

Astronomical cameras are film- or plate-holding units built onto high-power telescopes, typically reflecting systems. The telescopes run on precision, clock-driven mounts to keep the optical axis stationary with respect to the sky area as the Earth rotates during an exposure time,

which can run into several hours. For increased recording sensitivity, the telescope image may be intensified electronically.

Astronomical photographs taken through narrow-band colour filters—including infrared or ultraviolet transmitting filters—show selective emission characteristics of stars. In the case of the Sun and of planets, such photographs can reveal some surface details not observable by white light. Colour photographs reveal colours not directly visible because the intensity of starlight is too low to stimulate the eye's colour-vision mechanism.

Spectrography records the composition of light emitted by stars and other objects, the star image of the telescope being photographed through a diffraction grating, a device that disperses white light into constituent wavelengths. Elements present in the star or the gas mantle surrounding it can be identified from their characteristic spectral lines. Displacement of such lines from their known wavelength position can indicate the velocity with which the distant stellar systems recede from or approach the Earth.

Microfilming and microreproduction. Microfilming is the copying of documents, drawings, and other such matter at a reduced scale—typically 1:15 to 1:42—for compact storage. Complete microreproduction systems include methods of filing the film copies for easy retrieval and reenlargement. Various duplication methods allow microfilm records to be extensively distributed.

Documents, periodicals, and other printed matter are usually microfilmed on 16-mm film with an image size between 10×14 and 14×20 mm in a copying camera taking 100-foot lengths of film. Engineering drawings of high information content are microfilmed on 35-mm unperforated film with a standard image size of 32×45 mm. Films of up to 105 mm in width are also used. Automated microfilm cameras run continuously, documents being fed onto a moving band carried past the camera at a steady speed while the film runs past a slit at a matched rate.

Readers and reader printers are desk-top projectors that display the frames reenlarged to about natural size on a back projection screen. In a reader printer the image may also be projected on sensitized paper for full-size enlargements. Advanced readers have elaborate retrieval systems based on frame coding and run the microfilm rolls through at high speed until a specific searched image is reached.

Aperture cards or standard-size transparent jackets store microfilm images as single frames or groups of frames. Such unitized microfilms permit easier indexing and retrieval. Certain 35-mm microfilm cameras photograph the original document directly on film premounted in an aperture card and processed on the spot.

Widely used is the unitized microfiche system, which carries up to 98 frames, each about 9×12 mm, on a 4×6-inch sheet of film. The microfiche camera repositions the film frame by frame after every exposure. Microfiche with a larger frame can also be produced by jacketing strips of 16-mm microfilm in multichannel plastic jackets 4×6 inches in size.

For greater space saving, microfilm images may be reduced beyond 1:100 on high-resolution photochromic image materials. Extreme fine-grain silver copies then hold 3,000 to 4,000 individual frames on a single 4×6-inch film. This method, useful for complex catalogs and like purposes, offers easy retrieval of individual frames but requires a high-magnification reader.

THE PHOTOGRAPHY INDUSTRY

Equipment manufacture

Present-day manufacture of cameras and other photographic equipment is concentrated in mass-production plants that make most of the components (camera bodies, lenses, shutters, and other parts) on largely automated machines; the components are then assembled by semiskilled or skilled labour. Smaller manufacturers of low- and medium-priced cameras obtain components for assembly from such specialist suppliers as shutter manufacturers and lens producers. High-quality precision cameras are produced on a smaller scale with automated fabrication of the engineering components but much more extensive manual assembly by highly skilled technicians. Components and functions of every camera are tested at every production stage; less expensive cameras are usually batch-tested by a sampling procedure.

The raw material for lens manufacture covers a range of optical glasses of different optical characteristics. About 10 major worldwide glass producers supply the several dozen optical firms offering lenses of well-known brands. The glass is cast into blanks for specific lens elements and ground and polished to the required exact specifications, with the elements assembled in metal (sometimes plastic) mounts. Extensive production tests and optical performance checks safeguard quality standards.

Film base is produced either by coating a solution of the base material on large drums, where it solidifies (film casting), or by extrusion of plastics, such as polyester, in film extruders. For print materials, paper of suitable purity is coated with a barium sulfate emulsion in gelatin, to provide a smooth white surface, and then with the silver halide emulsion. Silver halide emulsions are made by mixing silver nitrate with a solution of alkali halide—typically potassium bromide and iodide—in gelatin. The silver halide then precipitates out as fine crystals. After cooling to a jelly, shredding, and washing, the emulsion is remelted and treated to increase speed and contrast. Colour sensitizers (and colour couplers for colour emulsions) and additives are introduced, and the gelatin emulsion is machine-coated on wide continuous webs of paper or film. Generally several coatings are applied—up to a dozen for certain colour films. Operations from emulsion mixing onward are carried on in total darkness. After cooling and drying, the material is batch-tested for consistent characteristics and then is cut and packed.

Photofinishing laboratories process most amateur and some professional photographers' films and prints. In the 1980s, virtually all of the total business of the laboratories in the United States was in colour processing.

Continuous processing of films

Photofinishing laboratories use machines that carry the films in spliced-together lengths or on racks through successive tanks of the processing solutions. Prints are usually made to standard formats on automatic enlargers, taking both the negatives and the paper in continuous rolls. The paper rolls of 250 or 500 feet are processed in continuous-strip processors, which deliver prints dry and ready for automatic cutting. Many printers have automatic exposure measurement based on overall negative density, with automatically controlled colour correction for colour negatives. High-capacity colour printers of this type can produce 2,000 to 3,000 prints per hour. Coding systems identify individual films and corresponding prints by customer or order number for final re-sorting. More exacting processing services grade colour negatives before printing by light transmission measurements through different colour filters; the resulting exposure data may be punched as edge codes in the film itself or programmed on perforated paper tape. When the tape is run through the printer together with the film, the perforations directly control the colour exposures and corrections. Advanced automatic printing systems may involve electronically controlled image enhancement.

Enlargements to special sizes and colour printing for professional photographers require individual enlarging by skilled personnel on conventional enlargers with advanced automation features of focusing, exposure measurement, and colour control. Other processing services include duplication of transparencies, various types of photocopying (partly on coin-operated copiers in public places), microfilming, and microfilm processing. (L.A.Ma./Ed.)

BIBLIOGRAPHY. General reference works on photography include INTERNATIONAL CENTER OF PHOTOGRAPHY, *Encyclopedia of Photography* (1984); and the volumes in the "Life Library of Photography" series by the editors of Time-Life Books, on the art and history of photography, types of photography, and techniques and processes. See also TURNER BROWNE and ELAINE PARTNOW, *Macmillan Biographical Encyclopedia of Photographic Artists & Innovators* (1983); MICHELE AUER and MICHEL AUER, *Encyclopedie internationale des photographes de 1839 a nos jours: Photographers Encyclopaedia International 1839 to the Present* (1985); and COLIN NAYLOR (ed.), *Contemporary Photographers*, 2nd ed. (1988).

The history and art of photography: Historical overviews of the development of the art of photography are provided

in ROBERT TAFT, *Photography and the American Scene: A Social History, 1839–1889* (1938, reprinted 1964); HELMUT GERNSHEIM, *Creative Photography: Aesthetic Trends, 1839–1960* (1962); PETER POLLACK, *The Picture History of Photography: From the Earliest Beginnings to the Present Day*, rev. and enlarged ed. (1969), especially valuable for its wealth of illustrations; WILLIAM WELLING, *Photography in America: The Formative Years, 1839–1900* (1978, reprinted 1987); PETR TAUSK, *Photography in the 20th Century* (1980; originally published in German, 1977); BEAUMONT NEWHALL, *The History of Photography: From 1839 to the Present*, rev. and enlarged ed. (1982), the stylistic development of the art of photography as related to the technological and scientific characteristics of the medium; NAOMI ROSENBLUM, *A World History of Photography* (1984); PETER TURNER (ed.), *American Images: Photography 1945–1980* (1985); and JEAN-CLAUDE LEMAGNY and ANDRÉ ROUILLÉ (eds.), *A History of Photography: Social and Cultural Perspectives* (1987; originally published in French, 1986).

Essays on art and photography can be found in CHARLES H. CAFFIN, *Photography as a Fine Art: The Achievements and Possibilities of Photographic Art in America* (1901, reissued 1972); IRVING PENN, *Moments Preserved: Eight Essays in Photographs and Words* (1960); NATHAN LYONS (ed.), *Photographers on Photography: A Critical Anthology* (1966), essays by photographers from the 1890s to the 1960s; JOHN SZARKOWSKI, *The Photographer's Eye* (1966, reissued 1980), a penetrating examination of photographic aesthetics; AARON SCHARF, *Art and Photography* (1968, reprinted 1986); VOLKER KAHMEN, *Photography as Art* (1974; originally published in German, 1973); JEAN-LUC DAVAL, *Photography, History of an Art* (1982; originally published in French, 1982); BRYAN HOLME (ed.), *Photography as Fine Art* (1983); and ANDY GRUNDBERG and KATHERINE MCCARTHY GAUSS, *Photography and Art: Interactions Since 1946* (1987), photographs from an exhibition.

Studies of particular schools and types of photography include ROBERT DOTY, *Photo Secession: Photography as a Fine Art* (1960, reprinted with title *Photo-Secession: Stieglitz and the Fine-Art Movement in Photography*, 1978), an account of events leading up to and following the founding of the society by Alfred Stieglitz; WILLIAM CULP DARRAH, *Stereo Views: A History of Stereographs in America and Their Collection* (1964); *French Primitive Photography* (1969), a catalog of an exhibition; RICHARD RUDISILL, *Mirror Image: The Influence of the Daguerreotype on American Society* (1971), a thorough survey; VAN DEREN COKE, *The Painter and the Photograph: From Delacroix to Warhol*, rev. and enlarged ed. (1972), an exhibition catalog; BEAUMONT NEWHALL, *The Daguerreotype in America*, 3rd rev. ed. (1976), a study of the industry as well as the art of daguerreotyping; BEN MADDOW, *Faces: A Narrative History of the Portrait in Photography* (1977); MARGARET HARKER, *The Linked Ring: The Secession Movement in Photography in Britain, 1892–1910* (1979); VAN DEREN COKE, *Avant-Garde Photography in Germany, 1919–1939*, trans. from German (1982); RAINER FABIAN and HANS-CHRISTIAN ADAM, *Masters of Early Travel Photography* (1983; originally published in German, 1981); and *Bauhaus Photography* (1985; originally published in German, 1982). See also BEAUMONT NEWHALL and NANCY NEWHALL (eds.), *Masters of Photography* (1958, reissued 1982); short biographical sketches of 19 photographers, with representative photographs by each.

Photojournalism as a separate form is discussed in WILSON HICKS, *Words and Pictures: An Introduction to Photojournalism* (1952, reissued 1973); KEN BAYNES (ed.), *Scoop, Scandal, and Strife: A Study of Photography in Newspapers* (1971); and JOHN R. WHITING, *Photography Is a Language* (1946, reprinted 1979). See also STANLEY RAYFIELD, *How Life Gets the Story: Behind the Scenes in Photojournalism* (1955), field experience of photographers from *Life*, presented in the format and photographic essay style of the magazine. (Be.N./H.E.R.G.)

The technology of photography: General works include *The Focal Encyclopedia of Photography*, rev. ed., 2 vol. (1965, reissued 1977); L.P. CLERC, *Photography: Theory and Practice*, rev. and enlarged ed. edited by D.A. SPENCER, 4 vol. (1970–71; originally published in French, 1926), a classic treatise; *The Theory of the Photographic Process*, 4th ed. edited by T.H. JAMES (1977), a classic work; C.B. NEBLETTE, *Neblette's Handbook of Photography and Reprography: Materials, Processes, and Systems*, 7th ed. edited by JOHN M. STURGE (1977); *Encyclopedia of Practical Photography*, 14 vol. (1977–79), edited by and published for the Eastman Kodak Company; JOHN HEDGECOE, *The Photographer's Handbook: A Complete Reference Manual of Techniques, Procedures, Equipment, and Styles*, 2nd ed. rev. (1982); and BRUCE PINKARD, *The Photographer's Bible: An Encyclopedic Reference Manual* (1983). See also ALBERT BONI (ed.), *Photographic Literature* (1962), and a supplemental volume, *Photographic Literature, 1960–1970* (1972), an exhaustive and valuable bibliography; WOLFGANG BAIER, *Quellendarstellungen zur Geschichte der Fotografie: A Source Book of Photographic History* (1963, reissued 1977), detailed bibliographies and references to the literature on photographic developments, with an introduction in English; and BEAUMONT NEWHALL, *Latent Image: The Discovery of Photography* (1967, reissued 1983), helpful for the understanding of the establishment of the medium.

The evolution of photographic techniques is traced in H. FOX TALBOT (WILLIAM HENRY FOX TALBOT), *The Pencil of Nature* (1844–46, reprinted 1969), the inventor's account, illustrated with 24 actual calotypes; GEORGES POTONNIÉE, *The History of the Discovery of Photography* (1936, reissued 1973; originally published in French, 1925), a detailed account of the early days of photography; JOSEF MARIA EDER, *History of Photography* (1945, reprinted 1978; originally published in German, 4th rev. ed., 2 vol., 1932), a pioneer Austrian work that deals primarily with the scientific and technological development of photography; BEAUMONT NEWHALL (ed.), *On Photography: A Sourcebook of Photo History in Facsimile* (1956), an anthology of the inventors' own accounts of various processes; D.B. THOMAS, *The First Negatives: An Account of the Discovery and Early Use of the Negative-Positive Photographic Process* (1964); JOSEPH S. FRIEDMAN, *The History of Color Photography*, 2nd ed. (1968); HELMUT GERNSHEIM, *The History of Photography from the Camera Obscura to the Beginning of the Modern Era*, 2nd ed. (1969), the first part of which was revised as *The Origins of Photography* (1982); and GAIL BUCKLAND, *Fox Talbot and the Invention of Photography* (1980).

Camera history and technology is outlined in LESLIE D. STROEBEL, *View Camera Technique*, 5th ed. (1986), on the use of studio and field cameras in industrial, commercial, and other applications; MICHEL AUER, *The Illustrated History of the Camera from 1839 to the Present*, trans. from French and adapted by D.B. TUBBS (1975); BRIAN COE, *Cameras: From Daguerreotypes to Instant Pictures* (1978); and EATON S. LOTHROP, JR., *A Century of Cameras from the Collection of the International Museum of Photography at George Eastman House*, rev. and expanded ed. (1982).

Lenses and optical principles are described in C.B. NEBLETTE and ALLEN E. MURRAY, *Photographic Lenses*, rev. ed. (1973); ARTHUR COX, *Photographic Optics*, 15th rev. ed. (1974), classic manual of lens principles and use; and SIDNEY F. RAY, *The Photographic Lens* (1979), an introduction.

Film and the techniques of taking pictures are examined in WALTER NURNBERG, *Lighting for Photography: Means and Methods*, 16th rev. ed. (1968, reissued 1971); and MICHAEL LANGFORD, *Basic Photography*, 5th ed. (1986), and *Advanced Photography: A Grammar of Techniques*, 4th ed. (1980), manuals of practical technique for professional photographers.

Film processing and printing are the subject of D.H.O. JOHN and G.T.J. FIELD, *A Textbook of Photographic Chemistry* (1963), basics of chemical reactions in black-and-white processing; C.I. JACOBSON and R.E. JACOBSON, *Developing: The Negative-Technique*, 18th ed. (1972), manual of all aspects of negative technique; C.I. JACOBSON and L.A. MANNHEIM, *Enlarging*, 22nd ed. (1975), manual of positive technique in black and white and colour; L.F.A. MASON, *Photographic Processing Chemistry*, 2nd ed. (1975), detailed treatment of processing mechanisms and reactions; GRANT HAIST, *Modern Photographic Processing*, 2 vol. (1979), chemistry and technology of black-and-white and colour processing; and JAN ARNOW, *Handbook of Alternative Photographic Processes* (1982).

Colour photography is treated in LOUIS WALTON SIPLEY, *A Half Century of Color* (1951); RALPH M. EVANS, W.T. HANSON, JR., and W. LYLE BREWER, *Principles of Color Photography* (1953), principles of colour rendering, response, and reproduction; D.A. SPENCER, *Colour Photography in Practice*, rev. ed. by L.A. MANNHEIM and VISCOUNT HANWORTH (1966, reissued 1975), containing both theory and practical techniques; R.W.G. HUNT, *The Reproduction of Colour*, 3rd ed. (1975), a standard handbook on colour photography, television, and printing, with moderately advanced mathematical treatment; and GERT KOSHOFER, *Farbfotographie*, 3 vol. (1981), a complete historical review, including a lexicon of equipment and materials.

Special photographic techniques and applications are the focus of HAROLD E. EDGERTON and JAMES R. KILLIAN, JR., *Flash! Seeing the Unseen by Ultra High-Speed Photography*, 2nd ed. (1954), and *Moments of Vision: The Stroboscopic Revolution in Photography* (1979, reprinted 1984); J. BERGNER, E. GELBKE, and W. MEHLISS, *Practical Photomicrography* (1966; originally published in German, 1961), a comprehensive manual; R.F. SAXE, *High-Speed Photography* (1966), a condensed but comprehensive survey; JOHN BRACKETT HERSEY (ed.), *Deep-Sea Photography* (1967); C.R. ARNOLD, P.J. ROLLS, and J.C.J. STEWART, *Applied Photography* (1971), on scientific applications; H. LOU GIBSON, *Photography by Infrared*, 3rd ed. (1978); and GJON MILI, *Gjon Mili: Photographs and Recollections* (1980), on stroboscopic photography. (H.E.R.G./L.A.Ma.)

Photosynthesis

Photosynthesis, which means putting together with light, is the process by which green plants and certain other organisms transform light energy into chemical energy. During photosynthesis in green plants, light energy is captured and used to convert water, carbon dioxide, and minerals into oxygen and energy-rich organic compounds.

It would be impossible to overestimate the importance of photosynthesis in the maintenance of life on Earth. If photosynthesis ceased, there would soon be little food or other organic matter on Earth. Most organisms would disappear, and in time the Earth's atmosphere would become nearly devoid of gaseous oxygen. The only organisms able to exist under such conditions would be the chemosynthetic bacteria, which can utilize the chemical energy of certain inorganic compounds and thus are not dependent on the conversion of light energy.

Photosynthesis also is responsible for the "fossil fuels" (*i.e.*, coal, oil, and gas) that power industrial society. In past ages, green plants and small organisms that fed on plants increased faster than they were consumed, and their remains were deposited in the Earth's crust by sedimentation and other geological processes. There, protected from oxidation, these organic remains were slowly converted to fossil fuels. These fuels not only provide much of the energy used in factories, homes, and transportation, but they also serve as the raw material for plastics and other synthetic products. Unfortunately, modern civilization is using up in a few centuries the excess of photosynthetic production accumulated over millions of years.

Requirements for food, materials, and energy in a world where human population is rapidly growing have created a need to increase both the amount of photosynthesis and the efficiency of converting photosynthetic output into products useful to people. One response to these needs—the so-called "Green Revolution"—has achieved enormous improvements in agricultural yield through the use of chemical fertilizers, pest and plant disease control, plant breeding, and mechanized tilling, harvesting, and crop processing. This effort has limited severe famines to a few areas of the world despite rapid population growth, but it has not eliminated widespread malnutrition.

A second agricultural revolution, based on plant genetic engineering, may lead to increases in plant productivity and thereby partially alleviate malnutrition. Since the 1970s, molecular biologists have possessed the means to manipulate a plant's genetic material (DNA) to achieve improvements in disease and drought resistance, product yield and quality, frost hardiness, and other desirable properties. In the future, such genetic engineering may result in improvements in the process of photosynthesis.

This article is divided into the following sections:

General characteristics 799
- Development of the idea
- Overall reaction of photosynthesis
- Basic products of photosynthesis
- Evolution of the process

Factors that influence the rate of photosynthesis 800
- Light intensity and temperature
- Carbon dioxide
- Water
- Minerals
- Internal factors

Energy efficiency of photosynthesis 801

Chloroplasts, the photosynthetic units of green plants 802
- Structural features
- Chemical composition of lamellae

The process of photosynthesis: the light reactions 803
- Light absorption and energy transfer
- The pathway of electrons
- Evidence of two light reactions
- Photosystems I and II
- Quantum requirements

The process of photosynthesis: the conversion of light energy to ATP 804

The process of photosynthesis: carbon fixation and reduction 805
- Elucidation of the carbon pathway
- The reductive pentose phosphate cycle
- Regulation of the cycle
- Products of carbon reduction
- Photorespiration
- Carbon fixation via C_4 acids

The molecular biology of photosynthesis 806

Bibliography 806

GENERAL CHARACTERISTICS

Development of the idea. The study of photosynthesis began in 1771, with observations made by the English chemist Joseph Priestley. Priestley had burned a candle in a closed container until the air within the container could no longer support combustion. He then placed a sprig of mint plant in the container and discovered that after several days the mint had produced some substance (later recognized as oxygen) that enabled the confined air to again support combustion. In 1779 the Dutch physician Jan Ingenhousz expanded upon Priestley's work, showing that the plant must be exposed to light if the combustible substance (*i.e.*, oxygen) was to be restored; he also demonstrated that this process required the presence of the green tissues of the plant.

In 1782 it was demonstrated that the combustion-supporting gas (oxygen) was formed at the expense of another gas, or "fixed air," which had been identified the year before as carbon dioxide. Gas-exchange experiments in 1804 showed that the gain in weight of a plant grown in a carefully weighed pot was the sum of carbon, which came entirely from absorbed carbon dioxide, and water taken up by plant roots. Almost half a century passed before the concept of chemical energy developed sufficiently to permit the discovery (in 1845) that light energy from the sun is stored as chemical energy in products formed during photosynthesis.

Overall reaction of photosynthesis. In chemical terms, photosynthesis is a light-energized oxidation–reduction process. (Oxidation refers to the removal of electrons from a molecule; reduction refers to the gain of electrons by a molecule.) In plant photosynthesis, the energy of light is used to drive the oxidation of water (H_2O), producing oxygen gas (O_2), hydrogen ions (H^+), and electrons. Most of the removed electrons and hydrogen ions ultimately are transferred to carbon dioxide (CO_2), which is reduced to organic products. Other electrons and hydrogen ions are used to reduce nitrate and sulfate to amino and sulfhydryl groups in amino acids, which are the building blocks of proteins. In most green cells, carbohydrates—especially starch and the sugar sucrose—are the major direct organic products of photosynthesis. The overall reaction in which carbohydrates—represented by the general formula (CH_2O)—are formed during plant photosynthesis can be indicated by the following equation:

$$CO_2 + 2H_2O \xrightarrow[\text{green plants}]{\text{light}} (CH_2O) + O_2 + H_2O.$$

This equation is merely a summary statement, for the process of photosynthesis actually involves numerous complex reactions. These reactions occur in two stages:

the "light" stage, consisting of photochemical (*i.e.,* light-dependent) reactions; and the "dark" stage, comprising chemical reactions controlled by enzymes (organic catalysts). During the first stage, the energy of light is absorbed and used to drive a series of electron transfers, resulting in the synthesis of the energy-rich compound adenosine triphosphate (ATP) and the electron donor reduced nicotine adenine dinucleotide phosphate (NADPH). During the dark stage, the ATP and NADPH formed in the light reactions are used to reduce carbon dioxide to organic carbon compounds. This assimilation of inorganic carbon into organic compounds is called carbon fixation.

During the 20th century, comparisons between photosynthetic processes in green plants and in certain photosynthetic sulfur bacteria provided important information about the photosynthetic mechanism. Sulfur bacteria use hydrogen sulfide (H_2S) as a source of hydrogen atoms and produce sulfur instead of oxygen during photosynthesis. The overall reaction is

$$CO_2 + 2H_2S \xrightarrow[\text{sulfur bacteria}]{\text{light}} (CH_2O) + S_2 + H_2O.$$

In the 1930s Dutch biologist Cornelis van Niel recognized that the utilization of carbon dioxide to form organic compounds was similar in the two types of photosynthetic organisms. Suggesting that differences existed in the light-dependent stage and in the nature of the compounds used as a source of hydrogen atoms, he proposed that hydrogen was transferred from hydrogen sulfide (in bacteria) or water (in green plants) to an unknown acceptor (called A), which was reduced to H_2A. During the dark reactions, which are similar in both bacteria and green plants, the reduced acceptor (H_2A) reacted with carbon dioxide (CO_2) to form carbohydrate (CH_2O) and to oxidize the unknown acceptor to A. This putative reaction can be represented as:

$$CO_2 + 2H_2A \xrightarrow{\text{light}} (CH_2O) + 2A + H_2O.$$

Van Niel's proposal was important because the popular (but incorrect) theory had been that oxygen was removed from carbon dioxide (rather than hydrogen from water) and that carbon then combined with water to form carbohydrate (rather than the hydrogen from water combining with CO_2 to form CH_2O).

By 1940 chemists were using heavy isotopes to follow the reactions of photosynthesis. Water marked with an isotope of oxygen (^{18}O) was used in early experiments. Plants that photosynthesized in the presence of water containing $H_2{}^{18}O$ produced oxygen gas containing ^{18}O; those that photosynthesized in the presence of normal water produced normal oxygen gas. These results provided strong support for van Niel's theory that the oxygen gas produced during photosynthesis is derived from water.

Origin of oxygen gas in photosynthesis

Basic products of photosynthesis. As has been stated, carbohydrates are the most important direct organic product of photosynthesis in the majority of green plants. The formation of a simple carbohydrate, glucose, is indicated by a chemical equation,

$$\underset{\text{carbon dioxide}}{6CO_2} + \underset{\text{water}}{12H_2O} \xrightarrow[\text{green plants}]{\text{light}} \underset{\text{glucose}}{C_6H_{12}O_6} + \underset{\text{oxygen}}{6O_2} + \underset{\text{water}}{6H_2O}.$$

Little free glucose is produced in plants; instead, glucose units are linked together to form starch or are joined with fructose, another sugar, to form sucrose (see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Carbohydrates*).

Not only carbohydrates, as was once thought, but also amino acids, proteins, lipids (or fats), pigments, and other organic components of green tissues are synthesized during photosynthesis. Minerals supply the elements (*e.g.,* nitrogen, N; phosphorus, P; sulfur, S) required to form these compounds. Chemical bonds are broken between oxygen (O) and carbon (C), hydrogen (H), nitrogen, and sulfur, and new bonds are formed in products that include gaseous oxygen (O_2) and organic compounds. More energy is required to break the bonds between oxygen and other elements (*e.g.,* in water, nitrate, and sulfate) than is released when new bonds form in the products. This difference in bond energy accounts for a large part of the light energy stored as chemical energy in the organic products formed during photosynthesis. Additional energy is stored in making complex molecules from simple ones.

Evolution of the process. Although life and the quality of the atmosphere today depend on photosynthesis, it is likely that green plants evolved long after the first living cells. When the Earth was young, electrical storms and solar radiation probably provided the energy for the synthesis of complex molecules from abundant simpler ones, such as water, ammonia, and methane. The first living cells probably evolved from these complex molecules (see LIFE). For example, the accidental joining together (condensation) of the amino acid glycine and the fatty acid acetate may have formed complex organic molecules known as porphyrins; these molecules, in turn, may have evolved further into coloured molecules called pigments; *e.g.,* chlorophylls of green plants, bacteriochlorophyll of photosynthetic bacteria, hemin (the red pigment of blood), and cytochromes, a group of pigment molecules essential in both photosynthesis and cellular respiration.

Primitive coloured cells then had to evolve mechanisms for using the light energy absorbed by their pigments. At first, the energy may have been used immediately to initiate reactions useful to the cell. As the process for utilization of light energy continued to evolve, however, a larger part of the absorbed light energy probably was stored as chemical energy, to be used to maintain life. Green plants, with their ability to use light energy to convert carbon dioxide and water to carbohydrates and oxygen, are the culmination of this evolutionary process.

The first oxygenic (oxygen-producing) cells probably were the cyanophytes, or "blue-green algae," which appeared about 2,000,000,000 to 3,000,000,000 years ago. These microscopic organisms are believed to have greatly increased the oxygen content of the atmosphere, making possible the development of aerobic (oxygen-using) organisms. Cyanophytes are prokaryotic cells; that is, they contain no distinct, membrane-enclosed subcellular particles (organelles), such as nuclei and chloroplasts. Green plants, by contrast, are composed of eukaryotic cells, in which the photosynthetic apparatus is contained within membrane-bound chloroplasts. There is a theory that the first photosynthetic eukaryotes were red algae that may have developed when nonphotosynthetic eukaryotic cells engulfed cyanophytes. Within the host cells, these cyanophytes are thought to have evolved into chloroplasts. Alternatively, the ancestors of chloroplasts in green plants may have been another oxygenic prokaryote like *Prochloron,* an organism that has been found only growing symbiotically inside ascidians.

Bacterial photosynthesis

There are a number of photosynthetic bacteria that are not oxygenic (*e.g.,* the sulfur bacteria previously discussed). The evolutionary pathway that led to these bacteria diverged from the one that resulted in oxygenic organisms. In addition to the absence of oxygen production, nonoxygenic photosynthesis differs from oxygenic photosynthesis in two other ways: light of longer wavelengths is absorbed and used by pigments called bacteriochlorophylls, and reduced compounds other than water (such as hydrogen sulfide or organic molecules) provide the electrons needed for the reduction of carbon dioxide.

FACTORS THAT INFLUENCE THE RATE OF PHOTOSYNTHESIS

The rate of photosynthesis is defined in terms of the rate of oxygen production either per unit mass (or area) of green plant tissues or per unit weight of total chlorophyll. The amount of light, the carbon dioxide supply, the temperature, the water supply, and the availability of minerals are the most important environmental factors that directly affect the rate of photosynthesis in land plants. The rate of photosynthesis also is determined by the plant species and its physiological state—*e.g.,* its health, its maturity, and whether or not it is in flower.

Light intensity and temperature. As has been mentioned, the complex mechanism of photosynthesis includes a photochemical, or light-dependent, stage and an enzymatic, or dark, stage that involves chemical reactions.

These stages can be distinguished by studying the rates of photosynthesis at various degrees of light saturation (*i.e.*, intensity) and at different temperatures. Over a range of moderate temperatures and at low to medium light intensities (relative to the normal range of the plant species), the rate of photosynthesis increases as the intensity increases and is independent of temperature. As the light intensity increases to higher levels, however, the rate becomes increasingly dependent on temperature and less dependent on intensity; light "saturation" is achieved at a specific light intensity, and the rate then is dependent only on temperature if all other factors are constant. In the light-dependent range before saturation, therefore, the rate of photosynthesis is determined by the rates of photochemical steps. At high light intensities, some of the chemical reactions of the dark stage become rate-limiting. At light saturation, rate increases with temperature until a point is reached beyond which no further rate increase can occur. In many land plants, moreover, a process called photorespiration occurs at high light intensities and temperatures. Photorespiration competes with photosynthesis and limits further increases in the rate of photosynthesis, especially if the supply of water is limited (see below *Photorespiration*).

Carbon dioxide. Included among the rate-limiting steps of the dark stage of photosynthesis are the chemical reactions by which organic compounds are formed using carbon dioxide as a carbon source. The rates of these reactions can be increased somewhat by increasing the carbon dioxide concentration. During the past century, the level of carbon dioxide in the atmosphere has been rising due to the extensive combustion of fossil fuels. The atmospheric level of carbon dioxide climbed from about 0.028 percent in 1860 to 0.0315 percent by 1958 (when improved measurements began), and to 0.034 percent by 1981. This increase in carbon dioxide directly increases plant photosynthesis, but the size of the increase depends on the species and physiological condition of the plant. Furthermore, if increasing levels of atmospheric carbon dioxide result in climatic changes, including increased global temperatures as some meteorologists predict, these changes will affect photosynthesis rates.

Water. For land plants, water availability can function as a limiting factor in photosynthesis and plant growth. Besides the requirement for water in the photosynthetic reaction itself, water is transpired from the leaves; that is, water evaporates from the leaves to the atmosphere via the stomates. These stomates are small openings through the leaf epidermis, or outer skin; they permit the entry of carbon dioxide but also allow the exit of water vapour. The stomates open and close according to the physiological needs of the leaf. In hot and arid climates the stomates may close to conserve water, but this closure limits the entry of carbon dioxide and hence the rate of photosynthesis, while the wasteful process of photorespiration may increase. If the level of carbon dioxide in the atmosphere increases, more carbon dioxide could enter through a smaller opening of the stomates, so that more photosynthesis could occur with a given supply of water.

Minerals. Several minerals are required for healthy plant growth and for maximum rates of photosynthesis. Nitrate or ammonia, sulfate, phosphate, iron, magnesium, and potassium are required in substantial amounts for the synthesis of amino acids, proteins, coenzymes, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), chlorophyll and other pigments, and other essential plant constituents. Smaller amounts of such elements as manganese, copper, and chlorine are required in photosynthesis. Some other trace elements are needed for various nonphotosynthetic functions in plants.

Internal factors. Each plant species adapts to a range of environmental factors. Within this normal range of conditions, complex regulatory mechanisms in the plant's cells adjust the activities of enzymes (*i.e.*, organic catalysts). These adjustments maintain a balance in the overall photosynthetic process and control it in accordance with the needs of the whole plant. With a given plant species, for example, doubling the carbon dioxide level might cause a temporary increase of nearly twofold in the rate of photosynthesis; a few hours later, however, the rate might fall to the original level because photosynthesis had made more sucrose than the rest of the plant could use. By contrast, another plant species provided with such carbon dioxide enrichment might be able to use more sucrose and would continue to photosynthesize and to grow faster throughout most of its life cycle.

ENERGY EFFICIENCY OF PHOTOSYNTHESIS

The energy efficiency of photosynthesis is the ratio of the energy stored to the energy of light absorbed. The chemical energy stored is the difference between that contained in gaseous oxygen and organic compound products and the energy of water, carbon dioxide, and other reactants. The amount of energy stored can only be estimated because many products are formed, and these vary with the plant species and environmental conditions. If the equation for glucose formation given earlier is used to approximate the actual storage process, the production of one mole (*i.e.*, 6.02×10^{23} molecules; abbreviated N) of oxygen and one-sixth mole of glucose results in the storage of about 117 kilocalories (kcal) of chemical energy. This amount must then be compared to the energy of light absorbed to produce one mole of oxygen in order to calculate the efficiency of photosynthesis.

Light can be described as a wave of particles known as photons; these are units of energy, or light quanta. The quantity N photons is called an einstein. The energy of light varies inversely with the length of the photon waves; that is, the shorter the wavelength, the greater the energy content. The energy (e) of a photon is given by the equation $e = hc/\lambda$, where c is the velocity of light, h is Planck's constant, and λ is the light wavelength. The energy (E) of an einstein is $E = Ne = Nhc/\lambda = 28{,}600/\lambda$, when E is in kilocalories and λ is given in nanometres (nm; 1 nm = 10^{-9} metres). An einstein of red light with a wavelength of 680 nm has an energy of about 42 kcal. Blue light has a shorter wavelength and therefore more energy than red light. Regardless of whether the light is blue or red, however, the same number of einsteins are required for photosynthesis per mole of oxygen formed. The part of the solar spectrum used by plants has an estimated mean wavelength of 570 nanometres; therefore, the energy of light used during photosynthesis is approximately 28,600/570, or 50 kilocalories per einstein.

Energy of light

In order to compute the amount of light energy involved in photosynthesis, one other value is needed: the number of einsteins absorbed per mole of oxygen evolved. This is called the quantum requirement. The minimum quantum requirement for photosynthesis under optimal conditions is about nine. Thus the energy used is 9×50, or 450 kilocalories per mole of oxygen evolved. Therefore, the estimated maximum energy efficiency of photosynthesis is the energy stored per mole of oxygen evolved—117 kilocalories—divided by 450; that is, 117/450, or 26 percent.

Quantum requirement

The actual percentage of solar energy stored by plants is much less than the maximum energy efficiency of photosynthesis. An agricultural crop in which the biomass (total dry weight) stores as much as 1 percent of total solar energy received on an annual area-wide basis is exceptional, although a few cases of higher yields (perhaps as much as 3.5 percent in sugarcane) are reported. There are several reasons for this difference between the predicted maximum efficiency of photosynthesis and the actual energy stored in biomass. First, more than half of the incident sunlight is composed of wavelengths too long to be absorbed, while some of the remainder is reflected or lost to the leaves. Consequently, plants can at best absorb only about 34 percent of the incident sunlight. Second, plants must carry out a variety of physiological processes in such nonphotosynthetic tissues as roots and stems; these processes, as well as cellular respiration in all parts of the plant, use up stored energy. Third, rates of photosynthesis in bright sunlight sometimes exceed the needs of the plants, resulting in the formation of excess sugars and starch. When this happens, the regulatory mechanisms of the plant slow down the process of photosynthesis, allowing more absorbed sunlight to go unused. Fourth, in many plants, energy is wasted by the process of photorespiration. Finally, the growing season may last only a few

months of the year; sunlight received during other seasons is not used. Furthermore, it should be noted that if only agricultural products (*e.g.,* seeds, fruits, and tubers, rather than total biomass) are considered as the end product of the energy conversion process of photosynthesis, the efficiency falls even further.

CHLOROPLASTS, THE PHOTOSYNTHETIC UNITS OF GREEN PLANTS

The process of plant photosynthesis takes place entirely within the chloroplasts. Detailed studies of the role of these organelles date from the work of the British biochemist Robert Hill. About 1940 Hill discovered that green particles obtained from broken cells could produce oxygen from water in the presence of light and a chemical compound, such as ferric oxalate, able to serve as an electron acceptor. This process is known as the Hill reaction.

The Hill reaction

During the 1950s Daniel Arnon and other American biochemists prepared plant cell fragments in which not only the Hill reaction but also the synthesis of the energy-storage compound ATP occurred. In addition, the coenzyme NADP was used as the final acceptor of electrons, replacing the nonphysiological electron acceptors used by Hill. His procedures were refined further so that individual small pieces of isolated chloroplast membranes, or lamellae, could perform the Hill reaction. These small pieces of lamellae were then fragmented into pieces so small that they performed only the light reactions of the photosynthetic process. It is now possible also to isolate the entire chloroplast so that it can carry out the complete process of photosynthesis, from light absorption, oxygen formation, and the reduction of carbon dioxide to the formation of glucose and other products.

Structural features. The intricate structural organization of the photosynthetic apparatus is essential for the efficient performance of the complex process of photosynthesis. The chloroplast is enclosed in a double outer membrane, and its size approximates a spheroid about 2,500 nanometres thick and 5,000 nanometres long. Some single-celled algae have one chloroplast that occupies more than half the cell volume. Leaf cells of higher plants contain many chloroplasts, each approximately the size of the one in some algal cells.

Size and shape

When thin sections of a chloroplast are examined under the electron microscope (Figure 1), several features are apparent. Chief among these are the intricate internal membranes (*i.e.,* the lamellae) and the stroma, a colourless matrix in which the lamellae are embedded. Also visible are starch granules, which appear as dense bodies.

The stroma is basically a solution of enzymes and small molecules. The dark reactions occur in the stroma, the soluble enzymes of which catalyze the conversion of carbon dioxide and minerals to carbohydrates and other organic compounds. The capacity for carbon fixation and reduction is lost if the outer membrane of the chloroplast is broken, allowing the stroma enzymes to leak out.

By courtesy of the University of California, Lawrence Berkeley Laboratory

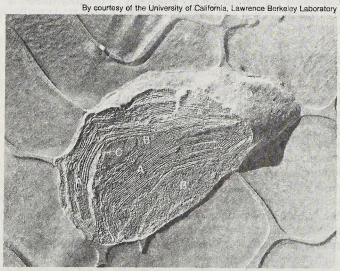


Figure 1: *Electron micrograph of an isolated spinach chloroplast prepared by the freeze-etching method.* (A) Lamellae, which are layers of lipid and protein with chlorophyll, other pigments, and enzymes, are the sites of light absorption and splitting of the water to give oxygen and hydrogen. (B) Stroma, containing soluble enzymes, is the site of conversion of carbon dioxide and hydrogen to carbohydrates. (C) Starch granule is a storage product of photosynthesis. (D) Surface of the chloroplast, and (E) frozen solution in which the chloroplast was suspended.

A single lamella, which contains all the photosynthetic pigments, is approximately 10–15 nanometres thick. The lamellae exist in more-or-less flat sheets, a few of which extend through much of the length of the chloroplast. Examination of cross sections of lamellae under the electron microscope shows that their edges are joined to form closed hollow disks that are called thylakoids ("saclike"). The chloroplasts of most higher plants have regions, called grana, in which the thylakoids are very tightly stacked. When viewed by electron microscopy at an oblique angle, the grana appear as stacks of disks. When viewed in cross section (Figure 1), it is apparent that some thylakoids extend from one grana through the stroma into other grana. The thin aqueous spaces inside the thylakoids are believed to be connected with each other via these stroma thylakoids. These thylakoid spaces are isolated from the stroma spaces by the relatively impermeable lamellae.

Grana

The light reactions occur exclusively in the thylakoids. The complex structural organization of lamellae is required for proper thylakoid function; intact thylakoids apparently are necessary for the formation of ATP. Thylakoids that have been broken down to smaller units can no longer form ATP, even when the conversion of light into chemical energy occurs during electron transport in these units. Such lamellar fragments can carry out the Hill reaction, with the transfer of electrons from water to $NADP^+$.

Chemical composition of lamellae. *Lipids.* Lamellae consist of about equal amounts of lipids and proteins. About one-fourth of the lipid portion of the lamellae consists of pigments and coenzymes; the remainder consists of various lipids, including polar compounds such as phospholipids and galactolipids. These polar lipid molecules have "head" groups that attract water (*i.e.,* are hydrophilic) and fatty acid "tails" that are oil soluble and repel water (*i.e.,* are hydrophobic). When polar lipids are placed in an aqueous environment, they can line up with the fatty acid tails side by side. A second layer of phospholipids forms tail-to-tail with the first, establishing a lipid bilayer in which the hydrophilic heads are in contact with the aqueous solution on each side of the bilayer. Sandwiched between the heads are the hydrophobic tails, creating a hydrophobic environment from which water is excluded. This lipid bilayer is an essential feature of all biological membranes (see CELLS: *The plasma membrane*). The hydrophobic parts of proteins and lipid-soluble cofactors and pigments are dissolved or embedded in the lipid bilayer. Lamellar membranes can function as electrical insulating material and permit a charge, or potential difference, to develop across the membrane. Such a charge can be a source of chemical or electrical energy.

Approximately one-fifth of the lamellar lipids are chlorophyll molecules; one type, chlorophyll *a,* is more abundant than the second type, chlorophyll *b.* The chlorophyll molecules are specifically bound to small protein molecules. Most of these chlorophyll-proteins are "light-harvesting" pigments. These absorb light and pass its energy on to special chlorophyll *a* molecules that are directly involved in the conversion of light energy to chemical energy. When one of these special chlorophyll *a* molecules is excited by light energy (as described later), it gives up an electron. There are two types of these special chlorophyll *a* molecules: one, called P_{680}, has an absorption spectrum that peaks at 684 nanometres; the other, called P_{700}, shows an absorption peak at 700 nanometres.

Although chlorophylls are the main light-absorbing molecules in green plants, other pigments such as carotenes and carotenoids (which are responsible for the yellow-orange colour of carrots) also can absorb light and may supplement chlorophyll as the light-absorbing molecules in some plant cells. The light energy absorbed by these pigments must be passed to chlorophyll before conversion to chemical energy can occur.

Pigments

Proteins. Many of the lamellar proteins are components of the chlorophyll–protein complexes described above.

Electron carrier molecules

Other proteins include enzymes and protein-containing coenzymes. Enzymes are required as organic catalysts for specific reactions within the lamellae. Protein coenzymes, also called cofactors, include important electron carrier molecules called cytochromes, which are iron-containing pigments with the pigment portions attached to protein molecules. During electron transfer, an electron is accepted by an iron atom in the pigment portion of a cytochrome molecule, which thus is reduced; then the electron is transferred to the iron atom in the next cytochrome carrier in the electron transfer chain, thus oxidizing the first cytochrome and reducing the next one in the chain.

In addition to the metal atoms found in the pigment portions of cytochrome molecules, metal atoms also are found in other protein molecules of the lamellae. In proteins with a total molecular weight of 900,000 (based on the weight of hydrogen as one), there are two atoms of manganese, 10 atoms of iron, and six atoms of copper. These metal atoms are required for the catalytic activity of some of the enzymes important in photosynthesis. The manganese atoms are involved in water-splitting and oxygen formation. Both copper- and iron-containing proteins function in electron transport between water and the final electron-acceptor molecule of the light stage of photosynthesis, an iron-containing protein called ferredoxin. Ferredoxin is a soluble component in the chloroplasts. In its reduced form, it gives electrons directly to the systems that reduce nitrate and sulfate and via NADPH to the system that reduces carbon dioxide. A copper-containing protein called plastocyanin (PC) carries electrons at one point in the electron transport chain. PC molecules are water soluble and can move through the inner space of the thylakoids, carrying electrons from one place to another.

Quinones. Small molecules called plastoquinones are found in substantial numbers in the lamellae. Like the cytochromes, quinones have important roles in carrying electrons between the components of the light reactions. Since they are lipid soluble, they can diffuse through the membrane. They can carry one or two electrons and, in their reduced form (with added electrons), they carry hydrogen atoms that can be released as hydrogen ions when the added electrons are passed on, for example, to a cytochrome.

THE PROCESS OF PHOTOSYNTHESIS: THE LIGHT REACTIONS

Light absorption and energy transfer. The light energy absorbed by a chlorophyll molecule excites some electrons within the structure of the molecule to higher energy levels, or excited states. Light of shorter wavelength (such as blue) has more energy than light of longer wavelength (such as red), so that absorption of blue light creates an excited state of higher energy. A molecule raised to this higher energy state quickly gives up the "extra" energy as heat and falls to its lowest excited state. This lowest excited state is similar to that of a molecule that has just absorbed the longest wavelength light capable of exciting it. In the case of chlorophyll *a*, this lowest excited state corresponds to that of a molecule that has absorbed red light of about 680 nanometres.

The return of a chlorophyll *a* molecule from its lowest excited state to its original low-energy state (ground state) requires the release of the extra energy of the excited state. This can occur in one of several ways. In photosynthesis, most of this energy is conserved as chemical energy by the transfer of an electron from a special chlorophyll *a* molecule (P_{680} or P_{700}) to an electron acceptor. When this electron transfer is blocked by inhibitors, such as the herbicide dichlorophenylmethylurea (DCMU), or by low temperature, the energy can be released as red light. Such re-emission of light is called fluorescence. The examination of fluorescence from photosynthetic material in which electron transfer has been blocked has proved to be a useful tool for scientists studying the light reactions.

The pathway of electrons. The general features of a widely accepted mechanism for photoelectron transfer, in which two light reactions occur during the transfer of electrons from water to carbon dioxide, were proposed by Robert Hill and Derek Bendall, in 1960. A modified scheme for this mechanism is shown in Figure 2. In this figure the vertical scale represents the relative potential (in volts) of various cofactors of the electron-transfer chain to be oxidized or reduced. Molecules that in their oxidized form have the strongest affinity for electrons (*i.e.*, are strong oxidizing agents) are near the bottom of the scale. Molecules that in their oxidized form are difficult to reduce are near the top of the scale; once they have accepted electrons, these molecules are strong reducing agents.

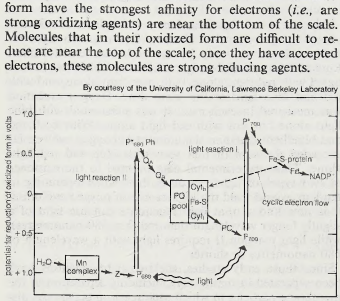


Figure 2: *Flow of electrons during the light reaction stage of photosynthesis.*
Arrows pointing upward represent light reactions that increase the chemical potential; arrows slanting downward represent flow of electrons via carriers in the membrane (see text).

Electron transfer chains

The actual photochemical steps are indicated by the two vertical arrows, which signify that the special pigments P_{680} and P_{700} receive light energy from the light-harvesting chlorophyll-protein molecules and are raised in energy from their ground state to excited states, symbolized as P^*_{680} and P^*_{700}. In their excited state, these pigments are extremely strong reducing agents that quickly transfer electrons to the first acceptor. These first acceptors also are strong reducing agents and rapidly pass electrons to more stable carriers. In light reaction II the first acceptor may be pheophytin (Ph; a molecule similar to chlorophyll), which also has a strong reducing potential and quickly transfers electrons to the next acceptor. Q_A and Q_B are special quinones, similar to plastoquinone. They receive electrons from pheophytin and pass them to the intermediate electron carriers, which include the plastoquinone (PQ) pool and the cytochromes b and f (Cyt_b and Cyt_f) associated in a complex with an iron-sulfur protein (Fe-S).

In light reaction I the identity of the first electron acceptor, X, is not known. It passes electrons on to iron-sulfur proteins (Fe-S-protein) in the lamellar membrane, after which the electrons flow to ferredoxin (Fd), a small, water-soluble iron-sulfur protein. When $NADP^+$ and a suitable enzyme are present, two ferredoxin molecules, carrying one electron each, transfer two electrons to $NADP^+$, which picks up a proton (*i.e.*, a hydrogen ion) and becomes NADPH.

Reduction of P_{680}^{+} and P_{700}^{+}

Each time a P_{680} or P_{700} molecule gives up an electron, it returns to its ground (unexcited) state, but with a positive charge due to the loss of the electron. These positively charged ions are extremely strong oxidizing agents that remove an electron from a suitable donor. The P_{680}^{+} of light reaction II is capable of taking electrons from water in the presence of appropriate catalysts. There is good evidence that two or more manganese atoms complexed with protein are involved in this catalysis, taking four electrons from two water molecules (with release of four hydrogen ions). The manganese-protein complex gives up these electrons one at a time via an unidentified carrier Z to P_{680}^{+}, reducing it to P_{680}. When manganese is selectively removed by chemical treatment, the thylakoids lose the capacity to oxidize water, but all other parts of the electron pathway remain intact.

In light reaction I, P_{700}^{+} recovers electrons from plastocyanin (PC), which in turn receives them from intermediate carriers, including the plastoquinone pool and cytochrome b and cytochrome f molecules. The pool of intermediate carriers may receive electrons from water via light reaction II and Q_A and Q_B. Transfer of electrons from water to ferredoxin via the two light reactions and interme-

diate carriers is called noncyclic electron flow. Alternately, electrons may be transferred only by light reaction I, in which case they are recycled from ferredoxin back to the intermediate carriers (dashed line, Figure 2). This process is called cyclic electron flow.

Evidence of two light reactions. Many lines of evidence support the concept of electron flow via two light reactions. An early study by the U.S. biochemist Robert Emerson employed the algae *Chlorella*, which was illuminated with red light alone, with blue light alone, and with red and blue light at the same time. Oxygen evolution was measured in each case. It was substantial with blue light alone but not with red light alone. With both red and blue light together, the amount of oxygen evolved far exceeded the sum of that seen with blue and red light alone. These experimental data pointed to the existence of two types of light reactions that, when operating in tandem, would yield the highest rate of oxygen evolution. It is now known that light reaction I can use light of a slightly longer wavelength than red ($\lambda = 680$ nanometres), while light reaction II requires light with a wavelength of 680 nanometres or shorter.

Since those early studies, the two light reactions have been separated in many ways, including separation of the membrane particles in which each reaction occurs. As discussed previously, lamellae can be disrupted mechanically into fragments that absorb light energy and break the bonds of water molecules (*i.e.*, oxidize water) to produce oxygen, hydrogen ions, and electrons. These electrons can be transferred to ferredoxin, the final electron acceptor of the light stage. No transfer of electrons from water to ferredoxin occurs if the herbicide DCMU is present. The subsequent addition of certain reduced dyes (*i.e.*, electron donors) restores the light reduction of $NADP^+$ but without oxygen production, suggesting that light reaction I but not light reaction II is functioning. It is now known that DCMU blocks the transfer of electrons from Q_A to the PQ pool (see Figure 2).

Lamellar fragments

When treated with certain detergents, lamellae can be broken down into smaller particles capable of carrying out single light reactions. One type of particle can absorb light energy, oxidize water, and produce oxygen (light reaction II), but a special dye molecule must be supplied to accept the electrons. In the presence of electron donors, such as a reduced dye, a second type of lamellar particle can absorb light and transfer electrons from the electron donor to ferredoxin (light reaction I).

Photosystems I and II. The structural and photochemical properties of the minimum particles capable of performing light reactions I and II have received much study. Treatment of lamellar fragments with neutral detergents releases these particles, designated photosystem I and photosystem II, respectively. Subsequent harsher treatment (with charged detergents) and separation of the individual polypeptides with electrophoretic techniques has helped identify the components of the photosystems. Each photosystem consists of a light-harvesting complex and a core complex. Each core complex contains a reaction centre with the pigment (either P_{700} or P_{680}) that can be photochemically oxidized, together with electron acceptors and electron donors. In addition, the core complex has some 40 to 60 chlorophyll molecules bound to proteins. In addition to the light absorbed by the chlorophyll molecules in the core complex, the reaction centres receive a major part of their excitation from the pigments of the light-harvesting complex.

Quantum requirements. The quantum requirements of the individual light reactions of photosynthesis are defined as the number of light photons absorbed for the transfer of one electron. The quantum requirement for each light reaction has been found to be approximately one photon. The total number of quanta required, therefore, to transfer the four electrons that result in the formation of one molecule of oxygen via the two light reactions should be four times two, or eight. It appears, however, that additional light is absorbed and used to form ATP by a cyclic photophosphorylation pathway (see next section). The actual quantum requirement, therefore, probably is nine to 10.

THE PROCESS OF PHOTOSYNTHESIS: THE CONVERSION OF LIGHT ENERGY TO ATP

The electron transfers of the light reactions provide the energy for the synthesis of two compounds vital to the dark reactions: NADPH and ATP. The previous section explained how noncyclic electron flow results in the reduction of $NADP^+$ to NADPH. In this section, the synthesis of the energy-rich compound ATP is described.

ATP is formed by the addition of a phosphate group to a molecule of adenosine diphosphate (ADP); or to state it in chemical terms, by the phosphorylation of ADP. This reaction requires a substantial input of energy, much of which is captured in the bond that links the added phosphate group to ADP. Because light energy powers this reaction in the chloroplasts, the production of ATP during photosynthesis is referred to as photophosphorylation.

Unlike the production of NADPH, the photophosphorylation of ADP occurs in conjunction with both cyclic and noncyclic electron flow (see Figure 2). In fact researchers speculate that the sole purpose of cyclic electron flow may be for photophosphorylation, since this process involves no net transfer of electrons to reducing agents. The relative amounts of cyclic and noncyclic flow may be adjusted in accordance with changing physiological needs for ATP and reduced ferrodoxin and NADPH in chloroplasts. In contrast to electron transfer in light reactions I and II, which can occur in membrane fragments, intact thylakoids are required for efficient photophosphorylation. This requirement stems from the special nature of the mechanism linking photophosphorylation to electron flow in the lamellae.

The chemiosmotic theory of ATP formation

The theory relating the formation of ATP to electron flow in the membranes of both chloroplasts and mitochondria (the organelles responsible for ATP formation during cellular respiration) was first proposed by the English biochemist Peter Mitchell. This chemiosmotic theory has been somewhat modified to fit later experimental facts, and there is still debate over many of the details. The general features, however, are widely accepted. A central feature is the formation of a hydrogen ion (proton) concentration gradient and an electrical charge across intact lamellae. The potential energy stored by the proton gradient and electrical charge is then used to drive the energetically unfavourable conversion of ADP and inorganic phosphate (P_i) to ATP and water.

The manganese-protein complex associated with light reaction II is exposed to the interior of the thylakoid. Consequently, the oxidation of water during light reaction II leads to release of hydrogen ions (protons) into the inner thylakoid space. Furthermore, it is likely that photoreaction II entails the transfer of electrons across the lamella toward its outer face, so that when plastoquinone molecules are reduced they can receive protons from the outside of the thylakoid. When these reduced plastoquinone molecules are oxidized, giving up electrons to the cytochrome-iron-sulfur complex, protons are released inside the thylakoid. Because the lamella is impermeable to them, the release of protons inside the thylakoid by oxidation of both water and plastoquinone leads to a higher concentration of protons inside the thylakoid than outside it. In other words, a proton gradient is established across the lamella. The movement of electrons (negatively charged particles) outward across the lamella during both light reactions results in the establishment of an electrical charge across the lamella. (Some scientists believe, however, that the proton gradient and electrical charge required for ATP formation need not be between inner and outer thylakoid space but only within the membrane.)

An enzyme complex located partly in and on the lamellae catalyzes the reaction in which ATP is formed from ADP and inorganic phosphate. The reverse of this reaction is catalyzed by an enzyme called ATP-ase, hence the enzyme complex is sometimes called an ATP-ase complex. It is also called the coupling factor. It consists of hydrophilic polypeptides (F_1), which project from the outer surface of the lamellae, and hydrophobic polypeptides (F_0), which are embedded inside the lamellae. Researchers hypothesize that F_0 forms a channel that permits protons to flow through the lamellar membrane to F_1. The enzymes in F_1

then catalyze ATP formation, using both the proton supply and the lamellar transmembrane charge.

In summary, the use of light energy for ATP formation occurs indirectly: a proton gradient and electrical charge—built up in or across the lamellae as a consequence of electron flow in the light reactions—provide the energy to drive the synthesis of ATP from ADP and P_i.

THE PROCESS OF PHOTOSYNTHESIS: CARBON FIXATION AND REDUCTION

The assimilation of carbon into organic compounds is the result of a complex series of enzymatically regulated chemical reactions—the dark reactions. This term is something of a misnomer, for these reactions can take place in either light or darkness. Furthermore, some of the enzymes involved in the so-called dark reactions become inactive in prolonged darkness.

Elucidation of the carbon pathway. Radioactive isotopes of carbon (^{14}C) and phosphorus (^{32}P) have been valuable in identifying the intermediate compounds formed during carbon assimilation. A photosynthesizing plant does not strongly discriminate between the natural carbon isotopes and ^{14}C. During photosynthesis in the presence of $^{14}CO_2$, the compounds formed become labelled with the radioisotope. During very short exposures, only the first intermediates in the carbon-fixing pathway become labelled. Early investigations showed that some radioactive products were formed even when the light was turned off and the $^{14}CO_2$ was added just afterward in the dark, confirming the nature of the carbon fixation as a "dark" reaction.

The U.S. biochemist Melvin Calvin, a Nobel Prize recipient for his work on the carbon reduction cycle, allowed green plants to photosynthesize in the presence of radioactive carbon dioxide for a few seconds under various experimental conditions. Products that became labelled with radioactive carbon during Calvin's experiments included a three-carbon compound called 3-phosphoglycerate (abbreviated PGA, see Figure 3), sugar phosphates, amino acids, sucrose, and carboxylic acids. When photosynthesis was stopped after two seconds, the principal radioactive product was PGA, which therefore was identified as the first compound formed during carbon dioxide fixation in green plants.

Identification of the first compound formed

Further studies with ^{14}C as well as with inorganic phosphate labelled with ^{32}P led to the mapping of the carbon fixation and reduction pathway called the reductive pentose phosphate cycle (RPP cycle). An additional pathway for carbon transport in certain plants was later discovered in other laboratories (see below *Carbon fixation via C_4 acids*). All the steps in these pathways can be carried out in the laboratory by isolated enzymes in the dark. Several steps require the ATP or NADPH generated by the light reactions. In addition, some of the enzymes are fully active only when conditions simulate those in green cells exposed to light. In vivo, these enzymes are active during photosynthesis but not in the dark.

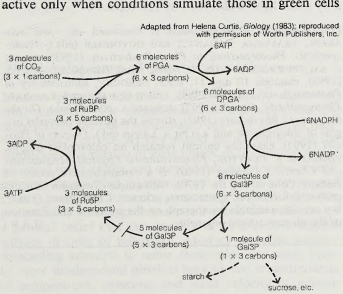


Figure 3: *Pathway of carbon dioxide fixation and reduction in photosynthesis, the reductive pentose phosphate cycle.* The diagram represents one complete turn of the cycle, with the net production of one molecule of Gal3P. The nine molecules of ATP and six molecules of NADPH come from the light reactions (see text).

The reductive pentose phosphate cycle. *Overall reaction.* The RPP cycle, in which carbon is fixed, reduced, and utilized, involves the formation of intermediate sugar phosphates in a cyclic sequence (Figure 3). One complete RPP cycle incorporates three molecules of carbon dioxide and produces one molecule of the three-carbon compound glyceraldehyde-3-phosphate (Gal3P). This three-carbon sugar phosphate usually is either exported from the chloroplasts or is converted to starch (dashed lines).

Figure 3 summarizes the main steps that take place during one complete turn of the RPP cycle. ATP and NADPH formed during the light reactions are utilized for key steps in this pathway and provide the energy and reducing equivalents (*i.e.*, electrons) to drive the sequence in the direction shown. For each molecule of carbon dioxide that is fixed, two molecules of NADPH and three molecules of ATP from the light reactions are required. The overall reaction can be represented as follows:

$$9ATP + 6NADPH + 3CO_2 \longrightarrow Gal3P + 6NADP^+ + 9ADP + 8P_i$$

The cycle is composed of four stages: (1) carboxylation, (2) reduction, (3) isomerization/condensation/dismutation, and (4) phosphorylation.

Carboxylation. The initial incorporation of carbon dioxide, which is catalyzed by the enzyme ribulose 1,5-bisphosphate carboxylase, proceeds by the addition of carbon dioxide to the five-carbon compound ribulose 1,5-bisphosphate (RuBP) and the splitting of the resulting unstable six-carbon compound into two molecules of PGA, a three-carbon compound. As indicated in Figure 3, this reaction occurs three times during each complete turn of the cycle; thus, six molecules of PGA are produced.

Reduction. The six molecules of PGA are first phosphorylated with ATP by the enzyme PGA-kinase, yielding six molecules of 1,3-diphosphoglycerate (DPGA). These then are reduced with NADPH and the enzyme glyceraldehyde-3-phosphate dehydrogenase to give six molecules of Gal3P. These reactions are the reverse of two steps of the process glycolysis in cellular respiration (see METABOLISM).

Isomerization/condensation/dismutation. For each complete RPP cycle, one of the Gal3P molecules, with its three carbon atoms, is the net product and may be transferred out of the chloroplast or converted to starch inside the chloroplast (dashed lines, Figure 3). For the cycle to regenerate, the other five Gal3P molecules (with a total of 15 carbon atoms) must be converted back to three molecules of five-carbon RuBP. The conversion of Gal3P to RuBP begins with a complex series of enzymatically regulated reactions (not shown in Figure 3) that lead to the synthesis of the five-carbon compound ribulose-5-phosphate (Ru5P). The occurrence of this complex series, which entails isomerization, condensation, and dismutation reactions, is indicated by the broken arrow in Figure 3.

Conversion of Gal3P to Ru5P

Phosphorylation. The three molecules of Ru5P are converted to the carboxylation substrate, RuBP, by the enzyme phosphoribulokinase, using ATP. This reaction, shown below, completes the cycle.

$$3Ru5P + 3ATP \longrightarrow 3RuBP + 3ADP$$

Regulation of the cycle. Photosynthesis cannot occur at night, but the respiratory process of glycolysis—which uses some of the same reactions as the RPP cycle, except in the reverse—does take place. Thus, some steps in the RPP cycle would be wasteful if allowed to occur in the dark because they would counteract the reactions of glycolysis. For this reason, some enzymes of the RPP cycle are "turned off" (*i.e.*, become inactive) in the dark.

Even in the presence of light, changes in physiological conditions frequently necessitate adjustments in the relative rates of reactions of the RPP cycle, so that enzymes for some reactions change in their catalytic activity. These alterations in enzyme activity typically are brought about by changes in levels of such chloroplast components as reduced ferredoxin, acids, and soluble components (*e.g.*, P_i and magnesium ions).

Products of carbon reduction. The most important use of Gal3P is its export from the chloroplasts to the cyto-

plasm of green cells, where it is used for biosynthesis of products needed by the plant. In land plants, a principal product is sucrose, which is translocated from the green cells of the leaves to other parts of the plant. Other key products include the carbon skeletons of certain primary amino acids, such as alanine, glutamate, and aspartate. To complete the synthesis of these compounds, amino groups are added to the appropriate carbon skeletons made from Gal3P. Sulfur amino acids such as cysteine are formed by adding sulfhydryl groups and amino groups. Other biosynthesis pathways lead from Gal3P to lipids, pigments, and most of the constituents of green cells.

Synthesis of starch

Starch synthesis and accumulation in the chloroplasts occurs particularly when photosynthetic carbon fixation exceeds the needs of the plant. Under such circumstances, sugar phosphates accumulate in the cytoplasm, binding cytoplasmic P_i. The export of Gal3P from the chloroplasts is tied to a one-for-one exchange of P_i for Gal3P, so less cytoplasmic P_i results in decreased export of Gal3P and decreased P_i in the chloroplast. These changes trigger alterations in the activities of regulated enzymes, leading in turn to increased starch synthesis. This starch can be broken down at night and used as a source of reduced carbon and energy for the physiological needs of the plant. Too much starch in the chloroplasts leads to diminished rates of photosynthesis, however. Thus, under what would seem to be the ideal photosynthetic conditions of a bright, warm day, many plants in fact have slower rates of photosynthesis in the afternoon.

Photorespiration. Under conditions of high light intensity, hot weather, and water limitation, the productivity of the RPP cycle is limited in many plants by the occurrence of photorespiration. This process converts sugar phosphates back to carbon dioxide; it is initiated by the oxygenation of RuBP (*i.e.,* the combination of gaseous oxygen [O_2] with RuBP). This oxygenation reaction yields only one molecule of PGA and one molecule of a two-carbon acid, phosphoglycollate, which is subsequently converted in part to carbon dioxide. The reaction of oxygen with RuBP is in direct competition with the carboxylation reaction (CO_2 + RuBP) that initiates the RPP cycle and is, in fact, catalyzed by the same protein, ribulose 1,5-bisphosphate carboxylase. The relative concentrations of oxygen and carbon dioxide within the chloroplasts determine whether oxygenation or carboxylation is favoured. The concentration of oxygen inside the chloroplasts may be higher than atmospheric (20 percent) owing to photosynthetic oxygen evolution, whereas the internal carbon dioxide concentration may be lower than atmospheric (0.035 percent) owing to photosynthetic uptake. Any increase in the internal carbon dioxide pressure tends to help the carboxylation reaction compete more effectively with oxygenation.

Carbon fixation via C_4 acids. Certain plants—including the important crops sugarcane and corn (maize), as well as other diverse species believed to have evolved in the drier tropical areas—have developed a special mechanism of carbon fixation that largely prevents photorespiration. The leaves of these plants have special anatomy and biochemistry. In particular, photosynthetic functions are divided between mesophyll and bundle sheath leaf cells. The carbon fixation pathway begins in the mesophyll cells, where carbon dioxide is added to the three-carbon acid phosphoenolpyruvate (PEPA) by an enzyme called phosphoenolpyruvate carboxylase. The product of this reaction is the four-carbon acid oxaloacetate, which is reduced to malate, another four-carbon acid, in one form of C_4 pathway. Malate then is translocated to bundle sheath cells, which are located near the vascular system of the leaf. There, malate enters the chloroplasts and is oxidized and decarboxylated (*i.e.,* loses CO_2) by malic enzyme. This yields carbon dioxide, which is fed into the RPP cycle of the bundle sheath cells, and pyruvate, a three-carbon acid that is translocated back to the mesophyll cells. In the mesophyll chloroplasts, the enzyme pyruvate orthophosphate dikinase (PPDK) uses ATP and P_i to convert pyruvate back to PEPA, completing the C_4 cycle. There are several variations of this pathway in different species. For example, the amino acids aspartate and alanine can substitute for malate and pyruvate in some species.

Advantage of the C_4 pathway

The C_4 pathway acts as a shuttle for carrying carbon dioxide into the chloroplasts of the bundle sheath cells, where it is used in carbohydrate synthesis. The resulting higher level of internal carbon dioxide in these chloroplasts serves to increase the ratio of carboxylation to oxygenation, thus minimizing photorespiration. Although the plant must expend extra energy to drive this shuttle, the energy loss is more than compensated by the near elimination of photorespiration under conditions where it would otherwise occur. Sugarcane and certain other plants that employ this pathway have the highest annual yields of biomass of all species.

THE MOLECULAR BIOLOGY OF PHOTOSYNTHESIS

Synthesis of photosynthetic enzymes

Oxygenic photosynthesis occurs in both prokaryotic cells (cyanophytes) and eukaryotic cells (algae and higher plants). In eukaryotic cells, which contain chloroplasts and a nucleus, the genetic information needed for the reproduction of the photosynthetic apparatus is contained partly in the chloroplast chromosome and partly in chromosomes of the nucleus. For example, the carboxylation enzyme ribulose 1,5-bisphosphate carboxylase is a large protein molecule comprising a complex of eight large polypeptide subunits and eight small polypeptide subunits. The gene for the large subunits is located in the chloroplast chromosome, while the gene for the small subunits is in the nucleus. Transcription of the DNA of the nuclear gene yields messenger RNA (MRNA) that encodes the information for the synthesis of the small polypeptides. During this synthesis, which occurs on the cytoplasmic ribosomes, some extra amino acid residues are added to form a recognition leader on the end of the polypeptide chain. This leader is recognized by special receptor sites on the outer chloroplast membrane; these receptor sites then allow the polypeptide to penetrate the membrane and enter the chloroplast. The leader is removed and the small subunits combine with the large subunits, which have been synthesized on chloroplast ribosomes according to MRNA transcribed from the chloroplast DNA. The expression of nuclear genes that code for proteins needed in the chloroplasts appears to be under control of events in the chloroplasts in some cases; for example, the synthesis of some nuclear-encoded chloroplast enzymes may occur only when light is absorbed by chloroplasts.

BIBLIOGRAPHY. Photosynthesis is discussed in YASH PAL ABROL, PRASANNA MOHANTY, and GOVINDJEE (eds.), *Photosynthesis: Photoreactions to Plant Productivity* (1993); FRANK B. SALISBURY and CLEON W. ROSS, *Plant Physiology,* 4th ed. (1992), chapters 10 and 11; R.P.F. GREGORY, *Biochemistry of Photosynthesis,* 3rd ed. (1989); and D.O. HALL and K.K. RAO, *Photosynthesis,* 4th ed. (1987). JEROME A. SCHIFF (ed.), *On the Origins of Chloroplasts* (1982), studies the possible bacterial origins of chloroplasts; and RALPH A. LEWIN (ed.), *Origins of Plastids* (1993), examines current research on chloroplast origins. RODERICK K. CLAYTON, *Photosynthesis: Physical Mechanisms and Chemical Patterns* (1980), is a comprehensive review of research done during the 1970s, with ample discussion of background information. J. KENNETH HOOBER, *Chloroplasts* (1984), is a concise, readable monograph on the structure and function of the photosynthetic organelle.

(J.A.B.)

Principles of Physical Science

Physical science, like all the natural sciences, is concerned with describing and relating to one another those experiences of the surrounding world that are shared by different observers and whose description can be agreed upon. One of its principal fields, physics, deals with the most general properties of matter, such as the behaviour of bodies under the influence of forces, and with the origins of those forces. In the discussion of this question, the mass and shape of a body are the only properties that play a significant role, its composition often being irrelevant. Physics, however, does not focus solely on the gross mechanical behaviour of bodies, but shares with chemistry the goal of understanding how the arrangement of individual atoms into molecules and larger assemblies confers particular properties. Moreover, the atom itself may be analyzed into its more basic constituents and their interactions.

The present opinion, rather generally held by physicists, is that these fundamental particles and forces, treated quantitatively by the methods of quantum mechanics, can reveal in detail the behaviour of all material objects. This is not to say that everything can be deduced mathematically from a small number of fundamental principles, since the complexity of real things defeats the power of mathematics or of the largest computers. Nevertheless, whenever it has been found possible to calculate the relationship between an observed property of a body and its deeper structure, no evidence has ever emerged to suggest that the more complex objects, even living organisms, require that special new principles be invoked, at least so long as only matter, and not mind, is in question. The physical scientist thus has two very different roles to play: on the one hand, he has to reveal the most basic constituents and the laws that govern them; and, on the other, he must discover techniques for elucidating the peculiar features that arise from complexity of structure without having recourse each time to the fundamentals.

This modern view of a unified science, embracing fundamental particles, everyday phenomena, and the vastness of the Cosmos, is a synthesis of originally independent disciplines, many of which grew out of useful arts. The extraction and refining of metals, the occult manipulations of alchemists, and the astrological interests of priests and politicians all played a part in initiating systematic studies that expanded in scope until their mutual relationships became clear, giving rise to what is customarily recognized as modern physical science.

For a survey of the major fields of physical science and their development, see the articles PHYSICAL SCIENCES, THE; and EARTH SCIENCES, THE. For coverage of various related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 111–133, 10/31, and 10/32, and the *Index*.

This article is divided into the following sections:

The development of quantitative science 807
The Newtonian paradigm 810
Interplay of experiment and theory 811
  Characteristic experimental procedures 812
    Unexpected observation
    The not-so-unexpected chance
    Qualitative tests to distinguish alternative theories
    Direct comparison of theory and experiment
    Compilation of data
    Tests of fundamental concepts
  Characteristic theoretical procedures 813
    Direct solution of fundamental equations
    Dissection
    Simplified models
    Recasting of basic theory
  General observations 814
Concepts fundamental to the attitudes and methods of physical science 815
  Fields 815
    Gradient
    Line integral
    Potential
    Conservative forces
    Field lines
    Gauss's theorem
    Images
    Divergence and Laplace's equation
    Non-conservative fields
    Examples of differential equations for fields
  Conservation laws and extremal principles 818
  Fundamental constituents of matter 820
    Development of the atomic theory
    Rise of quantum mechanics
    Developments in particle physics
  Simplicity and complexity 823
  Symmetry 823
  Entropy and disorder 824
  Chaos 825
Bibliography 827

The development of quantitative science

Modern physical science is characteristically concerned with numbers—the measurement of quantities and the discovery of the exact relationship between different measurements. Yet this activity would be no more than the compiling of a catalog of facts unless an underlying recognition of uniformities and correlations enabled the investigator to choose what to measure out of an infinite range of choices available. Proverbs purporting to predict weather are relics of science prehistory and constitute evidence of a general belief that the weather is, to a certain degree, subject to rules of behaviour. Modern scientific weather forecasting attempts to refine these rules and relate them to more fundamental physical laws so that measurements of temperature, pressure, and wind velocity at a large number of stations can be assembled into a detailed model of the atmosphere whose subsequent evolution can be predicted—not by any means perfectly but almost always more reliably than was previously possible.

Between proverbial weather lore and scientific meteorology lies a wealth of observations that have been classified and roughly systematized into the natural history of the subject—for example, prevailing winds at certain seasons, more or less predictable warm spells such as Indian summer, and correlation between Himalayan snowfall and intensity of monsoon. In every branch of science this preliminary search for regularities is an almost essential background to serious quantitative work, and in what follows it will be taken for granted as having been carried out.

Compared to the caprices of weather, the movements of the stars and planets exhibit almost perfect regularity, and so the study of the heavens became quantitative at a very early date, as evidenced by the oldest records from China and Babylon. Objective recording and analysis of these motions, when stripped of the astrological interpretations that may have motivated them, represent the beginning of scientific astronomy. The heliocentric planetary model (*c.* 1510) of the Polish astronomer Nicolaus Copernicus, which replaced the Ptolemaic geocentric model, and the precise description of the elliptical orbits of the planets (1609) by the German astronomer Johannes Kepler, based

Initial achievements of modern quantitative science

on the inspired interpretation of centuries of patient observation that had culminated in the work of Tycho Brahe of Denmark, may be regarded fairly as the first great achievements of modern quantitative science.

A distinction may be drawn between an observational science like astronomy, where the phenomena studied lie entirely outside the control of the observer, and an experimental science such as mechanics or optics, where the investigator sets up the arrangement to his own taste. In the hands of Isaac Newton not only was the study of colours put on a rigorous basis but a firm link also was forged between the experimental science of mechanics and observational astronomy by virtue of his law of universal gravitation and his explanation of Kepler's laws of planetary motion. Before proceeding as far as this, however, attention must be paid to the mechanical studies of Galileo Galilei, the most important of the founding fathers of modern physics, insofar as the central procedure of his work involved the application of mathematical deduction to the results of measurement.

It is nowadays taken for granted by scientists that every measurement is subject to error so that repetitions of apparently the same experiment give different results. In the intellectual climate of Galileo's time, however, when logical syllogisms that admitted no gray area between right and wrong were the accepted means of deducing conclusions, his novel procedures were far from compelling. In judging his work one must remember that the conventions now accepted in reporting scientific results were adopted long after Galileo's time. Thus if, as is said, he stated as a fact that two objects dropped from the leaning tower of Pisa reached the ground together with not so much as a hand's breadth between them, it need not be inferred that he performed the experiment himself or that, if he did, the result was quite so perfect. Some such experiment had indeed been performed a little earlier (1586) by the Flemish mathematician Simon Stevin, but Galileo idealized the result. A light ball and a heavy ball do not reach the ground together, nor is the difference between them always the same, for it is impossible to reproduce the ideal of dropping them exactly at the same instant. Nevertheless, Galileo was satisfied that it came closer to the truth to say that they fell together than that there was a significant difference between their rates. This idealization of imperfect experiments remains an essential scientific process, though nowadays it is considered proper to present (or at least have available for scrutiny) the primary observations, so that others may judge independently whether they are prepared to accept the author's conclusion as to what would have been observed in an ideally conducted experiment.

The principles may be illustrated by repeating, with the advantage of modern instruments, an experiment such as Galileo himself performed—namely, that of measuring the time taken by a ball to roll different distances down a gently inclined channel. The following account is of a real experiment designed to show in a very simple example how the process of idealization proceeds, and how the preliminary conclusions may then be subjected to more searching test.

Lines equally spaced at six centimetres (2.4 inches) were scribed on a brass channel, and the ball was held at rest beside the highest line by means of a card. An electronic timer was started at the instant the card was removed, and the timer was stopped as the ball passed one of the other lines. Seven repetitions of each timing showed that the measurements typically spread over a range of $\frac{1}{20}$ of a second, presumably because of human limitations. In such a case, where a measurement is subject to random error, the average of many repetitions gives an improved estimate of what the result would be if the source of random error were eliminated; the factor by which the estimate is improved is roughly the square root of the number of measurements. Moreover, the theory of errors attributable to the German mathematician Carl Friedrich Gauss allows one to make a quantitative estimate of the reliability of the result, as expressed in the table by the conventional symbol $\pm$. This does not mean that the first result in column 2 is guaranteed to lie between 0.671 and 0.685 but that, if this determination of the average of seven measurements were to be repeated many times, about two-thirds of the determinations would lie within these limits.

The Galileo Experiment

distance, x (in centimetres)	time, t (in seconds)	$(t+0.09)/\sqrt{x}$
6	0.678 ± 0.007	0.314
12	0.997 ± 0.006	0.314
18	1.240 ± 0.007	0.313
24	1.444 ± 0.011	0.313
30	1.627 ± 0.010	0.313

Representing measurements by means of graphs

The representation of measurements by a graph, as in Figure 1, was not available to Galileo but was developed shortly after his time as a consequence of the work of the French mathematician-philosopher René Descartes. The points appear to lie close to a parabola, and the curve that is drawn is defined by the equation $x = 12t^2$. The fit is not quite perfect, and it is worth trying to find a better formula. Since the operations of starting the timer when the card is removed to allow the ball to roll and stopping it as the ball passes a mark are different, there is a possibility that, in addition to random timing errors, a systematic error appears in each measured value of t; that is to say, each measurement t is perhaps to be interpreted as $t + t_0$, where t_0 is an as-yet-unknown constant timing error. If this is so, one might look to see whether the measured times were related to distance not by $x = at^2$, where a is a constant, but by $x = a(t + t_0)^2$. This may also be tested graphically by first rewriting the equation as $\sqrt{x} = \sqrt{a}(t + t_0)$, which states that when the values of $\sqrt{x}$ are plotted against measured values of t they should lie on a straight line. Figure 2 verifies this prediction rather closely; the line does not pass through the origin but rather cuts the horizontal axis at -0.09 second. From this, one deduces that $t_0 = 0.09$ second and that $(t + 0.09)x$ should be the same for all the pairs of measurements given in the table. The third column shows that this is certainly the case. Indeed, the constancy is better than might have been expected in view of the estimated errors. This must be regarded as a statistical accident; it does not imply any greater assurance in the correctness of the formula than if the figures in the last column had ranged, as they might very well have done, between 0.311 and 0.315. One would be surprised if a repetition of the whole experiment again yielded so nearly constant a result.

A possible conclusion, then, is that for some reason—probably observational bias—the measured times underestimate by 0.09 second the real time t it takes a ball, starting from rest, to travel a distance x. If so, under ideal conditions x would be strictly proportional to t^2. Further experiments, in which the channel is set at different but still gentle slopes, suggest that the general rule takes the form $x = at^2$, with a proportional to the slope. This tentative idealization of the experimental measurements may

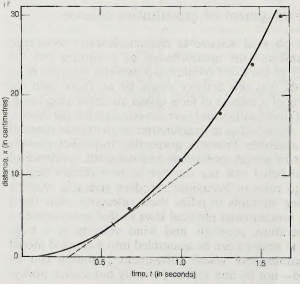


Figure 1: Data in the table of the Galileo experiment. The tangent to the curve is drawn at $t = 0.6$.

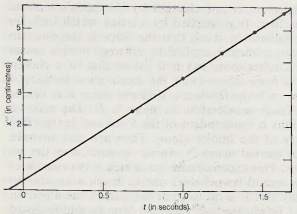


Figure 2: The data in the table of the Galileo experiment plotted differently.

need to be modified, or even discarded, in the light of further experiments. Now that it has been cast into mathematical form, however, it can be analyzed mathematically to reveal what consequences it implies. Also, this will suggest ways of testing it more searchingly.

From a graph such as Figure 1, which shows how x depends on t, one may deduce the instantaneous speed of the ball at any instant. This is the slope of the tangent drawn to the curve at the chosen value of t; at $t = 0.6$ second, for example, the tangent as drawn describes how x would be related to t for a ball moving at a constant speed of about 14 centimetres per second. The lower slope before this instant and the higher slope afterward indicate that the ball is steadily accelerating. One could draw tangents at various values of t and come to the conclusion that the instantaneous speed was roughly proportional to the time that had elapsed since the ball began to roll. This procedure, with its inevitable inaccuracies, is rendered unnecessary by applying elementary calculus to the supposed formula. The instantaneous speed v is the derivative of x with respect to t; if

$$x = at^2,\ v = \frac{dx}{dt} = 2at.$$

The implication that the velocity is strictly proportional to elapsed time is that a graph of v against t would be a straight line through the origin. On any graph of these quantities, whether straight or not, the slope of the tangent at any point shows how velocity is changing with time at that instant; this is the instantaneous acceleration f. For a straight-line graph of v against t, the slope and therefore the acceleration are the same at all times. Expressed mathematically, $f = dv/dt = d^2x/dt^2$; in the present case, f takes the constant value $2a$.

The preliminary conclusion, then, is that a ball rolling down a straight slope experiences constant acceleration and that the magnitude of the acceleration is proportional to the slope. It is now possible to test the validity of

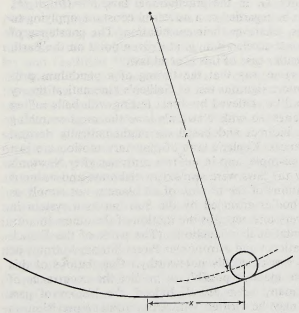


Figure 3: A ball rolling in a curved channel (see text).

the conclusion by finding what it predicts for a different experimental arrangement. If possible, an experiment is set up that allows more accurate measurements than those leading to the preliminary inference. Such a test is provided by a ball rolling in a curved channel so that its centre traces out a circular arc of radius r, as in Figure 3. Provided the arc is shallow, the slope at a distance x from its lowest point is very close to x/r, so that acceleration of the ball toward the lowest point is proportional to x/r. Introducing c to represent the constant of proportionality, this is written as a differential equation

$$\frac{d^2x}{dt^2} = -\frac{cx}{r}.$$

Here it is stated that, on a graph showing how x varies with t, the curvature d^2x/dt^2 is proportional to x and has the opposite sign, as illustrated in Figure 4. As the graph crosses the axis, x and therefore the curvature are zero, and the line is locally straight. This graph represents the oscillations of the ball between extremes of $\pm A$ after it has been released from $x = A$ at $t = 0$. The solution of the differential equation of which the diagram is the graphic representation is

$$x = A\cos(\omega t),$$

where ω, called the angular frequency, is written for $\sqrt{(c/r)}$. The ball takes time $T = 2\pi/\omega = 2\pi\sqrt{(r/c)}$ to return to its original position of rest, after which the oscillation is repeated indefinitely or until friction brings the ball to rest.

According to this analysis, the period, T, is independent of the amplitude of the oscillation, and this rather unexpected prediction is one that may be stringently tested. Instead of letting the ball roll on a curved channel, the same path is more easily and exactly realized by making it the bob of a simple pendulum. To test that the period is independent of amplitude two pendulums may be made as nearly identical as possible, so that they keep in step when swinging with the same amplitude. They are then swung with different amplitudes. It requires considerable care to detect any difference in period unless one amplitude is large, when the period is slightly longer. An observation that very nearly agrees with prediction, but not quite, does not necessarily show the initial supposition to be mistaken. In this case, the differential equation that predicted exact constancy of period was itself an approximation. When it is reformulated with the true expression for the slope replacing x/r, the solution (which involves quite heavy mathematics) shows a variation of period with amplitude that has been rigorously verified. Far from being discredited, the tentative assumption has emerged with enhanced support.

Galileo's law of acceleration, the physical basis of the expression $2\pi\sqrt{(c/r)}$ for the period, is further strengthened by finding that T varies inversely as the square root of r—*i.e.*, the length of the pendulum.

In addition, such measurements allow the value of the constant c to be determined with a high degree of precision, and it is found to coincide with the acceleration g of a freely falling body. In fact, the formula for the period of small oscillations of a simple pendulum of length r, $T = 2\pi\sqrt{(r/g)}$, is at the heart of some of the most precise methods for measuring g. This would not have happened unless the scientific community had accepted Galileo's description of the ideal behaviour and did not expect to be shaken in its belief by small deviations, so long as they could be understood as reflecting inevitable random discrepancies between the ideal and its experimental realization. The development of quantum mechanics in the first quarter of the 20th century was stimulated by the reluctant acceptance that this description systematically failed when applied to objects of atomic size. In this case, it was not a question, as with the variations of period, of translating the physical ideas into mathematics more precisely; the whole physical basis needed radical revision. Yet, the earlier ideas were not thrown out—they had been found to work well in far too many applications to be discarded. What emerged was a clearer understanding of the circumstances in which their absolute validity could safely be assumed.

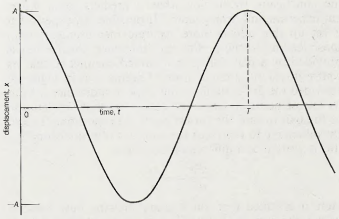


Figure 4: Oscillation of a simple pendulum (see text).

The experiments just described in detail as examples of scientific method were successful in that they agreed with expectation. They would have been just as successful if, in spite of being well conducted, they had disagreed because they would have revealed an error in the primary assumptions. The philosopher Karl Popper's widely accepted criterion for a scientific theory is that it must not simply pass such experimental tests as may be applied but that it must be formulated in such a way that falsification is in principle possible. For all its value as a test of scientific pretensions, however, it must not be supposed that the experimenter normally proceeds with Popper's criterion in mind. Normally he hopes to convince himself that his initial conception is correct. If a succession of tests agrees with (or fails to falsify) a hypothesis, it is regarded as reasonable to treat the hypothesis as true, at all events until it is discredited by a subsequent test. The scientist is not concerned with providing a guarantee of his conclusion, since, however many tests support it, there remains the possibility that the next one will not. His concern is to convince himself and his critical colleagues that a hypothesis has passed enough tests to make it worth accepting until a better one presents itself.

Popper's criterion for a scientific theory

The Newtonian paradigm

Up to this point the investigation has been concerned exclusively with kinetics—that is to say, providing an accurate mathematical description of motion, in this case of a ball on an inclined plane, with no implied explanation of the physical processes responsible. Newton's general dynamic theory, as expounded in his *Philosophiae Naturalis Principia Mathematica* of 1687, laid down in the form of his laws of motion, together with other axioms and postulates, the rules to follow in analyzing the motion of bodies interacting among themselves. This theory of classical mechanics is described in detail in the article MECHANICS, but some general comments may be offered here. For the present purpose, it seems sufficient to consider only bodies moving along a straight line and acted upon by forces parallel to the motion. Newton's laws are, in fact, considerably more general than this and encompass motion in curves as a result of forces deflecting a body from its initial direction.

Newton's laws of motion

Newton's first law may more properly be ascribed to Galileo. It states that a body continues at rest or in uniform motion along a straight line unless it is acted upon by a force, and it enables one to recognize when a force is acting. A tennis ball struck by a racket experiences a sudden change in its motion attributable to a force exerted by the racket. The player feels the shock of the impact. According to Newton's third law (action and reaction are equal and opposite), the force that the ball exerts on the racket is equal and opposite to that which the racket exerts on the ball. Moreover, a second balanced action and reaction acts between player and racket.

Newton's second law quantifies the concept of force, as well as that of inertia. A body acted upon by a steady force suffers constant acceleration. Thus a freely falling body or a ball rolling down a plane has constant acceleration, as has been seen, and this is to be interpreted in Newton's terms as evidence that the force of gravity, which causes the acceleration, is not changed by the body's motion. The same force (*e.g.*, applied by a string which includes a spring balance to check that the force is the same in different experiments) applied to different bodies causes different accelerations; and it is found that, if a chosen strength of force causes twice the acceleration in body A as it does in body B, then a different force also causes twice as much acceleration in A as in B. The ratio of accelerations is independent of the force and is therefore a property of the bodies alone. They are said to have inertia (or inertial mass) in inverse proportion to the accelerations. This experimental fact, which is the essence of Newton's second law, enables one to assign a number to every body that is a measure of its mass. Thus a certain body may be chosen as a standard of mass and assigned the number 1. Another body is said to have mass m if the body shows only a fraction $1/m$ of the acceleration of this standard when the two are subjected to the same force. By proceeding in this way, every body may be assigned a mass. It is because experiment allows this definition to be made that a given force causes every body to show acceleration f such that mf is the same for all bodies. This means that the product mf is determined only by the force and not by the particular body on which it acts, and mf is defined to be the numerical measure of the force. In this way a consistent set of measures of force and mass is arrived at, having the property that $F = mf$. In this equation F, m, and f are to be interpreted as numbers measuring the strength of the force, the magnitude of the mass, and the rate of acceleration; and the product of the numbers m and f is always equal to the number F. The product mv, called *motus* (motion) by Newton, is now termed momentum. Newton's second law states that the rate of change of momentum equals the strength of the applied force.

In order to assign a numerical measure m to the mass of a body, a standard of mass must be chosen and assigned the value $m = 1$. Similarly, to measure displacement a unit of length is needed, and for velocity and acceleration a unit of time also must be defined. Given these, the numerical measure of a force follows from mf without need to define a unit of force. Thus in the Système Internationale d'Unités (SI), in which the units are the standard kilogram, the standard metre, and the standard second, a force of magnitude unity is one that, applied to a mass of one kilogram, causes its velocity to increase steadily by one metre per second during every second the force is acting.

The idealized observation of Galileo that all bodies in free-fall accelerate equally implies that the gravitational force causing acceleration bears a constant relation to the inertial mass. According to Newton's postulated law of gravitation, two bodies of mass m_1 and m_2, separated by a distance r, exert equal attractive forces on each other (the equal action and reaction of the third law of motion) of magnitude proportional to m_1m_2/r^2. The constant of proportionality, G, in the gravitational law, $F = Gm_1m_2/r^2$, is thus to be regarded as a universal constant, applying to all bodies, whatever their constitution. The constancy of gravitational acceleration, g, at a given point on the Earth is a particular case of this general law.

Newton's law of gravitation

In the same way that the timing of a pendulum provided a more rigorous test of Galileo's kinematical theory than could be achieved by direct testing with balls rolling down planes, so with Newton's laws the most searching tests are indirect and based on mathematically derived consequences. Kepler's laws of planetary motion are just such an example, and in the two centuries after Newton's *Principia* the laws were applied to elaborate and arduous computations of the motion of all planets, not simply as isolated bodies attracted by the Sun but as a system in which every one perturbs the motion of the others by mutual gravitational interactions. (The work of the French mathematician and astronomer Pierre-Simon, Marquis de Laplace, was especially noteworthy.) Calculations of this kind have made it possible to predict the occurrence of eclipses many years ahead. Indeed, the history of past eclipses may be written with extraordinary precision so that, for instance, Thucydides' account of the lunar eclipse

that fatally delayed the Athenian expedition against Syracuse in 413 BC matches the calculations perfectly (see ECLIPSE, OCCULTATION, AND TRANSIT). Similarly, unexplained small departures from theoretical expectation of the motion of Uranus led John Couch Adams of England and Urbain-Jean-Joseph Le Verrier of France to predict in 1845 that a new planet (Neptune) would be seen at a particular point in the heavens. The discovery of Pluto in 1930 was achieved in much the same way.

There is no obvious reason why the inertial mass m that governs the response of a body to an applied force should also determine the gravitational force between two bodies, as described above. Consequently, the period of a pendulum is independent of its material and governed only by its length and the local value of g; this has been verified with an accuracy of a few parts per million. Still more sensitive tests, as originally devised by the Hungarian physicist Roland, Baron von Eötvös (1890), and repeated several times since, have demonstrated clearly that the accelerations of different bodies in a given gravitational environment are identical within a few parts in 10^{12}. An astronaut in free orbit can remain poised motionless in the centre of the cabin of his spacecraft, surrounded by differently constituted objects, all equally motionless (except for their extremely weak mutual attractions) because all of them are identically affected by the gravitational field in which they are moving. He is unaware of the gravitational force, just as those on the Earth are unaware of the Sun's attraction, moving as they do with the Earth in free orbit around the Sun. Albert Einstein made this experimental finding a central feature of his general theory of relativity (see RELATIVITY).

Newton believed that everything moved in relation to a fixed but undetectable spatial frame so that it could be said to have an absolute velocity. Time also flowed at the same steady pace everywhere. Even if there were no matter in the universe, the frame of the universe would still exist, and time would still flow even though there was no one to observe its passage. In Newton's view, when matter is present it is unaffected by its motion through space. If the length of a moving metre stick were compared with the length of one at rest, they would be found to be the same. Clocks keep universal time whether they are moving or not; therefore, two identical clocks, initially synchronized, would still be synchronized after one had been carried into space and brought back. The laws of motion take such a form that they are not changed by uniform motion. They were devised to describe accurately the response of bodies to forces whether in the heavens or on the Earth, and they lose no validity as a result of the Earth's motion at 30 kilometres per second in its orbit around the Sun. This motion, in fact, would not be discernible by an observer in a closed box. The supposed invariance of the laws of motion, in addition to standards of measurement, to uniform translation was called "Galilean invariance" by Einstein.

Galilean invariance

The impossibility of discerning absolute velocity led in Newton's time to critical doubts concerning the necessity of postulating an absolute frame of space and universal time, and the doubts of the philosophers George Berkeley and Gottfried Wilhelm Leibniz, among others, were still more forcibly presented in the severe analysis of the foundations of classical mechanics by the Austrian physicist Ernst Mach in 1883. James Clerk Maxwell's theory of electromagnetic phenomena (1865), including his description of light as electromagnetic waves, brought the problem to a state of crisis. It became clear that if light waves were propagated in the hypothetical ether that filled all space and provided an embodiment of Newton's absolute frame (see below), it would not be logically consistent to accept both Maxwell's theory and the ideas expressed in Galilean invariance, for the speed of light as it passed an observer would reveal how rapidly he was traveling through the ether.

Ingenious attempts by the physicists George FitzGerald of Ireland and Hendrik A. Lorentz of The Netherlands to devise a compromise to salvage the notion of ether were eventually superseded by Einstein's special theory of relativity (see RELATIVITY). Einstein proposed in 1905 that all laws of physics, not solely those of mechanics, must take the same form for observers moving uniformly relative to one another, however rapidly. In particular, if two observers, using identical metre sticks and clocks, set out to measure the speed of a light signal as it passes them, both would obtain the same value no matter what their relative velocity might be; in a Newtonian world, of course, the measured values would differ by the relative velocity of the two observers. This is but one example of the counterintuitive character of relativistic physics, but the deduced consequences of Einstein's postulate have been so frequently and so accurately verified by experiment that it has been incorporated as a fundamental axiom in physical theory.

Special theory of relativity

With the abandonment of the ether hypothesis, there has been a reversion to a philosophical standpoint reluctantly espoused by Newton. To him and to his contemporaries the idea that two bodies could exert gravitational forces on each other across immense distances of empty space was abhorrent. However, attempts to develop Descartes's notion of a space-filling fluid ether as a transmitting medium for forces invariably failed to account for the inverse square law. Newton himself adopted a pragmatic approach, deducing the consequences of his laws and showing how well they agreed with observation; he was by no means satisfied that a mechanical explanation was impossible, but he confessed in the celebrated remark, "Hypotheses non fingo" (Latin: "I frame no hypotheses"), that he had no solution to offer.

A similar reversion to the safety of mathematical description is represented by the rejection, during the early 1900s, of the explanatory ether models of the 19th century and their replacement by model-free analysis in terms of relativity theory. This certainly does not imply giving up the use of models as imaginative aids in extending theories, predicting new effects, or devising interesting experiments; if nothing better is available, however, a mathematical formulation that yields verifiably correct results is to be preferred over an intuitively acceptable model that does not.

Interplay of experiment and theory

The foregoing discussion should have made clear that progress in physics, as in the other sciences, arises from a close interplay of experiment and theory. In a well-established field like classical mechanics, it may appear that experiment is almost unnecessary and all that is needed is the mathematical or computational skill to discover the solutions of the equations of motion. This view, however, overlooks the role of observation or experiment in setting up the problem in the first place. To discover the conditions under which a bicycle is stable in an upright position or can be made to turn a corner, it is first necessary to invent and observe a bicycle. The equations of motion are so general and serve as the basis for describing so extended a range of phenomena that the mathematician must usually look at the behaviour of real objects in order to select those that are both interesting and soluble. His analysis may indeed suggest the existence of interesting related effects that can be examined in the laboratory; thus, the invention or discovery of new things may be initiated by the experimenter or the theoretician. To employ terms such as this has led, especially in the 20th century, to a common assumption that experimentation and theorizing are distinct activities, rarely performed by the same person. It is true that almost all active physicists pursue their vocation primarily in one mode or the other. Nevertheless, the innovative experimenter can hardly make progress without an informed appreciation of the theoretical structure, even if he is not technically competent to find the solution of particular mathematical problems. By the same token, the innovative theorist must be deeply imbued with the way real objects behave, even if he is not technically competent to put together the apparatus to examine the problem. The fundamental unity of physical science should be borne in mind during the following outline of characteristic examples of experimental and theoretical physics.

CHARACTERISTIC EXPERIMENTAL PROCEDURES

Unexpected observation. The discovery of X rays (1895) by Wilhelm Conrad Röntgen of Germany was certainly serendipitous. It began with his noticing that when an electric current was passed through a discharge tube a nearby fluorescent screen lit up, even though the tube was completely wrapped in black paper.

Ernest Marsden, a student engaged on a project, reported to his professor, Ernest Rutherford (then at the University of Manchester in England), that alpha particles from a radioactive source were occasionally deflected more than 90° when they hit a thin metal foil. Astonished at this observation, Rutherford deliberated on the experimental data to formulate his nuclear model of the atom (1911).

Heike Kamerlingh Onnes of The Netherlands, the first to liquefy helium, cooled a thread of mercury to within 4 K of absolute zero (4 K equals −269° C) to test his belief that electrical resistance would tend to vanish at zero. This was what the first experiment seemed to verify, but a more careful repetition showed that instead of falling gradually, as he expected, all trace of resistance disappeared abruptly just above 4 K. This phenomenon of superconductivity, which Kamerlingh Onnes discovered in 1911, defied theoretical explanation until 1957.

The not-so-unexpected chance. From 1807 the Danish physicist and chemist Hans Christian Ørsted came to believe that electrical phenomena could influence magnets, but it was not until 1819 that he turned his investigations to the effects produced by an electric current. On the basis of his tentative models he tried on several occasions to see if a current in a wire caused a magnet needle to turn when it was placed transverse to the wire, but without success. Only when it occurred to him, without forethought, to arrange the needle parallel on the wire did the long-sought effect appear.

A second example of this type of experimental situation involves the discovery of electromagnetic induction by the English physicist and chemist Michael Faraday. Aware that an electrically charged body induces a charge in a nearby body, Faraday sought to determine whether a steady current in a coil of wire would induce such a current in another short-circuited coil close to it. He found no effect except in instances where the current in the first coil was switched on or off, at which time a momentary current appeared in the other. He was in effect led to the concept of electromagnetic induction by changing magnetic fields.

Qualitative tests to distinguish alternative theories. At the time that Augustin-Jean Fresnel presented his wave theory of light to the French Academy (1815), the leading physicists were adherents of Newton's corpuscular theory. It was pointed out by Siméon-Denis Poisson, as a fatal objection, that Fresnel's theory predicted a bright spot at the very centre of the shadow cast by a circular obstacle. When this was in fact observed by François Arago, Fresnel's theory was immediately accepted.

Another qualitative difference between the wave and corpuscular theories concerned the speed of light in a transparent medium. To explain the bending of light rays toward the normal to the surface when light entered the medium, the corpuscular theory demanded that light go faster while the wave theory required that it go slower. Jean-Bernard-Léon Foucault showed that the latter was correct (1850).

The three categories of experiments or observations discussed above are those that do not demand high-precision measurement. The following, however, are categories in which measurement at varying degrees of precision is involved.

Direct comparison of theory and experiment. This is one of the commonest experimental situations. Typically, a theoretical model makes certain specific predictions, perhaps novel in character, perhaps novel only in differing from the predictions of competing theories. There is no fixed standard by which the precision of measurement may be judged adequate. As is usual in science, the essential question is whether the conclusion carries conviction, and this is conditioned by the strength of opinion regarding alternative conclusions.

Where strong prejudice obtains, opponents of a heterodox conclusion may delay acceptance indefinitely by insisting on a degree of scrupulosity in experimental procedure that they would unhesitatingly dispense with in other circumstances. For example, few experiments in paranormal phenomena, such as clairvoyance, which have given positive results under apparently stringent conditions, have made converts among scientists. In the strictly physical domain, the search for ether drift provides an interesting study. At the height of acceptance of the hypothesis that light waves are carried by a pervasive ether, the question of whether the motion of the Earth through space dragged the ether with it was tested (1887) by A.A. Michelson and Edward W. Morley of the United States by looking for variations in the velocity of light as it traveled in different directions in the laboratory. Their conclusion was that there was a small variation, considerably less than the Earth's velocity in its orbit around the Sun, and that the ether was therefore substantially entrained in the Earth's motion. According to Einstein's relativity theory (1905), no variation should have been observed, but during the next 20 years another American investigator, Dayton C. Miller, repeated the experiment many times in different situations and concluded that, at least on a mountaintop, there was a real "ether wind" of about 10 kilometres per second. Although Miller's final presentation was a model of clear exposition, with evidence scrupulously displayed and discussed, it has been set aside and virtually forgotten. This is partly because other experiments failed to show the effect; however, their conditions were not strictly comparable, since few, if any, were conducted on mountaintops. More significantly, other tests of relativity theory supported it in so many different ways as to lead to the consensus that one discrepant set of observations cannot be allowed to weigh against the theory.

Subjectivity

At the opposite extreme may be cited the 1919 expedition of the English scientist-mathematician Arthur Stanley Eddington to measure the very small deflection of the light from a star as it passed close to the Sun—a measurement that requires a total eclipse. The theories involved here were Einstein's general theory of relativity and the Newtonian particle theory of light, which predicted only half the relativistic effect. The conclusion of this exceedingly difficult measurement—that Einstein's theory was followed within the experimental limits of error, which amounted to ±30 percent—was the signal for worldwide feting of Einstein. If his theory had not appealed aesthetically to those able to appreciate it and if there had been any passionate adherents to the Newtonian view, the scope for error could well have been made the excuse for a long drawn-out struggle, especially since several repetitions at subsequent eclipses did little to improve the accuracy. In this case, then, the desire to believe was easily satisfied. It is gratifying to note that recent advances in radio astronomy have allowed much greater accuracy to be achieved, and Einstein's prediction is now verified within about 1 percent.

During the decade after his expedition, Eddington developed an extremely abstruse fundamental theory that led him to assert that the quantity $hc/2\pi e^2$ (h is Planck's constant, c the velocity of light, and e the charge on the electron) must take the value 137 exactly. At the time, uncertainties in the values of h and e allowed its measured value to be given as 137.29 ± 0.11; in accordance with the theory of errors, this implies that there was estimated to be about a 1 percent chance that a perfectly precise measurement would give 137. In the light of Eddington's great authority there were many prepared to accede to his belief. Since then the measured value of this quantity has come much closer to Eddington's prediction and is given as 137.03604 ± 0.00011. The discrepancy, though small, is 330 times the estimated error, compared with 2.6 times for the earlier measurement, and therefore a much more weighty indication against Eddington's theory. As the intervening years have cast no light on the virtual impenetrability of his argument, there is now hardly a physicist who takes it seriously.

Compilation of data. Technical design, whether of laboratory instruments or for industry and commerce, depends on knowledge of the properties of materials (density,

strength, electrical conductivity, etc.), some of which can only be found by very elaborate experiments (*e.g.*, those dealing with the masses and excited states of atomic nuclei). One of the important functions of standards laboratories is to improve and extend the vast body of factual information, but much also arises incidentally rather than as the prime objective of an investigation or may be accumulated in the hope of discovering regularities or to test the theory of a phenomenon against a variety of occurrences.

When chemical compounds are heated in a flame, the resulting colour can be used to diagnose the presence of sodium (orange), copper (green-blue), and many other elements. This procedure has long been used. Spectroscopic examination shows that every element has its characteristic set of spectral lines, and the discovery by the Swiss mathematician Johann Jakob Balmer of a simple arithmetic formula relating the wavelengths of lines in the hydrogen spectrum (1885) proved to be the start of intense activity in precise wavelength measurements of all known elements and the search for general principles. With the Danish physicist Niels Bohr's quantum theory of the hydrogen atom (1913) began an understanding of the basis of Balmer's formula; thenceforward spectroscopic evidence underpinned successive developments toward what is now a successful theory of atomic structure.

Direct and indirect testing

Tests of fundamental concepts. Coulomb's law states that the force between two electric charges varies as the inverse square of their separation. Direct tests, such as those performed with a special torsion balance by the French physicist Charles-Augustin de Coulomb, for whom the law is named, can be at best approximate. A very sensitive indirect test, devised by the English scientist and clergyman Joseph Priestley (following an observation by Benjamin Franklin) but first realized by the English physicist and chemist Henry Cavendish (1771), relies on the mathematical demonstration that no electrical changes occurring outside a closed metal shell—as, for example, by connecting it to a high voltage source—produce any effect inside if the inverse square law holds. Since modern amplifiers can detect minute voltage changes, this test can be made very sensitive. It is typical of the class of null measurements in which only the theoretically expected behaviour leads to no response and any hypothetical departure from theory gives rise to a response of calculated magnitude. It has been shown in this way that if the force between charges, r apart, is proportional not to $1/r^2$ but to $1/r^{2+x}$, then x is less than 2×10^{-9}.

According to the relativistic theory of the hydrogen atom proposed by the English physicist P.A.M. Dirac (1928), there should be two different excited states exactly coinciding in energy. Measurements of spectral lines resulting from transitions in which these states were involved hinted at minute discrepancies, however. Some years later (*c.* 1950) Willis E. Lamb, Jr., and Robert C. Retherford of the United States, employing the novel microwave techniques that wartime radar contributed to peacetime research, were able not only to detect the energy difference between the two levels directly but to measure it rather precisely as well. The difference in energy, compared to the energy above the ground state, amounts to only four parts in 10,000,000, but this was one of the crucial pieces of evidence that led to the development of quantum electrodynamics, a central feature of the modern theory of fundamental particles (see SUBATOMIC PARTICLES: *Quantum electrodynamics*).

CHARACTERISTIC THEORETICAL PROCEDURES

Application of established principles to new problems

Only at rare intervals in the development of a subject, and then only with the involvement of a few, are theoretical physicists engaged in introducing radically new concepts. The normal practice is to apply established principles to new problems so as to extend the range of phenomena that can be understood in some detail in terms of accepted fundamental ideas. Even when, as with the quantum mechanics of Werner Heisenberg (formulated in terms of matrices; 1925) and of Erwin Schrödinger (developed on the basis of wave functions; 1926), a major revolution is initiated, most of the accompanying theoretical activity involves investigating the consequences of the new hypothesis as if it were fully established in order to discover critical tests against experimental facts. There is little to be gained by attempting to classify the process of revolutionary thought because every case history throws up a different pattern. What follows is a description of typical procedures as normally used in theoretical physics. As in the preceding section, it will be taken for granted that the essential preliminary of coming to grips with the nature of the problem in general descriptive terms has been accomplished, so that the stage is set for systematic, usually mathematical, analysis.

Direct solution of fundamental equations. Insofar as the Sun and planets, with their attendant satellites, can be treated as concentrated masses moving under their mutual gravitational influences, they form a system that has not so overwhelmingly many separate units as to rule out step-by-step calculation of the motion of each. Modern high-speed computers are admirably adapted to this task and are used in this way to plan space missions and to decide on fine adjustments during flight. Most physical systems of interest, however, are either composed of too many units or are governed not by the rules of classical mechanics but rather by quantum mechanics, which is much less suited for direct computation.

Dissection. The mechanical behaviour of a body is analyzed in terms of Newton's laws of motion by imagining it dissected into a number of parts, each of which is directly amenable to the application of the laws or has been separately analyzed by further dissection so that the rules governing its overall behaviour are known. A very simple illustration of the method is given by the arrangement in Figure 5A, where two masses are joined by a light string passing over a pulley. The heavier mass, m_1, falls with constant acceleration, but what is the magnitude of the acceleration? If the string were cut, each mass would experience the force, m_1g or m_2g, due to its gravitational attraction and would fall with acceleration g. The fact that the string prevents this is taken into account by assuming that it is in tension and also acts on each mass. When the string is cut just above m_2, the state of accelerated motion just before the cut can be restored by applying equal and opposite forces (in accordance with Newton's third law) to the cut ends, as in Figure 5B; the string above the cut pulls the string below upward with a force T, while the string below pulls that above downward to the same extent. As yet, the value of T is not known. Now if the string is light, the tension T is sensibly the same everywhere along it, as may be seen by imagining a second cut, higher up,

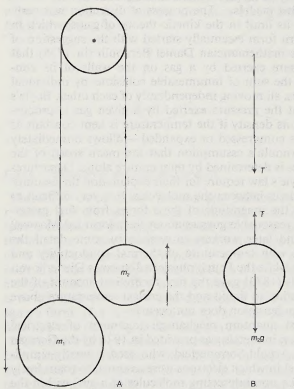


Figure 5: Dissection of a complex system into elementary parts (see text).

to leave a length of string acted upon by T at the bottom and possibly a different force T' at the second cut. The total force $T - T'$ on the string must be very small if the cut piece is not to accelerate violently, and, if the mass of the string is neglected altogether, T and T' must be equal. This does not apply to the tension on the two sides of the pulley, for some resultant force will be needed to give it the correct accelerative motion as the masses move. This is a case for separate examination, by further dissection, of the forces needed to cause rotational acceleration. To simplify the problem one can assume the pulley to be so light that the difference in tension on the two sides is negligible. Then the problem has been reduced to two elementary parts—on the right the upward force on m_2 is $T - m_2 g$, so that its acceleration upward is $T/m_2 - g$; and on the left the downward force on m_1 is $m_1 g - T$, so that its acceleration downward is $g - T/m_1$. If the string cannot be extended, these two accelerations must be identical, from which it follows that $T = 2m_1 m_2 g/(m_1 + m_2)$ and the acceleration of each mass is $g(m_1 - m_2)/(m_1 + m_2)$. Thus if one mass is twice the other ($m_1 = 2m_2$), its acceleration downward is $g/3$.

A liquid may be imagined divided into small volume elements, each of which moves in response to gravity and the forces imposed by its neighbours (pressure and viscous drag). The forces are constrained by the requirement that the elements remain in contact, even though their shapes and relative positions may change with the flow. From such considerations are derived the differential equations that describe fluid motion (see MECHANICS: *Fluid mechanics*).

The dissection of a system into many simple units in order to describe the behaviour of a complex structure in terms of the laws governing the elementary components is sometimes referred to, often with a pejorative implication, as reductionism. Insofar as it may encourage concentration on those properties of the structure that can be explained as the sum of elementary processes to the detriment of properties that arise only from the operation of the complete structure, the criticism must be considered seriously. The physical scientist is, however, well aware of the existence of the problem (see below *Simplicity and complexity*). If he is usually unrepentant about his reductionist stance, it is because this analytical procedure is the only systematic procedure he knows, and it is one that has yielded virtually the whole harvest of scientific inquiry. What is set up as a contrast to reductionism by its critics is commonly called the holistic approach, whose title confers a semblance of high-mindedness while hiding the poverty of tangible results it has produced.

Reductionism

Simplified models. The process of dissection was early taken to its limit in the kinetic theory of gases, which in its modern form essentially started with the suggestion of the Swiss mathematician Daniel Bernoulli (in 1738) that the pressure exerted by a gas on the walls of its container is the sum of innumerable collisions by individual molecules, all moving independently of each other. Boyle's law—that the pressure exerted by a given gas is proportional to its density if the temperature is kept constant as the gas is compressed or expanded—follows immediately from Bernoulli's assumption that the mean speed of the molecules is determined by temperature alone. Departures from Boyle's law require for their explanation the assumption of forces between the molecules. It is very difficult to calculate the magnitude of these forces from first principles, but reasonable guesses about their form led Maxwell (1860) and later workers to explain in some detail the variation with temperature of thermal conductivity and viscosity, while the Dutch physicist Johannes Diederik van der Waals (1873) gave the first theoretical account of the condensation to liquid and the critical temperature above which condensation does not occur.

The first quantum mechanical treatment of electrical conduction in metals was provided in 1928 by the German physicist Arnold Sommerfeld, who used a greatly simplified model in which electrons were assumed to roam freely (much like non-interacting molecules of a gas) within the metal as if it were a hollow container. The most remarkable simplification, justified at the time by its success rather than by any physical argument, was that the electrical force between electrons could be neglected. Since then, justification—without which the theory would have been impossibly complicated—has been provided in the sense that means have been devised to take account of the interactions whose effect is indeed considerably weaker than might have been supposed. In addition, the influence of the lattice of atoms on electronic motion has been worked out for many different metals. This development involved experimenters and theoreticians working in harness; the results of specially revealing experiments served to check the validity of approximations without which the calculations would have required excessive computing time.

These examples serve to show how real problems almost always demand the invention of models in which, it is hoped, the most important features are correctly incorporated while less-essential features are initially ignored and allowed for later if experiment shows their influence not to be negligible. In almost all branches of mathematical physics there are systematic procedures—namely, perturbation techniques—for adjusting approximately correct models so that they represent the real situation more closely.

Recourse to perturbation techniques

Recasting of basic theory. Newton's laws of motion and of gravitation and Coulomb's law for the forces between charged particles lead to the idea of energy as a quantity that is conserved in a wide range of phenomena (see below *Conservation laws and extremal principles*). It is frequently more convenient to use conservation of energy and other quantities than to start an analysis from the primitive laws. Other procedures are based on showing that, of all conceivable outcomes, the one followed is that for which a particular quantity takes a maximum or a minimum value—*e.g.*, entropy change in thermodynamic processes, action in mechanical processes, and optical path length for light rays.

GENERAL OBSERVATIONS

The foregoing accounts of characteristic experimental and theoretical procedures are necessarily far from exhaustive. In particular, they say too little about the technical background to the work of the physical scientist. The mathematical techniques used by the modern theoretical physicist are frequently borrowed from the pure mathematics of past eras. The work of Augustin-Louis Cauchy on functions of a complex variable, of Arthur Cayley and James Joseph Sylvester on matrix algebra, and of Bernhard Riemann on non-Euclidean geometry, to name but a few, were investigations undertaken with little or no thought for practical applications.

The experimental physicist, for his part, has benefited greatly from technological progress and from instrumental developments that were undertaken in full knowledge of their potential research application but were nevertheless the product of single-minded devotion to the perfecting of an instrument as a worthy thing-in-itself. The developments during World War II provide the first outstanding example of technology harnessed on a national scale to meet a national need. Postwar advances in nuclear physics and in electronic circuitry, applied to almost all branches of research, were founded on the incidental results of this unprecedented scientific enterprise. The semiconductor industry sprang from the successes of microwave radar and, in its turn, through the transistor, made possible the development of reliable computers with power undreamed of by the wartime pioneers of electronic computing. From all these, the research scientist has acquired the means to explore otherwise inaccessible problems. Of course, not all of the important tools of modern-day science were the by-products of wartime research. The electron microscope is a good case in point. Moreover, this instrument may be regarded as a typical example of the sophisticated equipment to be found in all physical laboratories, of a complexity that the research-oriented user frequently does not understand in detail, and whose design depended on skills he rarely possesses.

Impact of advances in technology and instrumentation

It should not be thought that the physicist does not give a just return for the tools he borrows. Engineering and technology are deeply indebted to pure science, while

much modern pure mathematics can be traced back to investigations originally undertaken to elucidate a scientific problem.

Concepts fundamental to the attitudes and methods of physical science

FIELDS

Vector notation

Newton's law of gravitation and Coulomb's electrostatic law both give the force between two particles as inversely proportional to the square of their separation and directed along the line joining them. The force acting on one particle is a vector. It can be represented by a line with arrowhead; the length of the line is made proportional to the strength of the force, and the direction of the arrow shows the direction of the force. If a number of particles are acting simultaneously on the one considered, the resultant force is found by vector addition; the vectors representing each separate force are joined head to tail, and the resultant is given by the line joining the first tail to the last head.

In what follows the electrostatic force will be taken as typical, and Coulomb's law is expressed in the form $F = q_1 q_2 r/4\pi\varepsilon_0 r^3$. The boldface characters F and r are vectors, F being the force which a point charge q_1 exerts on another point charge q_2. The combination r/r^3 is a vector in the direction of r, the line joining q_1 to q_2, with magnitude $1/r^2$ as required by the inverse square law. When r is rendered in lightface, it means simply the magnitude of the vector r, without direction. The combination $4\pi\varepsilon_0$ is a constant whose value is irrelevant to the present discussion. The combination $q_1 r/4\pi\varepsilon_0 r^3$ is called the electric field strength due to q_1 at a distance r from q_1 and is designated by E; it is clearly a vector parallel to r. At every point in space E takes a different value, determined by r, and the complete specification of $E(r)$—that is, the magnitude and direction of E at every point r—defines the electric field. If there are a number of different fixed charges, each produces its own electric field of inverse square character, and the resultant E at any point is the vector sum of the separate contributions. Thus the magnitude and direction of E may change in a complicated fashion from point to point. Any particle carrying charge q that is put in a place where the field is E experiences a force qE (provided the other charges are not displaced when it is inserted; if they are $E(r)$ must be recalculated for the actual positions of the charges).

A vector field, varying from point to point, is not always easily represented by a diagram, and it is often helpful for this purpose, as well as in mathematical analysis, to introduce the potential φ, from which E may be deduced. To appreciate its significance, the concept of vector gradient must be explained.

Gradient. The contours on a standard map are lines along which the height of the ground above sea level is constant. They usually take a complicated form, but if one imagines contours drawn at very close intervals of height and a small portion of the map to be greatly enlarged, the contours of this local region will become very nearly straight, like the two drawn in Figure 6 for heights h and $h + \delta h$.

Walking along any of these contours, one remains on the level. The slope of the ground is steepest along PQ, and, if the distance from P to Q is δl, the gradient is $\delta h/\delta l$ or dh/dl in the limit when δh and δl are allowed to go to zero. The vector gradient is a vector of this magnitude drawn parallel to PQ and is written as grad h, or ∇h. Walking along any other line PR at an angle θ to PQ, the slope is less in the ratio PQ/PR, or $\cos\theta$. The slope along PR is (grad h) $\cos\theta$ and is the component of the vector grad h along a line at an angle θ to the vector itself. This is an example of the general rule for finding components of vectors. In particular, the components parallel to the x and y directions have magnitude $\partial h/\partial x$ and $\partial h/\partial y$ (the partial derivatives, represented by the symbol ∂, mean, for instance, that $\partial h/\partial x$ is the rate at which h changes with distance in the x direction, if one moves so as to keep y constant; and $\partial h/\partial y$ is the rate of change in the y direction, x being constant). This result is expressed by

$$\text{grad } h = \left[\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y}\right],$$

the quantities in brackets being the components of the vector along the coordinate axes. Vector quantities that vary in three dimensions can similarly be represented by three Cartesian components, along x, y, and z axes; *e.g.*, $V = (V_x, V_y, V_z)$.

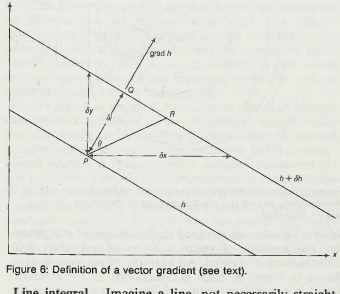


Figure 6: Definition of a vector gradient (see text).

Line integral. Imagine a line, not necessarily straight, drawn between two points A and B and marked off in innumerable small elements like δl in Figure 7, which is to be thought of as a vector. If a vector field takes a value V at this point, the quantity $V\delta l \cdot \cos\theta$ is called the scalar product of the two vectors V and δl and is written as $V \cdot \delta l$. The sum of all similar contributions from the different δl gives, in the limit when the elements are made infinitesimally small, the line integral $\int_A^B V \cdot dl$ along the line chosen.

Reverting to the contour map, it will be seen that $\int_A^B (\text{grad } h) \cdot dl$ is just the vertical height of B above A and that the value of the line integral is the same for all choices of line joining the two points. When a scalar quantity φ, having magnitude but not direction, is uniquely defined at every point in space, as h is on a two-dimensional map, the vector grad φ is then said to be irrotational, and $\varphi(r)$ is the potential function from which a vector field grad φ can be derived. Not all vector fields can be derived from a potential function, but the Coulomb and gravitational fields are of this form.

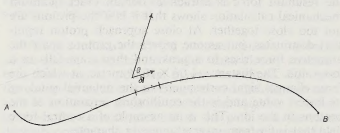


Figure 7: Definition of line integral (see text).

Potential. A potential function $\varphi(r)$ defined by $\varphi = A/r$, where A is a constant, takes a constant value on every sphere centred at the origin. The set of nesting spheres is the analogue in three dimensions of the contours of height on a map, and grad φ at a point r is a vector pointing normal to the sphere that passes through r; it therefore lies along the radius through r, and has magnitude $-A/r^2$. That is to say, grad $\varphi = -Ar/r^3$ and describes a field of inverse square form. If A is set equal to $q_1/4\pi\varepsilon_0$, the electrostatic field due to a charge q_1 at the origin is $E = -\text{grad } \varphi$.

When the field is produced by a number of point charges, each contributes to the potential $\varphi(r)$ in proportion to the size of the charge and inversely as the distance from the charge to the point r. To find the field strength E at r, the potential contributions can be added as numbers and contours of the resultant φ plotted; from these E

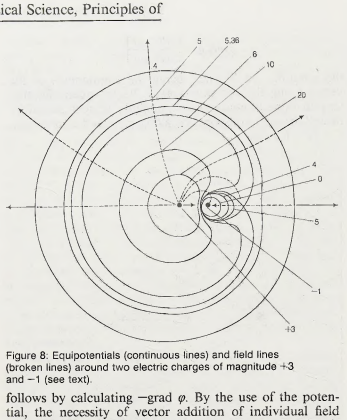


Figure 8. Equipotentials (continuous lines) and field lines (broken lines) around two electric charges of magnitude +3 and −1 (see text).

follows by calculating −grad φ. By the use of the potential, the necessity of vector addition of individual field contributions is avoided. An example of equipotentials is shown in Figure 8. Each is determined by the equation $3/r_1 - 1/r_2 =$ constant, with a different constant value for each, as shown. For any two charges of opposite sign, the equipotential surface, $\varphi = 0$, is a sphere, as no other is.

Conservative forces. The inverse square laws of gravitation and electrostatics are examples of central forces where the force exerted by one particle on another is along the line joining them and is also independent of direction. Whatever the variation of force with distance, a central force can always be represented by a potential; forces for which a potential can be found are called conservative. The work done by the force $F(r)$ on a particle as it moves along a line from A to B is the line integral $\int_A^B F \cdot dl$, or $-\int_A^B \text{grad}\ \varphi \cdot dl$ if F is derived from a potential φ, and this integral is just the difference between φ at A and B.

The ionized hydrogen molecule consists of two protons bound together by a single electron, which spends a large fraction of its time in the region between the protons. Considering the force acting on one of the protons, one sees that it is attracted by the electron, when it is in the middle, more strongly than it is repelled by the other proton. This argument is not precise enough to prove that the resultant force is attractive, but an exact quantum mechanical calculation shows that it is if the protons are not too close together. At close approach proton repulsion dominates, but as one moves the protons apart the attractive force rises to a peak and then soon falls to a low value. The distance, 1.06×10^{-10} metre, at which the force changes sign, corresponds to the potential φ taking its lowest value and is the equilibrium separation of the protons in the ion. This is an example of a central force field that is far from inverse square in character.

A similar attractive force arising from a particle shared between others is found in the strong nuclear force that holds the atomic nucleus together. The simplest example is the deuteron, the nucleus of heavy hydrogen, which consists either of a proton and a neutron or of two neutrons bound by a positive pion (a meson that has a mass 273 times that of an electron when in the free state). There is no repulsive force between the neutrons analogous to the Coulomb repulsion between the protons in the hydrogen ion, and the variation of the attractive force with distance follows the law $F = (g^2/r^2)e^{-r/r_0}$, in which g is a constant analogous to charge in electrostatics and r_0 is a distance of 1.4×10^{-15} metre, which is something like the separation of individual protons and neutrons in a nucleus. At separations closer than r_0, the law of force approximates to an inverse square attraction, but the exponential term kills the attractive force when r is only a few times r_0 (*e.g.*, when r is $5r_0$, the exponential reduces the force 150 times).

Strong nuclear force

Since strong nuclear forces at distances less than r_0 share an inverse square law with gravitational and Coulomb forces, a direct comparison of their strengths is possible. The gravitational force between two protons at a given distance is only about 5×10^{-39} times as strong as the Coulomb force at the same separation, which itself is 1,400 times weaker than the strong nuclear force. The nuclear force is therefore able to hold together a nucleus consisting of protons and neutrons in spite of the Coulomb repulsion of the protons. On the scale of nuclei and atoms, gravitational forces are quite negligible; they only make themselves felt when extremely large numbers of electrically neutral atoms are involved, as on a terrestrial or a cosmological scale.

Field lines. The vector field, $V = -\text{grad}\ \varphi$, associated with a potential φ is always directed normal to the equipotential surfaces, and the variations in space of its direction can be represented by continuous lines drawn accordingly, like those in Figure 8. The arrows show the direction of the force that would act on a positive charge; they thus point away from the charge +3 in its vicinity and toward the charge −1. If the field is of inverse square character (gravitational, electrostatic), the field lines may be drawn to represent both direction and strength of field. Thus from an isolated charge q a large number of radial lines may be drawn, filling the solid angle evenly. Since the field strength falls away as $1/r^2$ and the area of a sphere centred on the charge increases as r^2, the number of lines crossing unit area on each sphere varies as $1/r^2$, in the same way as the field strength. In this case, the density of lines crossing an element of area normal to the lines represents the field strength at that point. The result may be generalized to apply to any distribution of point charges. The field lines are drawn so as to be continuous everywhere except at the charges themselves, which act as sources of lines. From every positive charge q, lines emerge (*i.e.*, with outward-pointing arrows) in number proportional to q, while a similarly proportionate number enter negative charge $-q$. The density of lines then gives a measure of the field strength at any point. This elegant construction holds only for inverse square forces.

Gauss's theorem. At any point in space one may define an element of area dS by drawing a small, flat, closed loop. The area contained within the loop gives the magnitude of the vector area dS, and the arrow representing its direction is drawn normal to the loop. Then, if the electric field in the region of the elementary area is E, the flux through the element is defined as the product of the magnitude dS and the component of E normal to the element—*i.e.*, the scalar product $E \cdot dS$. A charge q at the centre of a sphere of radius r generates a field $\varepsilon = q/4\pi\varepsilon_0 r^2$ on the surface of the sphere whose area is $4\pi r^2$, and the total flux through the surface is $\int_S E \cdot dS = q/\varepsilon_0$. This is independent of r, and the German mathematician Karl Friedrich Gauss showed that it does not depend on q being at the centre nor even on the surrounding surface being spherical. The total flux of ε through a closed surface is equal to $1/\varepsilon_0$ times the total charge contained within it, irrespective of how that charge is arranged. It is readily seen that this result is consistent with the statement in the preceding paragraph—if every charge q within the surface is the source of q/ε_0 field lines, and these lines are continuous except at the charges, the total number leaving through the surface is Q/ε_0, where Q is the total charge. Charges outside the surface contribute nothing, since their lines enter and leave again.

Gauss's theorem in gravitational theory

Gauss's theorem takes the same form in gravitational theory, the flux of gravitational field lines through a closed surface being determined by the total mass within. This enables a proof to be given immediately of a problem that caused Newton considerable trouble. He was able to show, by direct summation over all the elements, that a uniform sphere of matter attracts bodies outside as if the whole mass of the sphere were concentrated at its centre. Now it is obvious by symmetry that the field has the same magnitude everywhere on the surface of the sphere, and this symmetry is unaltered by collapsing the mass to a point at the centre. According to Gauss's theorem, the

total flux is unchanged, and the magnitude of the field must therefore be the same. This is an example of the power of a field theory over the earlier point of view by which each interaction between particles was dealt with individually and the result summed.

Images. A second example illustrating the value of field theories arises when the distribution of charges is not initially known, as when a charge q is brought close to a piece of metal or other electrical conductor and experiences a force. When an electric field is applied to a conductor, charge moves in it; so long as the field is maintained and charge can enter or leave, this movement of charge continues and is perceived as a steady electric current. An isolated piece of conductor, however, cannot carry a steady current indefinitely because there is nowhere for the charge to come from or go to. When q is brought close to the metal, its electric field causes a shift of charge in the metal to a new configuration in which its field exactly cancels the field due to q everywhere on and inside the conductor. The force experienced by q is its interaction with the canceling field. It is clearly a serious problem to calculate E everywhere for an arbitrary distribution of charge, and then to adjust the distribution to make it vanish on the conductor. When, however, it is recognized that after the system has settled down, the surface of the conductor must have the same value of φ everywhere, so that $E = -\text{grad } \varphi$ vanishes on the surface, a number of specific solutions can easily be found.

In Figure 8, for instance, the equipotential surface $\varphi = 0$ is a sphere. If a sphere of uncharged metal is built to coincide with this equipotential, it will not disturb the field in any way. Moreover, once it is constructed, the charge -1 inside may be moved around without altering the field pattern outside, which therefore describes what the field lines look like when a charge $+3$ is moved to the appropriate distance away from a conducting sphere carrying charge -1. More usefully, if the conducting sphere is momentarily connected to the Earth (which acts as a large body capable of supplying charge to the sphere without suffering a change in its own potential), the required charge -1 flows to set up this field pattern. This result can be generalized as follows: if a positive charge q is placed at a distance r from the centre of a conducting sphere of radius a connected to the Earth, the resulting field outside the sphere is the same as if, instead of the sphere, a negative charge $q' = -(a/r)q$ had been placed at a distance $r' = r(1 - a^2/r^2)$ from q on a line joining it to the centre of the sphere. And q is consequently attracted toward the sphere with a force $qq'/4\pi\varepsilon_0 r'^2$, or $q^2 ar/4\pi\varepsilon_0(r^2 - a^2)^2$. The fictitious charge $-q'$ behaves somewhat, but not exactly, like the image of q in a spherical mirror, and hence this way of constructing solutions, of which there are many examples, is called the method of images.

Divergence and Laplace's equation. When charges are not isolated points but form a continuous distribution with a local charge density ρ being the ratio of the charge δq in a small cell to the volume δv of the cell, then the flux of E over the surface of the cell is $\rho\delta v/\varepsilon_0$, by Gauss's theorem, and is proportional to δv. The ratio of the flux to δv is called the divergence of E and is written div E. It is related to the charge density by the equation div $E = \rho/\varepsilon_0$. If E is expressed by its Cartesian components (ε_x, ε_y, ε_z),

$$\text{div } E = \frac{\partial \varepsilon_x}{\partial x} + \frac{\delta \varepsilon_y}{\partial y} + \frac{\partial \varepsilon_z}{\partial z}.$$

And since $E_x = -\partial\varphi/dx$, etc.,

$$\frac{\partial^2 \varphi}{\partial^2 x} + \frac{\partial^2 \varphi}{\partial^2 y} + \frac{\partial^2 \varphi}{\partial^2 z} = -\frac{\rho}{\varepsilon_0}.$$

The expression on the left side is usually written as $\nabla^2\varphi$ and is called the Laplacian of φ. It has the property, as is obvious from its relationship to ρ, of being unchanged if the Cartesian axes of x, y, and z are turned bodily into any new orientation.

If any region of space is free of charges, $\rho = 0$ and $\nabla^2\varphi = 0$ in this region. The latter is Laplace's equation, for which many methods of solution are available, providing a powerful means of finding electrostatic (or gravitational) field patterns.

Non-conservative fields. The magnetic field B is an example of a vector field that cannot in general be described as the gradient of a scalar potential. There are no isolated poles to provide, as electric charges do, sources for the field lines. Instead, the field is generated by currents and forms vortex patterns around any current-carrying conductor. Figure 9 shows the field lines for a single straight wire. If one forms the line integral $\int B \cdot dl$ around the closed path formed by any one of these field lines, each increment $B \cdot \delta l$ has the same sign and, obviously, the integral cannot vanish as it does for an electrostatic field. The value it takes is proportional to the total current enclosed by the path. Thus every path that encloses the conductor yields the same value for $\int B \cdot dl$; *i.e.*, $\mu_0 I$, where I is the current and μ_0 is a constant for any particular choice of units in which B, l, and I are to be measured. Magnetic fields

If no current is enclosed by the path, the line integral vanishes and a potential φ_B may be defined. Indeed, in the example shown in Figure 9, a potential may be defined even for paths that enclose the conductor, but it is many-valued because it increases by a standard increment $\mu_0 I$ every time the path encircles the current. A contour map of height would represent a spiral staircase (or, better, a spiral ramp) by a similar many-valued contour. The conductor carrying I is in this case the axis of the ramp. Like E in a charge-free region, where div $E = 0$, so also div $B = 0$; and where φ_B may be defined, it obeys Laplace's equation, $\nabla^2\varphi_B = 0$.

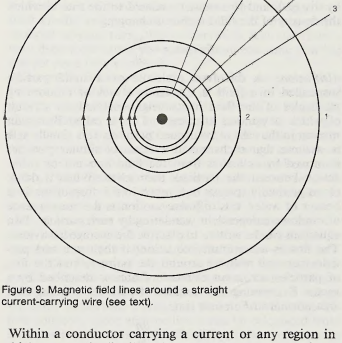


Figure 9: Magnetic field lines around a straight current-carrying wire (see text).

Within a conductor carrying a current or any region in which current is distributed rather than closely confined to a thin wire, no potential φ_B can be defined. For now the change in φ_B after traversing a closed path is no longer zero or an integral multiple of a constant $\mu_0 I$ but is rather μ_0 times the current enclosed in the path and therefore depends on the path chosen. To relate the magnetic field to the current, a new function is needed, the curl, whose name suggests the connection with circulating field lines.

The curl of a vector, say, curl B, is itself a vector quantity. To find the component of curl B along any chosen direction, draw a small closed path of area A lying in the plane normal to that direction, and evaluate the line integral $\int B \cdot dl$ around the path. As the path is shrunk in size, the integral diminishes with the area, and the limit of $A^{-1}\int B \cdot dl$ is the component of curl B in the chosen direction. The direction in which the vector curl B points is the direction in which $A^{-1}\int B \cdot dl$ is largest.

To apply this to the magnetic field in a conductor carrying current, the current density J is defined as a vector pointing along the direction of current flow, and the magnitude of J is such that JA is the total current flowing across a small area A normal to J. Now the line integral of B around the edge of this area is A curl B if A is very

small, and this must equal μ_0 times the contained current. It follows that

$$\text{curl } B = \mu_0 J.$$

Expressed in Cartesian coordinates,

$$\frac{\partial B_z}{\partial z} - \frac{\partial B_z}{\partial y} = \mu_0 J_x,$$

with similar expressions for J_y and J_z. These are the differential equations relating the magnetic field to the currents that generate it.

A magnetic field also may be generated by a changing electric field, and an electric field by a changing magnetic field. The description of these physical processes by differential equations relating curl B to $\partial E/\partial t$, and curl E to $\partial B/\partial t$ is the heart of Maxwell's electromagnetic theory and illustrates the power of the mathematical methods characteristic of field theories. Further examples will be found in the mathematical description of fluid motion, in which the local velocity $v(r)$ of fluid particles constitutes a field to which the notions of divergence and curl are naturally applicable.

Examples of differential equations for fields. *Continuity.* An incompressible fluid flows so that the net flux of fluid into or out of a given volume within the fluid is zero. Since the divergence of a vector describes the net flux out of an infinitesimal element, divided by the volume of the element, the velocity vector v in an incompressible fluid must obey the equation $\text{div } v = 0$. If the fluid is compressible, however, and its density $\rho(r)$ varies with position because of pressure or temperature variations, the net outward flux of mass from some small element is determined by $\text{div } (\rho v)$, and this must be related to the rate at which the density of the fluid within is changing:

$$\text{div } (\rho v) = -\frac{\partial \rho}{\partial t}.$$

Diffusion. A dissolved molecule or a small particle suspended in a fluid is constantly struck at random by molecules of the fluid in its neighbourhood, as a result of which it wanders erratically. This is called Brownian motion in the case of suspended particles. It is usually safe to assume that each one in a cloud of similar particles is moved by collisions from the fluid and not by interaction between the particles themselves. When a dense cloud gradually spreads out, much like a drop of ink in a beaker of water, this diffusive motion is the consequence of random, independent wandering by each particle. Two equations can be written to describe the average behaviour. The first is a continuity equation: if there are $n(r)$ particles per unit volume around the point r, and the flux of particles across an element of area is described by a vector F, meaning the number of particles crossing unit area normal to F in unit time,

Phenomenon of Brownian motion

$$\text{div } F = -\frac{\partial n}{\partial t}$$

describes the conservation of particles. Secondly, Fick's law states that the random wandering causes an average drift of particles from regions where they are denser to regions where they are rarer, and that the mean drift rate is proportional to the gradient of density and in the opposite sense to the gradient:

$$F = -D \text{ grad } n,$$

where D is a constant—the diffusion constant.

These two equations can be combined into one differential equation for the changes that n will undergo,

$$D\nabla^2 n = \frac{\partial n}{\partial t},$$

which defines uniquely how any initial distribution of particles will develop with time. Thus the spreading of a small drop of ink is rather closely described by the particular solution,

$$n = \left(\frac{C}{t^{3/2}}\right) \exp\left(\frac{-r^2}{4Dt}\right),$$

in which C is a constant determined by the total number of particles in the ink drop. When t is very small at the start of the process, all the particles are clustered near the origin of r, but, as t increases, the radius of the cluster increases in proportion to the square root of the time, while the density at the centre drops as the three-halves power to keep the total number constant. The distribution of particles with distance from the centre at three different times is shown in Figure 10. From this diagram one may calculate what fraction, after any chosen interval, has moved farther than some chosen distance from the origin. Moreover, since each particle wanders independently of the rest, it also gives the probability that a single particle will migrate farther than this in the same time. Thus a problem relating to the behaviour of a single particle, for which only an average answer can usefully be given, has been converted into a field equation and solved rigorously. This is a widely used technique in physics.

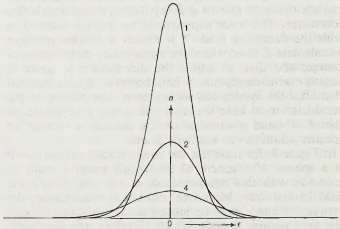


Figure 10: Diffusive spread of a cloud of particles initially concentrated at a point. The value given for each curve represents the time elapsed since n, the particles per unit volume around point r, began to disperse (see text).

Further examples of field equations. The equations describing the propagation of waves (electromagnetic, acoustic, deep water waves, and ripples) are discussed in relevant articles, as is the Schrödinger equation for probability waves that governs particle behaviour in quantum mechanics (see below *Fundamental constituents of matter*). The field equations that embody the special theory of relativity are more elaborate with space and time coordinates no longer independent of each other, though the geometry involved is still Euclidean. In the general theory of relativity, the geometry of this four-dimensional spacetime is non-Euclidean (see RELATIVITY).

CONSERVATION LAWS AND EXTREMAL PRINCIPLES

It is a consequence of Newton's laws of motion that the total momentum remains constant in a system completely isolated from external influences. The only forces acting on any part of the system are those exerted by other parts; if these are taken in pairs, according to the third law, A exerts on B a force equal and opposite to that of B on A. Since, according to the second law, the momentum of each changes at a rate equal to the force acting on it, the momentum change of A is exactly equal and opposite to that of B when only mutual forces between these two are considered. Because the effects of separate forces are additive, it follows that for the system as a whole no momentum change occurs. The centre of mass of the whole system obeys the first law in remaining at rest or moving at a constant velocity, so long as no external influences are brought to bear. This is the oldest of the conservation laws and is invoked frequently in solving dynamic problems.

Conservation of angular momentum

The total angular momentum (also called moment of momentum) of an isolated system about a fixed point is conserved as well. The angular momentum of a particle of mass m moving with velocity v at the instant when it is at a distance r from the fixed point is $mr \wedge v$. The quantity written as $r \wedge v$ is a vector (the vector product of r and v) having components with respect to Cartesian axes

$$r \wedge v = (yv_z - zv_y,\ zv_x - xv_z,\ xv_y - yv_x).$$

The meaning is more easily appreciated if all the particles

lie and move in a plane. The angular momentum of any one particle is the product of its momentum mv and the distance of nearest approach of the particle to the fixed point if it were to continue in a straight line. The vector is drawn normal to the plane. Conservation of total angular momentum does not follow immediately from Newton's laws but demands the additional assumption that any pair of forces, action and reaction, are not only equal and opposite but act along the same line. This is always true for central forces, but it holds also for the frictional force developed along sliding surfaces. If angular momentum were not conserved, one might find an isolated body developing a spontaneous rotation with respect to the distant stars or, if rotating like the Earth, changing its rotational speed without any external cause. Such small changes as the Earth experiences are explicable in terms of disturbances from without—*e.g.,* tidal forces exerted by the Moon. The law of conservation of angular momentum is not called into question.

Nevertheless, there are noncentral forces in nature, as, for example, when a charged particle moves past a bar magnet. If the line of motion and the axis of the magnet lie in a plane, the magnet exerts a force on the particle perpendicular to the plane while the magnetic field of the moving particle exerts an equal and opposite force on the magnet. At the same time, it exerts a couple tending to twist the magnet out of the plane. Angular momentum is not conserved unless one imagines that the balance of angular momentum is distributed in the space around the magnet and charge and changes as the particle moves past. The required result is neatly expressed by postulating the possible existence of magnetic poles that would generate a magnetic field analogous to the electric field of a charge (a bar magnet behaves roughly like two such poles of opposite sign, one near each end). Then there is associated with each pair, consisting of a charge q and a pole P, angular momentum $\mu_0 Pq/4\pi$, as if the electric and magnetic fields together acted like a gyroscope spinning about the line joining P and q. With this contribution included in the sum, angular momentum is always conserved.

The device of associating mechanical properties with the fields, which up to this point had appeared merely as convenient mathematical constructions, has even greater implications when conservation of energy is considered. This conservation law, which is regarded as basic to physics, seems at first sight, from an atomic point of view, to be almost trivial. If two particles interact by central forces, for which a potential function φ may be defined such that grad φ gives the magnitude of the force experienced by each, it follows from Newton's laws of motion that the sum of φ and of their separate kinetic energies, defined as $1/2mv^2$, remains constant. This sum is defined to be the total energy of the two particles and, by its definition, is automatically conserved. The argument may be extended to any number of particles interacting by means of central forces; a potential energy function may always be found, depending only on the relative positions of the particles, which may be added to the sum of the kinetic energies (depending only on the velocities) to give a total energy that is conserved.

Conservation of energy

The concept of potential energy, thus introduced as a formal device, acquires a more concrete appearance when it is expressed in terms of electric and magnetic field strengths for particles interacting by virtue of their charges. The quantities $1/2\varepsilon_0 E^2$ and $B^2/2\mu_0$ may be interpreted as the contributions per unit volume of the electric and magnetic fields to the potential energy, and, when these are integrated over all space and added to the kinetic energy, the total energy thus expressed is a conserved quantity. These expressions were discovered during the heyday of ether theories, according to which all space is permeated by a medium capable of transmitting forces between particles (see above). The electric and magnetic fields were interpreted as descriptions of the state of strain of the ether, so that the location of stored energy throughout space was no more remarkable than it would be in a compressed spring. With the abandonment of the ether theories following the rise of relativity theory, this visualizable model ceased to have validity.

The idea of energy as a real constituent of matter has, however, become too deeply rooted to be abandoned lightly, and most physicists find it useful to continue treating electric and magnetic fields as more than mathematical constructions. Far from being empty, free space is viewed as a storehouse for energy, with E and B providing not only an inventory but expressions for its movements as represented by the momentum carried in the fields. Wherever E and B are both present, and not parallel, there is a flux of energy, amounting to $E \wedge B/\mu_0$, crossing unit area and moving in a direction normal to the plane defined by E and B. This energy in motion confers momentum on the field, $E \wedge B/\mu_0 c$, per unit volume as if there were mass associated with the field energy. Indeed, the English physicist J.J. Thomson showed in 1881 that the energy stored in the fields around a moving charged particle varies as the square of the velocity as if there were extra mass carried with the electric field around the particle. Herein lie the seeds of the general mass–energy relationship developed by Einstein in his special theory of relativity; $E = mc^2$ expresses the association of mass with every form of energy. Neither of two separate conservation laws, that of energy and that of mass (the latter particularly the outcome of countless experiments involving chemical change), is in this view perfectly true, but together they constitute a single conservation law, which may be expressed in two equivalent ways—conservation of mass, if to the total energy E is ascribed mass E/c^2, or conservation of energy, if to each mass m is ascribed energy mc^2. The delicate measurements by Eötvös and later workers (see above) show that the gravitational forces acting on a body do not distinguish different types of mass, whether intrinsic to the fundamental particles or resulting from their kinetic and potential energies. For all its apparently artificial origins, then, this conservation law enshrines a very deep truth about the material universe, one that has not yet been fully explored.

Conservation of mass-energy

An equally fundamental law, for which no exception is known, is that the total electrical charge in an isolated system is conserved. In the production of a negatively charged electron by an energetic gamma ray, for example, a positively charged positron is produced simultaneously. An isolated electron cannot disappear, though an electron and a positron, whose total charge is zero and whose mass is $2m_e$ (twice the mass of an electron), may simultaneously be annihilated. The energy equivalent of the destroyed mass appears as gamma ray energy $2m_e c^2$.

For macroscopic systems—*i.e.,* those composed of objects massive enough for their atomic structure to be discounted in the analysis of their behaviour—the conservation law for energy assumes a different aspect. In the collision of two perfectly elastic objects, to which billiard balls are a good approximation, momentum and energy are both conserved. Given the paths and velocities before collision, those after collision can be calculated from the conservation laws alone. In reality, however, although momentum is always conserved, the kinetic energy of the separating balls is less than what they had on approach. Soft objects, indeed, may adhere on collision, losing most of their kinetic energy. The lost energy takes the form of heat, raising the temperature (if only imperceptibly) of the colliding objects. From the atomic viewpoint the total energy of a body may be divided into two portions: on the one hand, the external energy consisting of the potential energy associated with its position and the kinetic energy of motion of its centre of mass and its spin; and, on the other, the internal energy due to the arrangement and motion of its constituent atoms. In an inelastic collision the sum of internal and external energies is conserved, but some of the external energy of bodily motion is irretrievably transformed into internal random motions. The conservation of energy is expressed in the macroscopic language of the first law of thermodynamics—namely, energy is conserved provided that heat is taken into account. The irreversible nature of the transfer from external energy of organized motion to random internal energy is a manifestation of the second law of thermodynamics.

The irreversible degradation of external energy into random internal energy also explains the tendency of all

systems to come to rest if left to themselves. If there is a configuration in which the potential energy is less than for any slightly different configuration, the system may find stable equilibrium here because there is no way in which it can lose more external energy, either potential or kinetic. This is an example of an extremal principle—that a state of stable equilibrium is one in which the potential energy is a minimum with respect to any small changes in configuration. It may be regarded as a special case of one of the most fundamental of physical laws, the principle of increase of entropy, which is a statement of the second law of thermodynamics in the form of an extremal principle—the equilibrium state of an isolated physical system is that in which the entropy takes the maximum possible value. This matter is discussed further below and, in particular, in the article THERMODYNAMICS.

Manifestations of the extremal principle

The earliest extremal principle to survive in modern physics was formulated by the French mathematician Pierre de Fermat in about 1660. As originally stated, the path taken by a ray of light between two fixed points in an arrangement of mirrors, lenses, and so forth, is that which takes the least time. The laws of reflection and refraction may be deduced from this principle if it is assumed as Fermat did, correctly, that in a medium of refractive index μ light travels more slowly than in free space by a factor μ. Strictly, the time taken along a true ray path is either less or greater than for any neighbouring path. If all paths in the neighbourhood take the same time, the two chosen points are such that light leaving one is focused on the other. The perfect example is exhibited by an elliptical mirror, such as the one in Figure 11; all paths from F_1 to the ellipse and thence to F_2 have the same length. In conventional optical terms, the ellipse has the property that every choice of paths obeys the law of reflection, and every ray from F_1 converges after reflection onto F_2. Also shown in the figure are two reflecting surfaces tangential to the ellipse that do not have the correct curvature to focus light from F_1 onto F_2. A ray is reflected from F_1 to F_2 only at the point of contact. For the flat reflector the path taken is the shortest of all in the vicinity, while for the reflector that is more strongly curved than the ellipse it is the longest. Fermat's principle and its application to focusing by mirrors and lenses finds a natural explanation in the wave theory of light (see LIGHT).

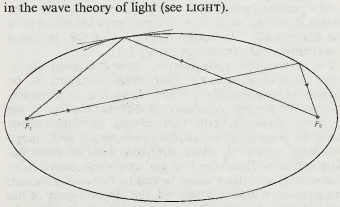


Figure 11: An elliptic mirror focusing all rays of light from F_1 onto F_2 (see text).

Principle of least action

A similar extremal principle in mechanics, the principle of least action, was proposed by the French mathematician and astronomer Pierre-Louis Moreau de Maupertuis but rigorously stated only much later, especially by the Irish mathematician and scientist William Rowan Hamilton in 1835. Though very general, it is well enough illustrated by a simple example, the path taken by a particle between two points A and B in a region where the potential $\varphi(r)$ is everywhere defined. Once the total energy E of the particle has been fixed, its kinetic energy T at any point P is the difference between E and the potential energy φ at P. If any path between A and B is assumed to be followed, the velocity at each point may be calculated from T, and hence the time t between the moment of departure from A and passage through P. The action for this path is found by evaluating the integral $\int_A^B (T-\varphi)dt$, and the actual path taken by the particle is that for which the action is minimal. It may be remarked that both Fermat and Maupertuis were guided by Aristotelian notions of economy in nature that have been found, if not actively misleading, too imprecise to retain a place in modern science.

Fermat's and Hamilton's principles are but two examples out of many whereby a procedure is established for finding the correct solution to a problem by discovering under what conditions a certain function takes an extremal value. The advantages of such an approach are that it brings into play the powerful mathematical techniques of the calculus of variations and, perhaps even more important, that in dealing with very complex situations it may allow a systematic approach by computational means to a solution that may not be exact but is near enough the right answer to be useful.

Fermat's principle, stated as a theorem concerning light rays but later restated in terms of the wave theory, found an almost exact parallel in the development of wave mechanics. The association of a wave with a particle by the physicists Louis-Victor de Broglie and Erwin Schrödinger was made in such a way that the principle of least action followed by an analogous argument.

FUNDAMENTAL CONSTITUENTS OF MATTER

Development of the atomic theory. The idea that matter is composed of atoms goes back to the Greek philosophers, notably Democritus, and has never since been entirely lost sight of, though there have been periods when alternative views were more generally preferred. Newton's contemporaries, Robert Hooke and Robert Boyle, in particular, were atomists, but their interpretation of the sensation of heat as random motion of atoms was overshadowed for more than a century by the conception of heat as a subtle fluid dubbed caloric. It is a tribute to the strength of caloric theory that it enabled the French scientist Sadi Carnot to arrive at his great discoveries in thermodynamics. In the end, however, the numerical rules for the chemical combination of different simple substances, together with the experiments on the conversion of work into heat by Benjamin Thompson (Count Rumford) and James Prescott Joule, led to the downfall of the theory of caloric. Nevertheless, the rise of ether theories to explain the transmission of light and electromagnetic forces through apparently empty space postponed for many decades the general reacceptance of the concept of atoms. The discovery in 1858 by the German scientist and philosopher Hermann von Helmholtz of the permanence of vortex motions in perfectly inviscid fluids encouraged the invention—throughout the latter half of the 19th century and especially in Great Britain—of models in which vortices in a structureless ether played the part otherwise assigned to atoms. In recent years the recognition that certain localized disturbances in a fluid, the so-called solitary waves, might persist for a very long time has led to attempts, so far unsuccessful, to use them as models of fundamental particles.

These attempts to describe the basic constituents of matter in the familiar language of fluid mechanics were at least atomic theories in contrast to the anti-atomistic movement at the end of the 19th century in Germany under the influence of Ernst Mach and Wilhelm Ostwald. For all their scientific eminence, their argument was philosophical rather than scientific, springing as it did from the conviction that the highest aim of science is to describe the relationship between different sensory perceptions without the introduction of unobservable concepts. Nonetheless, an inspection of the success of their contemporaries using atomic models shows why this movement failed. It suffices to mention the systematic construction of a kinetic theory of matter in which the physicists Ludwig Boltzmann of Austria and J. Willard Gibbs of the United States were the two leading figures. To this may be added Hendrik Lorentz's electron theory, which explained in satisfying detail many of the electrical properties of matter; and, as a crushing argument for atomism, the discovery and explanation of X-ray diffraction by Max von Laue of Germany and his collaborators, a discovery that was quickly developed, following the lead of the British physicist William Henry Bragg and his son Lawrence, into a systematic technique for mapping the precise atomic structure of crystals.

While the concept of atoms was thus being made in-

dispensable, the ancient belief that they were probably structureless and certainly indestructible came under devastating attack. J.J. Thomson's discovery of the electron in 1897 soon led to the realization that the mass of an atom largely resides in a positively charged part, electrically neutralized by a cloud of much lighter electrons. A few years later Ernest Rutherford and Frederick Soddy showed how the emission of alpha and beta particles from radioactive elements causes them to be transformed into elements of different chemical properties. By 1913, with Rutherford as the leading figure, the foundations of the modern theory of atomic structure were laid. It was determined that a small, massive nucleus carries all the positive charge whose magnitude, expressed as a multiple of the fundamental charge of the proton, is the atomic number. An equal number of electrons carrying a negative charge numerically equal to that of the proton form a cloud whose diameter is several thousand times that of the nucleus around which they swarm. The atomic number determines the chemical properties of the atom, and in alpha decay a helium nucleus, whose atomic number is 2, is emitted from the radioactive nucleus, leaving one whose atomic number is reduced by 2. In beta decay the nucleus in effect gains one positive charge by emitting a negative electron and thus has its atomic number increased by unity.

Modern theory of atomic structure

The nucleus, itself a composite body, was soon being described in various ways, none completely wrong but none uniquely right. Pivotal was James Chadwick's discovery in 1932 of the neutron, a nuclear particle with very nearly the same mass as the proton but no electric charge. After this discovery, investigators came to view the nucleus as consisting of protons and neutrons, bound together by a force of limited range, which at close quarters was strong enough to overcome the electrical repulsion between the protons. A free neutron survives for only a few minutes before disintegrating into a readily observed proton and electron, along with an elusive neutrino, which has no charge and zero, or at most extremely small, mass. The disintegration of a neutron also may occur inside the nucleus, with the expulsion of the electron and neutrino; this is the beta-decay process. It is common enough among the heavy radioactive nuclei but does not occur with all nuclei because the energy released would be insufficient for the reorganization of the resulting nucleus. Certain nuclei have a higher-than-ideal ratio of protons to neutrons and may adjust the proportion by the reverse process, a proton being converted into a neutron with the expulsion of a positron and an antineutrino. For example, a magnesium nucleus containing 12 protons and 11 neutrons spontaneously changes to a stable sodium nucleus with 11 protons and 12 neutrons. The positron resembles the electron in all respects except for being positively rather than negatively charged. It was the first antiparticle to be discovered. Its existence had been predicted, however, by Dirac after he had formulated the quantum mechanical equations describing the behaviour of an electron (see below). This was one of the most spectacular achievements of a spectacular albeit brief epoch, during which the basic conceptions of physics were revolutionized.

Rise of quantum mechanics. The idea of the quantum was introduced by the German physicist Max Planck in 1900 in response to the problems posed by the spectrum of radiation from a hot body, but the development of quantum theory soon became closely tied to the difficulty of explaining by classical mechanics the stability of Rutherford's nuclear atom. Bohr led the way in 1913 with his model of the hydrogen atom, but it was not until 1925 that the arbitrary postulates of his quantum theory found consistent expression in the new quantum mechanics that was formulated in apparently different but in fact equivalent ways by Heisenberg, Schrödinger, and Dirac (see MECHANICS: *Quantum mechanics*). In Bohr's model the motion of the electron around the proton was analyzed as if it were a classical problem, mathematically the same as that of a planet around the Sun, but it was additionally postulated that, of all the orbits available to the classical particle, only a discrete set was to be allowed, and Bohr devised rules for determining which orbits they were. In Schrödinger's wave mechanics the problem is also written down in the first place as if it were a classical problem, but, instead of proceeding to a solution of the orbital motion, the equation is transformed by an explicitly laid down procedure from an equation of particle motion to an equation of wave motion. The newly introduced mathematical function Ψ, the amplitude of Schrödinger's hypothetical wave, is used to calculate not how the electron moves but rather what the probability is of finding the electron in any specific place if it is looked for there.

Wave mechanics

Schrödinger's prescription reproduced in the solutions of the wave equation the postulates of Bohr but went much further. Bohr's theory had come to grief when even two electrons, as in the helium atom, had to be considered together, but the new quantum mechanics encountered no problems in formulating the equations for two or any number of electrons moving around a nucleus. Solving the equations was another matter; yet, numerical procedures were applied with devoted patience to a few of the simpler cases and demonstrated beyond cavil that the only obstacle to solution was calculational and not an error of physical principle. Modern computers have vastly extended the range of application of quantum mechanics not only to heavier atoms but also to molecules and assemblies of atoms in solids, and always with such success as to inspire full confidence in the prescription.

From time to time many physicists feel uneasy that it is necessary first to write down the problem to be solved as though it were a classical problem and them to subject it to an artificial transformation into a problem in quantum mechanics. It must be realized, however, that the world of experience and observation is not the world of electrons and nuclei. When a bright spot on a television screen is interpreted as the arrival of a stream of electrons, it is still only the bright spot that is perceived and not the electrons. The world of experience is described by the physicist in terms of visible objects, occupying definite positions at definite instants of time—in a word, the world of classical mechanics. When the atom is pictured as a nucleus surrounded by electrons, this picture is a necessary concession to human limitations; there is no sense in which one can say that, if only a good enough microscope were available, this picture would be revealed as genuine reality. It is not that such a microscope has not been made; it is actually impossible to make one that will reveal this detail. The process of transformation from a classical description to an equation of quantum mechanics, and from the solution of this equation to the probability that a specified experiment will yield a specified observation, is not to be thought of as a temporary expedient pending the development of a better theory. It is better to accept this process as a technique for predicting the observations that are likely to follow from an earlier set of observations. Whether electrons and nuclei have an objective existence in reality is a metaphysical question to which no definite answer can be given. There is, however, no doubt that to postulate their existence is, in the present state of physics, an inescapable necessity if a consistent theory is to be constructed to describe economically and exactly the enormous variety of observations on the behaviour of matter. The habitual use of the language of particles by physicists induces and reflects the conviction that, even if the particles elude direct observation, they are as real as any everyday object.

Following the initial triumphs of quantum mechanics, Dirac in 1928 extended the theory so that it would be compatible with the special theory of relativity. Among the new and experimentally verified results arising from this work was the seemingly meaningless possibility that an electron of mass m might exist with any negative energy between $-mc^2$ and $-\infty$. Between $-mc^2$ and $+mc^2$, which is in relativistic theory the energy of an electron at rest, no state is possible. It became clear that other predictions of the theory would not agree with experiment if the negative-energy states were brushed aside as an artifact of the theory without physical significance. Eventually Dirac was led to propose that all the states of negative energy, infinite in number, are already occupied with electrons and that these, filling all space evenly, are imperceptible. If, however, one of the negative-energy electrons is

given more than $2mc^2$ of energy, it can be raised into a positive-energy state, and the hole it leaves behind will be perceived as an electron-like particle, though carrying a positive charge. Thus this act of excitation leads to the simultaneous appearance of a pair of particles—an ordinary negative electron and a positively charged but otherwise identical positron. This process was observed in cloud-chamber photographs by Carl David Anderson of the United States in 1932. The reverse process was recognized at the same time; it can be visualized either as an electron and a positron mutually annihilating one another, with all their energy (two lots of rest energy, each mc^2, plus their kinetic energy) being converted into gamma rays (electromagnetic quanta), or as an electron losing all this energy as it drops into the vacant negative-energy state that simulates a positive charge. When an exceptionally energetic cosmic-ray particle enters the Earth's atmosphere, it initiates a chain of such processes in which gamma rays generate electron–positron pairs; these in turn emit gamma rays which, though of lower energy, are still capable of creating more pairs, so that what reaches the Earth's surface is a shower of many millions of electrons and positrons.

Electron–positron pairs

Not unnaturally, the suggestion that space was filled to infinite density with unobservable particles was not easily accepted in spite of the obvious successes of the theory. It would have seemed even more outrageous had not other developments already forced theoretical physicists to contemplate abandoning the idea of empty space. Quantum mechanics carries the implication that no oscillatory system can lose all its energy; there must always remain at least a "zero-point energy" amounting to $h\nu/2$ for an oscillator with natural frequency ν (h is Planck's constant). This also seemed to be required for the electromagnetic oscillations constituting radio waves, light, X rays, and gamma rays. Since there is no known limit to the frequency ν, their total zero-point energy density is also infinite; like the negative-energy electron states, it is uniformly distributed throughout space, both inside and outside matter, and presumed to produce no observable effects.

Developments in particle physics. It was at about this moment, say 1930, in the history of the physics of fundamental particles that serious attempts to visualize the processes in terms of everyday notions were abandoned in favour of mathematical formalisms. Instead of seeking modified procedures from which the awkward, unobservable infinities had been banished, the thrust was toward devising prescriptions for calculating what observable processes could occur and how frequently and how quickly they would occur. An empty cavity which would be described by a classical physicist as capable of maintaining electromagnetic waves of various frequencies, ν, and arbitrary amplitude now remains empty (zero-point oscillation being set aside as irrelevant) except insofar as photons, of energy $h\nu$, are excited within it. Certain mathematical operators have the power to convert the description of the assembly of photons into the description of a new assembly, the same as the first except for the addition or removal of one. These are called creation or annihilation operators, and it need not be emphasized that the operations are performed on paper and in no way describe a laboratory operation having the same ultimate effect. They serve, however, to express such physical phenomena as the emission of a photon from an atom when it makes a transition to a state of lower energy. The development of these techniques, especially after their supplementation with the procedure of renormalization (which systematically removes from consideration various infinite energies that naive physical models throw up with embarrassing abundance), has resulted in a rigorously defined procedure that has had dramatic successes in predicting numerical results in close agreement with experiment. It is sufficient to cite the example of the magnetic moment of the electron. According to Dirac's relativistic theory, the electron should possess a magnetic moment whose strength he predicted to be exactly one Bohr magneton ($eh/4\pi m$, or 9.27×10^{-24} joule per tesla). In practice, this has been found to be not quite right, as, for instance, in the experiment of Lamb and Rutherford mentioned earlier; more recent determinations give 1.0011596522 Bohr magnetons. Calculations by means of the theory of quantum electrodynamics give 1.0011596525 in impressive agreement.

Creation and annihilation operators

This account represents the state of the theory in about 1950, when it was still primarily concerned with problems related to the stable fundamental particles, the electron and the proton, and their interaction with electromagnetic fields. Meanwhile, studies of cosmic radiation at high altitudes—those conducted on mountains or involving the use of balloon-borne photographic plates—had revealed the existence of the pi-meson (pion), a particle 273 times as massive as the electron, which disintegrates into the mu-meson (muon), 207 times as massive as the electron, and a neutrino. Each muon in turn disintegrates into an electron and two neutrinos. The pion has been identified with the hypothetical particle postulated in 1935 by the Japanese physicist Yukawa Hideki as the particle that serves to bind protons and neutrons in the nucleus. Many more unstable particles have been discovered in recent years. Some of them, just as in the case of the pion and the muon, are lighter than the proton, but many are more massive. An account of such particles is given in the article SUBATOMIC PARTICLES.

The term particle is firmly embedded in the language of physics, yet a precise definition has become harder as more is learned. When examining the tracks in a cloud-chamber or bubble-chamber photograph, one can hardly suspend disbelief in their having been caused by the passage of a small charged object. However, the combination of particle-like and wavelike properties in quantum mechanics is unlike anything in ordinary experience, and, as soon as one attempts to describe in terms of quantum mechanics the behaviour of a group of identical particles (*e.g.*, the electrons in an atom), the problem of visualizing them in concrete terms becomes still more intractable. And this is before one has even tried to include in the picture the unstable particles or to describe the properties of a stable particle like the proton in relation to quarks. These hypothetical entities, worthy of the name particle to the theoretical physicist, are apparently not to be detected in isolation, nor does the mathematics of their behaviour encourage any picture of the proton as a molecule-like composite body constructed of quarks. Similarly, the theory of the muon is not the theory of an object composed, as the word is normally used, of an electron and two neutrinos. The theory does, however, incorporate such features of particle-like behaviour as will account for the observation of the track of a muon coming to an end and that of an electron starting from the end point. At the heart of all fundamental theories is the concept of countability. If a certain number of particles is known to be present inside a certain space, that number will be found there later, unless some have escaped (in which case they could have been detected and counted) or turned into other particles (in which case the change in composition is precisely defined). It is this property, above all, that allows the idea of particles to be preserved.

Concept of the particle

Undoubtedly, however, the term is being strained when it is applied to photons that can disappear with nothing to show but thermal energy or be generated without limit by a hot body so long as there is energy available. They are a convenience for discussing the properties of a quantized electromagnetic field, so much so that the condensed-matter physicist refers to the analogous quantized elastic vibrations of a solid as phonons without persuading himself that a solid really consists of an empty box with particle-like phonons running about inside. If, however, one is encouraged by this example to abandon belief in photons as physical particles, it is far from clear why the fundamental particles should be treated as significantly more real, and, if a question mark hangs over the existence of electrons and protons, where does one stand with atoms or molecules? The physics of fundamental particles does indeed pose basic metaphysical questions to which neither philosophy nor physics has answers. Nevertheless, the physicist has confidence that his constructs and the mathematical processes for manipulating them represent a technique for correlating the outcomes of observation and experiment with such precision and over so wide a

range of phenomena that he can afford to postpone deeper inquiry into the ultimate reality of the material world.

SIMPLICITY AND COMPLEXITY

The search for fundamental particles and the mathematical formalism with which to describe their motions and interactions has in common with the search for the laws governing gravitational, electromagnetic, and other fields of force the aim of finding the most economical basis from which, in principle, theories of all other material processes may be derived. Some of these processes are simple—a single particle moving in a given field of force, for example—if the term refers to the nature of the system studied and not to the mathematical equipment that may sometimes be brought to bear. A complex process, on the other hand, is typically one in which many interacting particles are involved and for which it is hardly ever possible to proceed to a complete mathematical solution. A computer may be able to follow in detail the movement of thousands of atoms interacting in a specified way, but a wholly successful study along these lines does no more than display on a large scale and at an assimilable speed what nature achieves on its own. Much can be learned from these studies, but, if one is primarily concerned with discovering what will happen in given circumstances, it is frequently quicker and cheaper to do the experiment than to model it on a computer. In any case, computer modeling of quantum mechanical, as distinct from Newtonian, behaviour becomes extremely complicated as soon as more than a few particles are involved.

The art of analyzing complex systems is that of finding the means to extract from theory no more information than one needs. It is normally of no value to discover the speed of a given molecule in a gas at a given moment; it is, however, very valuable to know what fraction of the molecules possess a given speed. The correct answer to this question was found by Maxwell, whose argument was ingenious and plausible. More rigorously, Boltzmann showed that it is possible to proceed from the conservation laws governing molecular encounters to general statements, such as the distribution of velocities, which are largely independent of how the molecules interact. In thus laying the foundations of statistical mechanics, Boltzmann provided an object lesson in how to avoid recourse to the fundamental laws, replacing them with a new set of rules appropriate to highly complex systems. This point is discussed further in *Entropy and disorder* below.

Analyzing complex systems by means of simplified models

The example of statistical mechanics is but one of many that together build up a hierarchical structure of simplified models whose function is to make practicable the analysis of systems at various levels of complexity. Ideally, the logical relationship between each successive pair of levels should be established so that the analyst may have confidence that the methods he applies to his special problem are buttressed by the enormous corpus of fact and theory that comprises physical knowledge at all levels. It is not in the nature of the subject for every connection to be proved with mathematical rigour, but, where this is lacking, experiment will frequently indicate what trust may be placed in the intuitive steps of the argument.

For instance, it is out of the question to solve completely the quantum mechanical problem of finding the stationary states in which an atomic nucleus containing perhaps 50 protons or neutrons can exist. Nevertheless, the energy of these states can be measured and models devised in which details of particle position are replaced by averages, such that when the simplified model is treated by the methods of quantum mechanics the measured energy levels emerge from the calculations. Success is attained when the rules for setting up the model are found to give the right result for every nucleus. Similar models had been devised earlier by the English physicist Douglas R. Hartree to describe the cloud of electrons around the nucleus. The increase in computing power made it feasible to add extra details to the model so that it agreed even better with the measured properties of atoms. It is worth noting that when the extranuclear electrons are under consideration it is frequently unnecessary to refer to details of the nucleus, which might just as well be a point charge; even if this is too simplistic, a small number of extra facts usually suffices. In the same way, when the atoms combine chemically and molecules in a gas or a condensed state interact, most of the details of electronic structure within the atom are irrelevant or can be included in the calculation by introducing a few extra parameters; these are often treated as empirical properties. Thus the degree to which an atom is distorted by an electric field is often a significant factor in its behaviour, and the investigator dealing with the properties of assemblies of atoms may prefer to use the measured value rather than the atomic theorist's calculation of what it should be. However, he knows that enough of these calculations have been successfully carried out for his use of measured values in any specific case to be a time-saver rather than a denial of the validity of his model.

These examples from atomic physics can be multiplied at all levels so that a connected hierarchy exists, ranging from fundamental particles and fields, through atoms and molecules, to gases, liquids, and solids that were studied in detail and reduced to quantitative order well before the rise of atomic theory. Beyond this level lie the realms of the Earth sciences, the planetary systems, the interior of stars, galaxies, and the Cosmos as a whole. And with the interior of stars and the hypothetical early universe, the entire range of models must be brought to bear if one is to understand how the chemical elements were built up or to determine what sort of motions are possible in the unimaginably dense, condensed state of neutron stars.

The following sections make no attempt to explore all aspects and interconnections of complex material systems, but they highlight a few ideas which pervade the field and which indicate the existence of principles that find little place in the fundamental laws yet are the outcome of their operation.

SYMMETRY

The normal behaviour of a gas on cooling is to condense into a liquid and then into a solid, though the liquid phase may be left out if the gas starts at a low enough pressure. The solid phase of a pure substance is usually crystalline, having the atoms or molecules arranged in a regular pattern so that a suitable small sample may define the whole. The unit cell is the smallest block out of which the pattern can be formed by stacking replicas. The checkerboard in Figure 12 illustrates the idea; here the unit cell has been chosen out of many possibilities to contain one white square and one black, dissected into quarters. For crystals, of course, the unit cell is three-dimensional. A very wide variety of arrangements is exhibited by different substances, and it is the great triumph of X-ray crystallography to have provided the means for

Concept of the unit cell

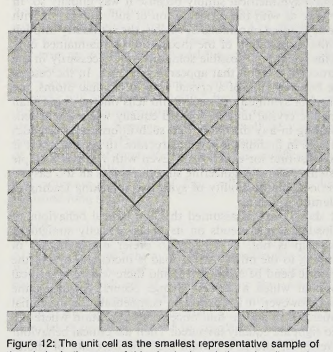


Figure 12: The unit cell as the smallest representative sample of the whole. In the case of this checkerboard, the unit cell consists of one white square, and one shaded square dissected into quarters.

determining experimentally what arrangement is involved in each case.

One may ask whether mathematical techniques exist for deducing the correct result independently of experiment, and the answer is almost always no. An individual sulfur atom, for example, has no features that reflect its preference, in the company of others, for forming rings of eight. This characteristic can only be discovered theoretically by calculating the total energy of different-sized rings and of other patterns and determining after much computation that the ring of eight has the lowest energy of all. Even then the investigator has no assurance that there is no other arrangement which confers still lower energy. In one of the forms taken by solid sulfur, the unit cell contains 128 atoms in a complex of rings. It would be an inspired guess to hit on this fact without the aid of X rays or the expertise of chemists, and mathematics provides no systematic procedure as an alternative to guessing or relying on experiment.

Nevertheless, it may be possible in simpler cases to show that calculations of the energy are in accord with the observed crystal forms. Thus, when silicon is strongly compressed, it passes through a succession of different crystal modifications for each of which the variation with pressure of the energy can be calculated. The pressure at which a given change of crystal form takes place is that at which the energy takes the same value for both modifications involved. As this pressure is reached, one gives way to the other for the possession of the lower energy. The fact that the calculation correctly describes not only the order in which the different forms occur but also the pressures at which the changeovers take place indicates that the physical theory is in good shape; only the power is lacking in the mathematics to predict behaviour from first principles.

The changes in symmetry that occur at the critical points where one modification changes to another are complex examples of a widespread phenomenon for which simple analogues exist. A perfectly straight metal strip, firmly fixed to a base so that it stands perfectly upright, remains straight as an increasing load is placed on its upper end until a critical load is reached. Any further load causes the strip to heel over and assume a bent form, and it only takes a minute disturbance to determine whether it will bend to the left or to the right. The fact that either outcome is equally likely reflects the left–right symmetry of the arrangement, but once the choice is made the symmetry is broken. The subsequent response to changing load and the small vibrations executed when the strip is struck lightly are characteristic of the new unsymmetrical shape. If one wishes to calculate the behaviour, it is essential to avoid assuming that an arrangement will always remain symmetrical simply because it was initially so. In general, as with the condensation of sulfur atoms or with the crystalline transitions in silicon, the symmetry implicit in the formulation of the theory will be maintained only in the totality of possible solutions, not necessarily in the particular solution that appears in practice. In the case of the condensation of a crystal from individual atoms, the spherical symmetry of each atom tells one no more than that the crystal may be formed equally well with its axis pointing in any direction; and such information provides no help in finding the crystal structure. In general, there is no substitute for experiment. Even with relatively simple systems such as engineering structures, it is all too easy to overlook the possibility of symmetry breaking leading to calamitous failure.

Symmetry breaking

It should not be assumed that the critical behaviour of a loaded strip depends on its being perfectly straight. If the strip is not, it is likely to prefer one direction of bending to the other. As the load is increased, so will the intrinsic bend be exaggerated, and there will be no critical point at which a sudden change occurs. By tilting the base, however, it is possible to compensate for the initial imperfection and to find once more a position where left and right are equally favoured. Then the critical behaviour is restored, and at a certain load the necessity of choice is present as with a perfect strip. The study of this and numerous more complex examples is the province of the so-called catastrophe theory. A catastrophe, in the special sense used here, is a situation in which a continuously varying input to a system gives rise to a discontinuous change in the response at a critical point. The discontinuities may take many forms, and their character may be sensitive in different ways to small changes in the parameters of the system. Catastrophe theory is the term used to describe the systematic classification, by means of topological mathematics, of these discontinuities. Wide-ranging though the theory may be, it cannot at present include in its scope most of the symmetry-breaking transitions undergone by crystals.

Catastrophe theory

ENTROPY AND DISORDER

As is explained in detail in the article THERMODYNAMICS, the laws of thermodynamics make possible the characterization of a given sample of matter—after it has settled down to equilibrium with all parts at the same temperature—by ascribing numerical measures to a small number of properties (pressure, volume, energy, and so forth). One of these is entropy. As the temperature of the body is raised by adding heat, its entropy as well as its energy is increased. On the other hand, when a volume of gas enclosed in an insulated cylinder is compressed by pushing on the piston, the energy in the gas increases while the entropy stays the same or, usually, increases a little. In atomic terms, the total energy is the sum of all the kinetic and potential energies of the atoms, and the entropy, it is commonly asserted, is a measure of the disorderly state of the constituent atoms. The heating of a crystalline solid until it melts and then vaporizes is a progress from a well-ordered, low-entropy state to a disordered, high-entropy state. The principal deduction from the second law of thermodynamics (or, as some prefer, the actual statement of the law) is that, when an isolated system makes a transition from one state to another, its entropy can never decrease. If a beaker of water with a lump of sodium on a shelf above it is sealed in a thermally insulated container and the sodium is then shaken off the shelf, the system, after a period of great agitation, subsides to a new state in which the beaker contains hot sodium hydroxide solution. The entropy of the resulting state is higher than the initial state, as can be demonstrated quantitatively by suitable measurements.

The idea that a system cannot spontaneously become better ordered but can readily become more disordered, even if left to itself, appeals to one's experience of domestic economy and confers plausibility on the law of increase of entropy. As far as it goes, there is much truth in this naive view of things, but it cannot be pursued beyond this point without a much more precise definition of disorder. Thermodynamic entropy is a numerical measure that can be assigned to a given body by experiment; unless disorder can be defined with equal precision, the relation between the two remains too vague to serve as a basis for deduction. A precise definition is to be found by considering the number, labeled W, of different arrangements that can be taken up by a given collection of atoms, subject to their total energy being fixed. In quantum mechanics, W is the number of different quantum states that are available to the atoms with this total energy (strictly, in a very narrow range of energies). It is so vast for objects of everyday size as to be beyond visualization; for the helium atoms contained in one cubic centimetre of gas at atmospheric pressure and at 0° C the number of different quantum states can be written as 1 followed by 170 million million million zeroes (written out, the zeroes would fill nearly one trillion sets of the *Encyclopædia Britannica*).

Thermodynamic entropy

The science of statistical mechanics, as founded by the aforementioned Ludwig Boltzmann and J. Willard Gibbs, relates the behaviour of a multitude of atoms to the thermal properties of the material they constitute. Boltzmann and Gibbs, along with Max Planck, established that the entropy, S, as derived through the second law of thermodynamics, is related to W by the formula $S = k \ln W$, where k is the Boltzmann constant (1.380662×10^{-23} joule per kelvin) and $\ln W$ is the natural (Naperian) logarithm of W. By means of this and related formulas it is possible in principle, starting with the quantum mechanics of the

constituent atoms, to calculate the measurable thermal properties of the material. Unfortunately, there are rather few systems for which the quantum mechanical problems succumb to mathematical analysis, but among these are gases and many solids, enough to validate the theoretical procedures linking laboratory observations to atomic constitution.

When a gas is thermally isolated and slowly compressed, the individual quantum states change their character and become mixed together, but the total number W does not alter. In this change, called adiabatic, entropy remains constant. On the other hand, if a vessel is divided by a partition, one side of which is filled with gas while the other side is evacuated, piercing the partition to allow the gas to spread throughout the vessel greatly increases the number of states available so that W and the entropy rise. The act of piercing requires little effort and may even happen spontaneously through corrosion. To reverse the process, waiting for the gas to accumulate accidentally on one side and then stopping the leak, would mean waiting for a time compared to which the age of the universe would be imperceptibly short. The chance of finding an observable decrease in entropy for an isolated system can be ruled out.

This does not mean that a part of a system may not decrease in entropy at the expense of at least as great an increase in the rest of the system. Such processes are indeed commonplace but only when the system as a whole is not in thermal equilibrium. Whenever the atmosphere becomes supersaturated with water and condenses into a cloud, the entropy per molecule of water in the droplets is less than it was prior to condensation. The remaining atmosphere is slightly warmed and has a higher entropy. The spontaneous appearance of order is especially obvious when the water vapour condenses into snow crystals. A domestic refrigerator lowers the entropy of its contents while increasing that of its surroundings. Most important of all, the state of nonequilibrium of the Earth irradiated by the much hotter Sun provides an environment in which the cells of plants and animals may build order—*i.e.*, lower their local entropy at the expense of their environment. The Sun provides a motive power that is analogous (though much more complex in detailed operation) to the electric cable connected to the refrigerator. There is no evidence pointing to any ability on the part of living matter to run counter to the principle of increasing (overall) disorder as formulated in the second law of thermodynamics.

Irreversible tendency toward disorder

The irreversible tendency toward disorder provides a sense of direction for time which is absent from space. One may traverse a path between two points in space without feeling that the reverse journey is forbidden by physical laws. The same is not true for time travel, and yet the equations of motion, whether in Newtonian or quantum mechanics, have no such built-in irreversibility. A motion picture of a large number of particles interacting with one another looks equally plausible whether run forward or backward. To illustrate and resolve this paradox it is convenient to return to the example of a gas enclosed in a vessel divided by a pierced partition. This time, however, only 100 atoms are involved (not 3×10^{19} as in one cubic centimetre of helium), and the hole is made so small that atoms pass through only rarely and no more than one at a time. This model is easily simulated on a computer, and Figure 13 shows a typical sequence during which there are 500 transfers of atoms across the partition. The number on one side starts at the mean of 50 and fluctuates randomly while not deviating greatly from the mean. Where the fluctuations are larger than usual, as indicated by the arrows, there is no systematic tendency for their growth to the peak to differ in form from the decay from it. This is in accord with the reversibility of the motions when examined in detail.

If one were to follow the fluctuations for a very long time and single out those rare occasions when a particular number occurred that was considerably greater than 50, say 75, one would find that the next number is more likely to be 74 than 76. Such would be the case because, if there are 75 atoms on one side of the partition, there will be only 25 on the other, and it is three times more likely that one atom will leave the 75 than that one will be gained from the 25. Also, since the detailed motions are reversible, it is three times more likely that the 75 was preceded by a 74 rather than by a 76. In other words, if one finds the system in a state that is far from the mean, it is highly probable that the system has just managed to get there and is on the point of falling back. If the system has momentarily fluctuated into a state of lower entropy, the entropy will be found to increase again immediately.

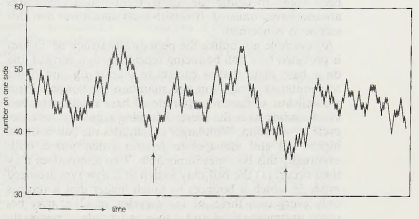


Figure 13: Fluctuations in the number of particles, out of 100, on one side of a perforated partition dividing a box into equal halves (see text).

It might be thought that this argument has already conceded the possibility of entropy decreasing. It has indeed, but only for a system on the minute scale of 100 atoms. The same computation carried out for 3×10^{19} atoms would show that one would have to wait interminably (*i.e.*, enormously longer than the age of the universe) for the number on one side to fluctuate even by as little as one part per million. A physical system as big as the Earth, let alone the entire Galaxy—if set up in thermodynamic equilibrium and given unending time in which to evolve—might eventually have suffered such a huge fluctuation that the condition known today could have come about spontaneously. In that case man would find himself, as he does, in a universe of increasing entropy as the fluctuation recedes. Boltzmann, it seems, was prepared to take this argument seriously on the grounds that sentient creatures could only appear as the aftermath of a large enough fluctuation. What happened during the inconceivably prolonged waiting period is irrelevant. Modern cosmology shows, however, that the universe is ordered on a scale enormously greater than is needed for living creatures to evolve, and Boltzmann's hypothesis is correspondingly rendered improbable in the highest degree. Whatever started the universe in a state from which it could evolve with an increase of entropy, it was not a simple fluctuation from equilibrium. The sensation of time's arrow is thus referred back to the creation of the universe, an act that lies beyond the scrutiny of the physical scientist.

It is possible, however, that in the course of time the universe will suffer "heat death," having attained a condition of maximum entropy, after which tiny fluctuations are all that will happen. If so, these will be reversible, like the graph of Figure 13, and will give no indication of a direction of time. Yet, because this undifferentiated cosmic soup will be devoid of structures necessary for consciousness, the sense of time will in any case have vanished long since.

CHAOS

Many systems can be described in terms of a small number of parameters and behave in a highly predictable manner. Were this not the case, the laws of physics might never have been elucidated. If one maintains the swing of a pendulum by tapping it at regular intervals, say once per swing, it will eventually settle down to a regular oscillation. Now let it be jolted out of its regularity; in due course it will revert to its previous oscillation as if nothing had disturbed it. Systems that respond in this well-behaved manner have been studied extensively and have frequently

been taken to define the norm, from which departures are somewhat unusual. It is with such departures that this section is concerned.

An example not unlike the periodically struck pendulum is provided by a ball bouncing repeatedly in a vertical line on a base plate that is caused to vibrate up and down to counteract dissipation and maintain the bounce. With a small but sufficient amplitude of base motion the ball synchronizes with the plate, returning regularly once per cycle of vibration. With larger amplitudes the ball bounces higher but still manages to remain synchronized until eventually this becomes impossible. Two alternatives may then occur: (1) the ball may switch to a new synchronized mode in which it bounces so much higher that it returns only every two, three, or more cycles, or (2) it may become unsynchronized and return at irregular, apparently random, intervals. Yet, the behaviour is not random in the way that raindrops strike a small area of surface at irregular intervals. The arrival of a raindrop allows one to make no prediction of when the next will arrive; the best one can hope for is a statement that there is half a chance that the next will arrive before the lapse of a certain time. By contrast, the bouncing ball is described by a rather simple set of differential equations that can be solved to predict without fail when the next bounce will occur and how fast the ball will be moving on impact, given the time of the last bounce and the speed of that impact. In other words, the system is precisely determinate, yet to the casual observer it is devoid of regularity. Systems that are determinate but irregular in this sense are called chaotic; like so many other scientific terms, this is a technical expression that bears no necessary relation to the word's common usage.

Determinate yet irregular systems

The coexistence of irregularity with strict determinism can be illustrated by an arithmetic example, one that lay behind some of the more fruitful early work in the study of chaos, particularly by the physicist Mitchell J. Feigenbaum following an inspiring exposition by Robert M. May. Suppose one constructs a sequence of numbers starting with an arbitrarily chosen x_0 (between 0 and 1) and writes the next in the sequence, x_1, as $Ax_0(1 - x_0)$; proceeding in the same way to $x_2 = Ax_1(1 - x_1)$, one can continue indefinitely, and the sequence is completely determined by the initial value x_0 and the value chosen for A. Thus, starting from $x_0 = 0.9$ with $A = 2$, the sequence rapidly settles to a constant value: 0.09, 0.18, 0.2952, 0.4161, 0.4859, 0.4996, 0.5000, 0.5000, and so forth.

When A lies between 2 and 3, it also settles to a constant but takes longer to do so. It is when A is increased above 3 that the sequence shows more unexpected features. At first, until A reaches 3.42, the final pattern is an alternation of two numbers, but with further small increments of A it changes to a cycle of 4, followed by 8, 16, and so forth at ever-closer intervals of A. By the time A reaches 3.57, the length of the cycle has grown beyond bounds—it shows no periodicity however long one continues the sequence. This is the most elementary example of chaos, but it is easy to construct other formulas for generating number sequences that can be studied rapidly with the aid of the smallest programmable computer. By such "experimental arithmetic" Feigenbaum found that the transition from regular convergence through cycles of 2, 4, 8, and so forth to chaotic sequences followed strikingly similar courses for all, and he gave an explanation that involved great subtlety of argument and was almost rigorous enough for pure mathematicians.

The chaotic sequence shares with the chaotic bouncing of the ball in the earlier example the property of limited predictability, as distinct from the strong predictability of the periodically driven pendulum and of the regular sequence found when A is less than 3. Just as the pendulum, having been disturbed, eventually settles back to its original routine, so the regular sequence, for a given choice of A, settles to the same final number whatever initial value x_0 may be chosen. By contrast, when A is large enough to generate chaos, the smallest change in x_0 leads eventually to a completely different sequence, and the smallest disturbance to the bouncing ball switches it to a different but equally chaotic pattern. This is illustrated for the number sequence in Figure 14, where two sequences are plotted (successive points being joined by straight lines) for $A = 3.7$ and x_0 chosen to be 0.9 and 0.9000009, a difference of one part per million. For the first 35 terms the sequences differ by too little to appear on the graph, but a record of the numbers themselves shows them diverging steadily until by the 40th term the sequences are unrelated. Although the sequence is completely determined by the first term, one cannot predict its behaviour for any considerable number of terms without extremely precise knowledge of the first term. The initial divergence of the two sequences is roughly exponential, each pair of terms being different by an amount greater than that of the preceding pair by a roughly constant factor. Put another way, to predict the sequence in this particular case out to n terms, one must know the value of x_0 to better than $n/8$ places of decimals. If this were the record of a chaotic physical system (*e.g.,* the bouncing ball), the initial state would be determined by measurement with an accuracy of perhaps 1 percent (*i.e.,* two decimal places), and prediction would be valueless beyond 16 terms. Different systems, of course, have different measures of their "horizon of predictability," but all chaotic systems share the property that every extra place of decimals in one's knowledge of the starting point only pushes the horizon a small extra distance away. In practical terms, the horizon of predictability is an impassable barrier. Even if it is possible to determine the initial conditions with extremely high precision, every physical system is susceptible to random disturbances from outside that grow exponentially in a chaotic situation until they have swamped any initial prediction. It is highly probable that atmospheric movements, governed by well-defined equations, are in a state of chaos. If so, there can be little hope of extending indefinitely the range of weather forecasting except in the most general terms. There are clearly certain features of climate, such as annual cycles of temperature and rainfall, which are exempt from the ravages of chaos. Other large-scale processes may still allow long-range prediction, but the more detail one asks for in a forecast, the sooner will it lose its validity.

Horizon of predictability

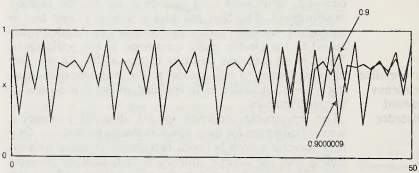


Figure 14: Sensitivity of a chaotic number sequence to initial value, illustrating the horizon of predictability (see text).

Linear systems for which the response to a force is strictly proportional to the magnitude of the force do not show chaotic behaviour. The pendulum, if not too far from the vertical, is a linear system, as are electrical circuits containing resistors that obey Ohm's law or capacitors and inductors for which voltage and current also are proportional. The analysis of linear systems is a well-established technique that plays an important part in the education of a physicist. It is relatively easy to teach, since the range of behaviour exhibited is small and can be encapsulated in a few general rules. Nonlinear systems, on the other hand, are bewilderingly versatile in their modes of behaviour and are, moreover, very commonly unamenable to elegant mathematical analysis. Until large computers became readily available, the natural history of nonlinear systems was little explored and the extraordinary prevalence of chaos unappreciated. To a considerable degree physicists have been persuaded, in their innocence, that predictability is a characteristic of a well-established theoretical structure; given the equations defining a system, it is only a matter of computation to determine how it will behave. However, once it becomes clear how many systems are sufficiently nonlinear to be considered for chaos,

Prevalence of chaos in nonlinear systems

it has to be recognized that prediction may be limited to short stretches set by the horizon of predictability. Full comprehension is not to be achieved by establishing firm fundamentals, important though they are, but must frequently remain a tentative process, a step at a time, with frequent recourse to experiment and observation in the event that prediction and reality have diverged too far.

BIBLIOGRAPHY. ERIC M. ROGERS, *Physics for the Inquiring Mind: The Methods, Nature, and Philosophy of Physical Science* (1960), is especially good on the origins of astronomy and mechanics, with minimal mathematics. Of the many general student texts, the *Berkeley Physics Course,* 5 vol. (1965–71), covering mechanics, electricity and magnetism, waves, quantum physics, and statistical physics; and DAVID HALLIDAY and ROBERT RESNICK, *Fundamentals of Physics,* 3rd ed. (1988), are recommended. *The Feynman Lectures on Physics,* 3 vol. (1963–65), by RICHARD P. FEYNMAN, ROBERT B. LEIGHTON, and MATTHEW SANDS, instructs students and teachers in the whole range of physical concepts, with characteristically revealing insights. See also JEFFERSON HANE WEAVER (ed.), *The World of Physics: A Small Library of the Literature of Physics from Antiquity to the Present,* 3 vol. (1987), an anthology covering the history of the major concepts of physics.

Expositions of more limited scope, reflecting on general principles for the benefit of nonspecialists, include H. BONDI, *Assumption and Myth in Physical Theory* (1967); RICHARD P. FEYNMAN, *The Character of Physical Law* (1965); and J.M. ZIMAN, *Public Knowledge: An Essay Concerning the Social Dimension of Science* (1968). At a more advanced level, M.S. LONGAIR, *Theoretical Concepts in Physics* (1984); and PETER GALISON, *How Experiments End* (1987), illustrate typical research procedures by means of case studies. ERNST MACH, *The Science of Mechanics,* 6th ed. (1974; originally published in German, 9th ed., 1933), is both a detailed history and a classic critique of fundamental assumptions. E.T. WHITTAKER, *A History of the Theories of Aether and Electricity,* vol. 1, *The Classical Theories,* rev. and enlarged ed. (1951, reprinted 1973), is equally detailed but less philosophically oriented.

Special topics in more recent physics are treated by ALBERT EINSTEIN, *Relativity: The Special & the General Theory* (1920; originally published in German, 1917), and many later editions; WOLFGANG RINDLER, *Essential Relativity: Special, General, and Cosmological,* rev. 2nd ed. (1979); STEVEN WEINBERG, *The Discovery of Subatomic Particles* (1983), and *The First Three Minutes: A Modern View of the Origin of the Universe,* updated ed. (1988); NATHAN SPIELBERG and BRYON D. ANDERSON, *Seven Ideas that Shook the Universe* (1985); P.C.W. DAVIES, *The Forces of Nature,* 2nd ed. (1986); A. ZEE, *Fearful Symmetry: The Search for Beauty in Modern Physics* (1986); and TONY HEY and PATRICK WALTERS, *The Quantum Universe* (1987).

The principles of catastrophe theory are presented, without mathematical detail, in V.I. ARNOLD, *Catastrophe Theory,* 2nd rev. and expanded ed. (1986; originally published in Russian, 2nd ed. enlarged, 1983), which is notably scornful of speculative applications. A full treatment is provided in TIM POSTON and IAN STEWART, *Catastrophe Theory and Its Applications* (1978).

Introductions to chaotic processes are found in A.B. PIPPARD, *Response and Stability: An Introduction to the Physical Theory* (1985); and JAMES GLEICK, *Chaos: Making a New Science* (1987). More systematic is J.M.T. THOMPSON and H.B. STEWART, *Nonlinear Dynamics and Chaos: Geometrical Methods for Engineers and Scientists* (1986). Anthologies of influential early papers are BAI-LIN HAO (comp.), *Chaos* (1984); and PREDRAG CVITANOVIĆ (comp.), *Universality in Chaos* (1984).

(A.B.P.)

The Physical Sciences

The physical sciences are the ongoing refined and systematic efforts to understand the inorganic world and the results of those efforts. They are generally divided into four broad areas: physics, chemistry, astronomy, and the Earth sciences. The historical development of the first three is surveyed in this article; due attention is also given to their scope, principal concerns, and methods. A complementary treatment of the Earth sciences is provided in a separate article.

Physics, in its modern sense, was founded in the mid-19th century as a synthesis of several older sciences—namely, those of mechanics, optics, acoustics, electricity, magnetism, heat, and the physical properties of matter. The synthesis was based in large part on the recognition that the different forces of nature are related and are, in fact, interconvertible because they are forms of energy.

The boundary between physics and chemistry is somewhat arbitrary. As it has developed in the 20th century, physics is concerned with the structure and behaviour of individual atoms and their components, while chemistry deals with the properties and reactions of molecules. These latter depend on energy, especially heat, as well as on atoms; hence, there is a strong link between physics and chemistry. Chemists tend to be more interested in the specific properties of different elements and compounds, whereas physicists are concerned with general properties shared by all matter.

Astronomy is the science of the entire universe beyond the Earth; it includes the Earth's gross physical properties, such as its mass and rotation, insofar as they interact with other bodies in the solar system. Until the 18th century, astronomers were concerned primarily with the Sun, Moon, planets, and comets. During the last two centuries, however, the study of stars, galaxies, nebulas, and the interstellar medium has become increasingly important. Celestial mechanics, the science of the motion of planets and other solid objects within the solar system, was the first testing ground for Newton's laws of motion and thereby helped to establish the fundamental principles of classical (that is, pre-20th-century) physics. Astrophysics, the study of the physical properties of celestial bodies, arose during the 19th century and is closely connected with the determination of the chemical composition of those bodies. In the 20th century physics and astronomy have become more intimately linked through cosmological theories, especially those based on the theory of relativity.

For coverage of topics related to physics, chemistry, and astronomy in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 111–133, 10/31, and 10/32.

The article is divided into the following sections:

History of the physical sciences 828
  Heritage of antiquity and the Middle Ages 828
    Ancient Middle Eastern and Greek astronomy
    Greek physics
    Islāmic and medieval science
  The scientific revolution 830
    Astronomy
    Physics
    Chemistry
  Science from the Enlightenment to the 20th century 832
    Celestial mechanics and astronomy
    Optics
    Electricity and magnetism
    Chemistry
  Developments and trends of the 20th century 835
    Astronomy
    Physics
    Chemistry
Physics 840
  The scope of physics 840
    Mechanics
    The study of gravitation
    The study of heat, thermodynamics, and statistical mechanics
    The study of electricity and magnetism
    Optics
    Atomic and chemical physics
    Condensed-matter physics
    Nuclear physics
    Particle physics
    Quantum mechanics
    Relativistic mechanics
    Conservation laws and symmetry
    Fundamental forces and fields
  The methodology of physics 844
  Relations between physics and other disciplines and society 845
    Influence of physics on related disciplines
    Influence of related disciplines on physics
    The physicist in society
Chemistry 845
  The scope of chemistry 846
    Analytical chemistry
    Inorganic chemistry
    Organic chemistry
    Biochemistry
    Polymer chemistry
    Physical chemistry
    Industrial chemistry
  The methodology of chemistry 849
    Studies of molecular structure
    Investigations of chemical transformations
  Chemistry and society 851
Astronomy 852
  The scope of astronomy 852
    Determining astronomical distances
    Study of the solar system
    Study of the stars
    Study of the Milky Way Galaxy
    Study of other galaxies and related phenomena
    Cosmology
  The methodology of astronomy 857
    Telescopic observations
    Use of radiation detectors
    Theoretical approaches
  Impact of astronomy 858
Bibliography 858

History of the physical sciences

HERITAGE OF ANTIQUITY AND THE MIDDLE AGES

The physical sciences ultimately derive from the rationalistic materialism that emerged in classical Greece, itself an outgrowth of magical and mythical views of the world. The Greek philosophers of the 6th and 5th centuries BC abandoned the animism of the poets and explained the world in terms of ordinarily observable natural processes. These early philosophers posed the broad questions that still underlie science: How did the world order emerge from chaos? What is the origin of multitude and variety in the world? How can motion and change be accounted for? What is the underlying relation between form and matter? Greek philosophy answered these questions in terms that provided the framework for science for approximately 2,000 years.

Ancient Middle Eastern and Greek astronomy. Western astronomy had its origins in Egypt and Mesopotamia. Egyptian astronomy, which was neither a very well-developed nor an influential study, was largely concerned with time reckoning. Its main lasting contribution was the civil

Egyptian and Mesopotamian origins

calendar of 365 days, consisting of 12 months of 30 days each and five additional festival days at the end of each year. This calendar played an important role in the history of astronomy, allowing astronomers to calculate the number of days between any two sets of observations.

Babylonian astronomy, dating back to about 1800 BC, constitutes one of the earliest systematic, scientific treatments of the physical world. In contrast to the Egyptians, the Babylonians were interested in the accurate prediction of astronomical phenomena, especially the first appearance of the new Moon. Using the zodiac as a reference, by the 4th century BC, they developed a complex system of arithmetic progressions and methods of approximation by which they were able to predict first appearances. At no point in the Babylonian astronomical literature is there the least evidence of the use of geometric models. The mass of observations they collected and their mathematical methods were important contributions to the later flowering of astronomy among the Greeks.

The Pythagoreans (5th century BC) were responsible for one of the first Greek astronomical theories. Believing that the order of the cosmos is fundamentally mathematical, they held that it is possible to discover the harmonies of the universe by contemplating the regular motions of the heavens. Postulating a central fire about which all the heavenly bodies including the Earth and Sun revolve, they constructed the first physical model of the solar system. Subsequent Greek astronomy derived its character from a comment ascribed to Plato, in the 4th century BC, who is reported to have instructed the astronomers to "save the phenomena" in terms of uniform circular motion. That is to say, he urged them to develop predictively accurate theories using only combinations of uniform circular motion. As a result, Greek astronomers never regarded their geometric models as true or as being physical descriptions of the machinery of the heavens. They regarded them simply as tools for predicting planetary positions.

First physical model of the solar system

Eudoxus of Cnidus (4th century BC) was the first of the Greek astronomers to rise to Plato's challenge. He developed a theory of homocentric spheres, a model that represented the universe by sets of nesting concentric spheres the motions of which combined to produce the planetary and other celestial motions. Using only uniform circular motions, Eudoxus was able to "save" the rather complex planetary motions with some success. His theory required four homocentric spheres for each planet and three each for the Sun and Moon. The system was modified by Callippus, a student of Eudoxus, who added spheres to improve the theory, especially for Mercury and Venus. Aristotle, in formulating his cosmology, adopted Eudoxus' homocentric spheres as the actual machinery of the heavens. The Aristotelian cosmos was like an onion consisting of a series of some 55 spheres nested about the Earth, which was fixed at the centre. In order to unify the system, Aristotle added spheres in order to "unroll" the motions of a given planet so that they would not be transmitted to the next inner planet.

The theory of homocentric spheres failed to account for two sets of observations: (1) brightness changes suggesting that planets are not always the same distance from the Earth, and (2) bounded elongations (*i.e.,* Venus is never observed to be more than about 48° and Mercury never more than about 24° from the Sun). Heracleides of Pontus (4th century BC) attempted to solve these problems by having Venus and Mercury revolve about the Sun, rather than the Earth, and having the Sun and other planets revolve in turn about the Earth, which he placed at the centre. In addition, to account for the daily motions of the heavens, he held that the Earth rotates on its axis. Heracleides' theory had little impact in antiquity except perhaps on Aristarchus of Samos (3rd century BC), who apparently put forth a heliocentric hypothesis similar to the one Copernicus was to propound in the 16th century.

An early heliocentric hypothesis

Hipparchus (fl. 130 BC) made extensive contributions to both theoretical and observational astronomy. Basing his theories on an impressive mass of observations, he was able to work out theories of the Sun and Moon that were more successful than those of any of his predecessors. His primary conceptual tool was the eccentric circle, a circle in which the Earth is at some point eccentric to the geometric centre. He used this device to account for various irregularities and inequalities observed in the motions of the Sun and Moon. He also proved that the eccentric circle is mathematically equivalent to a geometric figure called an epicycle-deferent system, a proof probably first made by Apollonius of Perga a century earlier.

Among Hipparchus' observations, one of the most significant was that of the precession of the equinoxes—*i.e.,* a gradual apparent increase in longitude between any fixed star and the equinoctial point (either of two points on the celestial sphere where the celestial equator crosses the ecliptic). Thus the north celestial pole, the point on the celestial sphere defined as the apparent centre of rotation of the stars, moves relative to the stars in its vicinity. In the heliocentric theory, this effect is ascribed to a change in the Earth's rotational axis, which traces out a conical path around the axis of the orbital plane.

Claudius Ptolemy (fl. AD 140) applied the theory of epicycles to compile a systematic account of Greek astronomy. He elaborated theories for each of the planets, as well as for the Sun and Moon. His theory generally fitted the data available to him with a good degree of accuracy, and his book, the *Almagest,* became the vehicle by which Greek astronomy was transmitted to astronomers of the Middle Ages and Renaissance. It essentially molded astronomy for the next millennium and a half.

Greek physics. Several kinds of physical theories emerged in ancient Greece, including both generalized hypotheses about the ultimate structure of nature and more specific theories that considered the problem of motion from both metaphysical and mathematical points of view. Attempting to reconcile the antithesis between the underlying unity and apparent multitude and diversity of nature, the Greek atomists Leucippus (mid-5th century BC), Democritus (late 5th century BC), and Epicurus (late 4th and early 3rd century BC) asserted that nature consists of immutable atoms moving in empty space. According to this theory, the various motions and configurations of atoms and clusters of atoms are the causes of all the phenomena of nature.

The first atomists

In contrast to the particulate universe of the atomists, the Stoics (principally Zeno, of Citium, bridging 4th and 3rd centuries BC, Chrysippus [3rd century BC], and Poseidonius of Apamea [fl. *c.* 100 BC]) insisted on the continuity of nature, conceiving of both space and matter as continuous and as infused with an active, airlike spirit—pneuma—which serves to unify the frame of nature. The inspiration for the Stoic emphasis on pneumatic processes probably arose from earlier experiences with the "spring" (*i.e.,* compressibility and pressure) of the air. Neither the atomic theory nor Stoic physics survived the criticism of Aristotle and his theory.

In his physics, Aristotle was primarily concerned with the philosophical question of the nature of motion as one variety of change. He assumed that a constant motion requires a constant cause; that is to say, as long as a body remains in motion, a force must be acting on that body. He considered the motion of a body through a resisting medium as proportional to the force producing the motion and inversely proportional to the resistance of the medium. Aristotle used this relationship to argue against the possibility of the existence of a void, for in a void resistance is zero, and the relationship loses meaning. He considered the cosmos to be divided into two qualitatively different realms, governed by two different kinds of laws. In the terrestrial realm, within the sphere of the Moon, rectilinear up-and-down motion is characteristic. Heavy bodies, by their nature, seek the centre and tend to move downward in a natural motion. It is unnatural for a heavy body to move up, and such unnatural or violent motion requires an external cause. Light bodies, in direct contrast, move naturally upward. In the celestial realm, uniform circular motion is natural, thus producing the motions of the heavenly bodies.

The physics of Aristotle

Archimedes (3rd century BC) fundamentally applied mathematics to the solution of physical problems and brilliantly employed physical assumptions and insights leading to mathematical demonstrations, particularly in problems

of statics and hydrostatics. He was thus able to derive the law of the lever rigorously and to deal with problems of the equilibrium of floating bodies.

Islāmic and medieval science. Greek science reached a zenith with the work of Ptolemy in the 2nd century AD. The lack of interest in theoretical questions in the Roman world reduced science in the Latin West to the level of predigested handbooks and encyclopaedias that had been distilled many times. Social pressures, political persecution, and the anti-intellectual bias of some of the early Church Fathers drove the few remaining Greek scientists and philosophers to the East. There they ultimately found a welcome when the rise of Islām in the 7th century stimulated interest in scientific and philosophical subjects. Most of the important Greek scientific texts were preserved in Arabic translations. Although the Muslims did not alter the foundations of Greek science, they made several important contributions within its general framework. When interest in Greek learning revived in western Europe during the 12th and 13th centuries, scholars turned to Islāmic Spain for the scientific texts. A spate of translations resulted in the revival of Greek science in the West and coincided with the rise of the universities. Working within a predominantly Greek framework, scientists of the late Middle Ages reached high levels of sophistication and prepared the ground for the scientific revolution of the 16th and 17th centuries.

Revival of Greek science

Mechanics was one of the most highly developed sciences pursued in the Middle Ages. Operating within a fundamentally Aristotelian framework, medieval physicists criticized and attempted to improve many aspects of Aristotle's physics.

The problem of projectile motion was a crucial one for Aristotelian mechanics, and the analysis of this problem represents one of the most impressive medieval contributions to physics. Because of the assumption that continuation of motion requires the continued action of a motive force, the continued motion of a projectile after losing contact with the projector required explanation. Aristotle himself had proposed explanations of the continuation of projectile motion in terms of the action of the medium. The ad hoc character of these explanations rendered them unsatisfactory to most of the medieval commentators, who nevertheless retained the fundamental assumption that continued motion requires a continuing cause.

The most fruitful alternative to Aristotle's attempts to explain projectile motion resulted from the concept of impressed force. According to this view, there is an incorporeal motive force that is imparted to the projectile, causing it to continue moving. Such views were espoused by John Philoponus of Alexandria (fl. 6th century), Avicenna, the Persian philosopher (d. 1037), and the Arab Abū al Barakāt al-Baghdādi (d. 1164). In the 14th century the French philosopher Jean Buridan developed a new version of the impressed-force theory, calling the quality impressed on the projectile "impetus." Impetus, a permanent quality for Buridan, is measurable by the initial velocity of the projectile and by the quantity of matter contained in it. Buridan employed this concept to suggest an explanation of the everlasting motions of the heavens.

During the 1300s certain Oxford scholars pondered the philosophical problem of how to describe the change that occurs when qualities increase or decrease in intensity and came to consider the kinematic aspects of motion. Dealing with these problems in a purely hypothetical manner without any attempt to describe actual motions in nature or to test their formulas experimentally, they were able to derive the result that in a uniformly accelerated motion, distance increases as the square of the time.

Although medieval science was deeply influenced by Aristotle's philosophy, adherence to his point of view was by no means dogmatic. During the 13th century, theologians at the University of Paris were disturbed by certain statements in Aristotle that seemed to imply limitations of God's powers as well as other statements, such as the eternity of the world, which stood in apparent contradiction to scripture. In 1277 Pope John XXI condemned 219 propositions, many from Aristotle and St. Thomas Aquinas, which had clearly theological consequences. Many of these condemned propositions had scientific implications as well. For example, one of these propositions states, "That the first cause (*i.e.*, God) could not make several worlds." Although it is unlikely that anyone in the Middle Ages actually asserted the existence of many worlds, the condemnation led to the discussion of that possibility, as well as other important problems such as the possibility that the Earth moved.

Criticism of Aristotle

THE SCIENTIFIC REVOLUTION

During the 15th, 16th, and 17th centuries, scientific thought underwent a revolution. A new view of nature emerged, replacing the Greek view that had dominated science for almost 2,000 years. Science became an autonomous discipline, distinct from both philosophy and technology, and it came to be regarded as having utilitarian goals. By the end of this period, it may not be too much to say that science had replaced Christianity as the focal point of European civilization. Out of the ferment of the Renaissance and Reformation there arose a new view of science, bringing about the following transformations: the reeducation of common sense in favour of abstract reasoning; the substitution of a quantitative for a qualitative view of nature; the view of nature as a machine rather than as an organism; the development of an experimental method that sought definite answers to certain limited questions couched in the framework of specific theories; the acceptance of new criteria for explanation, stressing the "how" rather than the "why" that had characterized the Aristotelian search for final causes.

Emergence of a new conception of science

Astronomy. The scientific revolution began in astronomy. Although there had been earlier discussions of the possibility of the Earth's motion, the Polish astronomer Nicolaus Copernicus was the first to propound a comprehensive heliocentric theory, equal in scope and predictive capability to Ptolemy's geocentric system. Motivated by the desire to satisfy Plato's dictum, Copernicus was led to overthrow traditional astronomy because of its alleged violation of the principle of uniform circular motion and its lack of unity and harmony as a system of the world. Relying on virtually the same data as Ptolemy had possessed, Copernicus turned the world inside out, putting the Sun at the centre and setting the Earth into motion around it. Copernicus' theory, published in 1543, possessed a qualitative simplicity that Ptolemaic astronomy appeared to lack. To achieve comparable levels of quantitative precision, however, the new system became just as complex as the old. Perhaps the most revolutionary aspect of Copernican astronomy lay in Copernicus' attitude toward the reality of his theory. In contrast to Platonic instrumentalism, Copernicus asserted that to be satisfactory astronomy must describe the real, physical system of the world.

The reception of Copernican astronomy amounted to victory by infiltration. By the time large-scale opposition to the theory had developed in the church and elsewhere, most of the best professional astronomers had found some aspect or other of the new system indispensable. Copernicus' book *De revolutionibus orbium coelestium, libri VI* (1543; *On the Revolutions of the Celestial Spheres*) became a standard reference for advanced problems in astronomical research, particularly for its mathematical techniques. Thus, it was widely read by mathematical astronomers, in spite of its central cosmological hypothesis, which was widely ignored. In 1551 the German astronomer Erasmus Reinhold published the *Tabulae prutenicae* ("Prutenic Tables"), computed by Copernican methods. The tables were more accurate and more up-to-date than their 13th-century predecessor and became indispensable to both astronomers and astrologers.

Reception of the Copernican view

During the 16th century the Danish astronomer Tycho Brahe, rejecting both the Ptolemaic and Copernican systems, was responsible for major changes in observation, unwittingly providing the data that ultimately decided the argument in favour of the new astronomy. Using larger, stabler, and better calibrated instruments, he observed regularly over extended periods, thereby obtaining a continuity of observations that were accurate for planets to within about one minute of arc—several times better than any previous observation. Several of Tycho's observations

contradicted Aristotle's system: a nova that appeared in 1572 exhibited no parallax (meaning that it lay at a very great distance) and was thus not of the sublunary sphere and therefore contrary to the Aristotelian assertion of the immutability of the heavens; similarly, a succession of comets appeared to be moving freely through a region that was supposed to be filled with solid, crystalline spheres. Tycho devised his own world system—a modification of Heracleides'—to avoid various undesirable implications of the Ptolemaic and Copernican systems.

At the beginning of the 17th century, the German astronomer Johannes Kepler placed the Copernican hypothesis on firm astronomical footing. Converted to the new astronomy as a student and deeply motivated by a neo-Pythagorean desire for finding the mathematical principles of order and harmony according to which God had constructed the world, Kepler spent his life looking for simple mathematical relationships that described planetary motions. His painstaking search for the real order of the universe forced him finally to abandon the Platonic ideal of uniform circular motion in his search for a physical basis for the motions of the heavens.

Kepler's laws

In 1609 Kepler announced two new planetary laws derived from Tycho's data: (1) the planets travel around the Sun in elliptical orbits, one focus of the ellipse being occupied by the Sun; and (2) a planet moves in its orbit in such a manner that a line drawn from the planet to the Sun always sweeps out equal areas in equal times. With these two laws, Kepler abandoned uniform circular motion of the planets on their spheres, thus raising the fundamental physical question of what holds the planets in their orbits. He attempted to provide a physical basis for the planetary motions by means of a force analogous to the magnetic force, the qualitative properties of which had been recently described in England by William Gilbert in his influential treatise, *De Magnete, Magneticisque Corporibus et de Magno Magnete Tellure* (1600; *William Gilbert of Colchester . . . on the Great Magnet of the Earth*). The impending marriage of astronomy and physics had been announced. In 1618 Kepler stated his third law, which was one of many laws concerned with the harmonies of the planetary motions: (3) the square of the period in which a planet orbits the Sun is proportional to the cube of its mean distance from the Sun.

Galileo's observation

A powerful blow was dealt to traditional cosmology by Galileo Galilei, who early in the 17th century used the telescope, a recent invention of Dutch lens grinders, to look toward the heavens. In 1610 Galileo announced observations that contradicted many traditional cosmological assumptions. He observed that the Moon is not a smooth, polished surface, as Aristotle had claimed, but that it is jagged and mountainous. Earthshine on the Moon revealed that the Earth, like the other planets, shines by reflected light. Like the Earth, Jupiter was observed to have satellites; hence, the Earth had been demoted from its unique position. The phases of Venus proved that that planet orbits the Sun, not the Earth.

Physics. *Mechanics.* The battle for Copernicanism was fought in the realm of mechanics as well as astronomy. The Ptolemaic–Aristotelian system stood or fell as a monolith, and it rested on the idea of Earth's fixity at the centre of the cosmos. Removing the Earth from the centre destroyed the doctrine of natural motion and place, and circular motion of the Earth was incompatible with Aristotelian physics.

Galileo's contributions to the science of mechanics were related directly to his defense of Copernicanism. Although in his youth he adhered to the traditional impetus physics, his desire to mathematize in the manner of Archimedes led him to abandon the traditional approach and develop the foundations for a new physics that was both highly mathematizable and directly related to the problems facing the new cosmology. Interested in finding the natural acceleration of falling bodies, he was able to derive the law of free fall (the distance, s, varies as the square of the time, t^2). Combining this result with his rudimentary form of the principle of inertia, he was able to derive the parabolic path of projectile motion. Furthermore, his principle of inertia enabled him to meet the traditional physical objections to the Earth's motion: since a body in motion tends to remain in motion, projectiles and other objects on the terrestrial surface will tend to share the motions of the Earth, which will thus be imperceptible to someone standing on the Earth.

The 17th-century contributions to mechanics of the French philosopher René Descartes, like his contributions to the scientific endeavour as a whole, were more concerned with problems in the foundations of science than with the solution of specific technical problems. He was principally concerned with the conceptions of matter and motion as part of his general program for science—namely, to explain all the phenomena of nature in terms of matter and motion. This program, known as the mechanical philosophy, came to be the dominant theme of 17th-century science.

The rise of mechanical philosophy

Descartes rejected the idea that one piece of matter could act on another through empty space; instead, forces must be propagated by a material substance, the "ether," that fills all space. Although matter tends to move in a straight line in accordance with the principle of inertia, it cannot occupy space already filled by other matter, so the only kind of motion that can actually occur is a vortex in which each particle in a ring moves simultaneously.

According to Descartes, all natural phenomena depend on the collisions of small particles, and so it is of great importance to discover the quantitative laws of impact. This was done by Descartes's disciple, the Dutch physicist Christiaan Huygens, who formulated the laws of conservation of momentum and of kinetic energy (the latter being valid only for elastic collisions).

The work of Sir Isaac Newton represents the culmination of the scientific revolution at the end of the 17th century. His monumental *Philosophiae Naturalis Principia Mathematica* (1687; *Mathematical Principles of Natural Philosophy*) solved the major problems posed by the scientific revolution in mechanics and in cosmology. It provided a physical basis for Kepler's laws, unified celestial and terrestrial physics under one set of laws, and established the problems and methods that dominated much of astronomy and physics for well over a century. By means of the concept of force, Newton was able to synthesize two important components of the scientific revolution, the mechanical philosophy and the mathematization of nature.

The impact of Newton

Newton was able to derive all these striking results from his three laws of motion:

1. Every body continues in its state of rest or of motion in a straight line unless it is compelled to change that state by force impressed on it;
2. The change of motion is proportional to the motive force impressed and is made in the direction of the straight line in which that force is impressed;
3. To every action there is always opposed an equal reaction: or, the mutual actions of two bodies upon each other are always equal.

The second law was put into its modern form $F = ma$ (where a is acceleration) by the Swiss mathematician Leonhard Euler in 1750. In this form, it is clear that the rate of change of velocity is directly proportional to the force acting on a body and inversely proportional to its mass.

In order to apply his laws to astronomy, Newton had to extend the mechanical philosophy beyond the limits set by Descartes. He postulated a gravitational force acting between any two objects in the universe, even though he was unable to explain how this force could be propagated.

The universality of gravitation

By means of his laws of motion and a gravitational force proportional to the inverse square of the distance between the centres of two bodies, Newton could deduce Kepler's laws of planetary motion. Galileo's law of free fall is also consistent with Newton's laws. The same force that causes objects to fall near the surface of the Earth also holds the Moon and planets in their orbits.

Newton's physics led to the conclusion that the shape of the Earth is not precisely spherical but should bulge at the Equator. The confirmation of this prediction by French expeditions in the mid-18th century helped persuade most European scientists to change from Cartesian to Newtonian physics. Newton also used the nonspherical shape of the Earth to explain the precession of the equinoxes,

using the differential action of the Moon and Sun on the equatorial bulge to show how the axis of rotation would change its direction.

Optics. The science of optics in the 17th century expressed the fundamental outlook of the scientific revolution by combining an experimental approach with a quantitative analysis of phenomena. Optics had its origins in Greece, especially in the works of Euclid (*c.* 300 BC), who stated many of the results in geometric optics that the Greeks had discovered, including the law of reflection: the angle of incidence is equal to the angle of reflection. In the 13th century, such men as Roger Bacon, Robert Grosseteste, and John Pecham, relying on the work of the Arab Alhazen (d. 1039), considered numerous optical problems, including the optics of the rainbow. It was Kepler, taking his lead from the writings of these 13th-century opticians, who set the tone for the science in the 17th century. Kepler introduced the point by point analysis of optical problems, tracing rays from each point on the object to a point on the image. Just as the mechanical philosophy was breaking the world into atomic parts, so Kepler approached optics by breaking organic reality into what he considered to be ultimately real units. He developed a geometric theory of lenses, providing the first mathematical account of Galileo's telescope.

Descartes sought to incorporate the phenomena of light into mechanical philosophy by demonstrating that they can be explained entirely in terms of matter and motion. Using mechanical analogies, he was able to derive mathematically many of the known properties of light, including the law of reflection and the newly discovered law of refraction.

Newton's theory of colours

Many of the most important contributions to optics in the 17th century were the work of Newton, especially the theory of colours. Traditional theory considered colours to be the result of the modification of white light. Descartes, for example, thought that colours were the result of the spin of the particles that constitute light. Newton upset the traditional theory of colours by demonstrating in an impressive set of experiments that white light is a mixture out of which separate beams of coloured light can be separated. He associated different degrees of refrangibility with rays of different colours, and in this manner he was able to explain the way prisms produce spectra of colours from white light.

His experimental method was characterized by a quantitative approach, since he always sought measurable variables and a clear distinction between experimental findings and mechanical explanations of those findings. His second important contribution to optics dealt with the interference phenomena that came to be called "Newton's rings." Although the colours of thin films (*e.g.,* oil on water) had been previously observed, no one had attempted to quantify the phenomena in any way. Newton observed quantitative relations between the thickness of the film and the diameters of the rings of colour, a regularity he attempted to explain by his theory of fits of easy transmission and fits of easy reflection. Notwithstanding the fact that he generally conceived of light as being particulate, Newton's theory of fits involves periodicity and vibrations of ether, the hypothetical fluid substance permeating all space (see above).

Huygens was the second great optical thinker of the 17th century. Although he was critical of many of the details of Descartes's system, he wrote in the Cartesian tradition, seeking purely mechanical explanations of phenomena. Huygens regarded light as something of a pulse phenomenon, but he explicitly denied the periodicity of light pulses. He developed the concept of wave front, by means of which he was able to derive the laws of reflection and refraction from his pulse theory and to explain the recently discovered phenomenon of double refraction.

Chemistry. Chemistry had manifold origins, coming from such diverse sources as philosophy, alchemy, metallurgy, and medicine. It emerged as a separate science only with the rise of mechanical philosophy in the 17th century. Aristotle had regarded the four elements earth, water, air, and fire as the ultimate constituents of all things. Transmutable each into the other, all four elements were believed to exist in every substance. Originating in Egypt and the Middle East, alchemy had a double aspect: on the one hand it was a practical endeavour aimed to make gold from baser substances, while on the other it was a cosmological theory based on the correspondence between man and the universe at large. Alchemy contributed to chemistry a long tradition of experience with a wide variety of substances. Paracelsus, a 16th-century Swiss natural philosopher, was a seminal figure in the history of chemistry, putting together in an almost impenetrable combination the Aristotelian theory of matter, alchemical correspondences, mystical forms of knowledge, and chemical therapy in medicine. His influence was widely felt in succeeding generations.

During the first half of the 17th century, there were few established doctrines that chemists generally accepted as a framework. As a result, there was little cumulative growth of chemical knowledge. Chemists tended to build detailed systems, "chemical philosophies," attempting to explain the entire universe in chemical terms. Most chemists accepted the traditional four elements (air, earth, water, fire), or the Paracelsian principles (salt, sulfur, mercury), or both, as the bearers of real qualities in substances; they also exhibited a marked tendency toward the occult.

Effect of mechanical theory on chemistry

The interaction between chemistry and mechanical philosophy altered this situation by providing chemists with a shared language. The mechanical philosophy had been successfully employed in other areas; it seemed consistent with an experimental empiricism and seemed to provide a way to render chemistry respectable by translating it into the terms of the new science. Perhaps the best example of the influence of the mechanical philosophy is the work of Robert Boyle. The thrust of his work was to understand the chemical properties of matter, to provide experimental evidence for the mechanical philosophy, and to demonstrate that all chemical properties can be explained in mechanical terms. He was an excellent laboratory chemist and developed a number of important techniques, especially colour-identification tests.

SCIENCE FROM THE ENLIGHTENMENT TO THE 20TH CENTURY

Seminal contributions to science are those that change the tenor of the questions asked by succeeding generations. The works of Newton formed just such a contribution. The mathematical rigour of the *Principia* and the experimental approach of the *Opticks* became models for scientists of the 18th and 19th centuries. Celestial mechanics developed in the wake of his *Principia,* extending its scope and refining its mathematical methods. The more qualitative, experimental, and hypothetical approach of Newton's *Opticks* influenced the sciences of optics, electricity and magnetism, and chemistry.

Celestial mechanics and astronomy. Eighteenth-century theoretical astronomy in large measure derived both its point of view and its problems from the *Principia.* In this work Newton had provided a physics for the Copernican worldview by, among other things, demonstrating the implications of his gravitational theory for a two-body system consisting of the Sun and a planet. While Newton himself had grave reservations as to the wider scope of his theory, the 18th century witnessed various attempts to extend it to the solution of problems involving three gravitating bodies.

Early in the 18th century the English astronomer Edmond Halley, having noted striking similarities in the comets that had been observed in 1531, 1607, and 1682, argued that they were the periodic appearances every 75 years or so of but a single comet that he predicted would return in 1758. Months before its expected return, the French mathematician Alexis Clairaut employed rather tedious and brute-force mathematics to calculate the effects of the gravitational attraction of Jupiter and Saturn on the otherwise elliptical orbit of Comet Halley. Clairaut was finally able to predict in the fall of 1758 that Comet Halley would reach perihelion in April of 1759, with a leeway of one month. Its actual return, in March, was an early confirmation of the scope and power of the Newtonian theory.

Three-body problem

It was, however, the three-body problem of either two planets and the Sun or the Sun–Earth–Moon system that provided the most persisting and profound test of Newton's theory. This problem, involving more regular members of the solar system (*i.e.*, those describing nearly circular orbits having the same sense of revolution and in nearly the same plane), permitted certain simplifying assumptions and thereby invited more general and elegant mathematical approaches than the comet problem. An illustrious group of 18th-century continental mathematicians (including Clairaut; the Bernoulli family and Leonhard Euler of Switzerland; and Jean Le Rond d'Alembert, Joseph-Louis, comte de Lagrange, and Pierre-Simon, marquis de Laplace, of France) attacked these astronomical problems, as well as related ones in Newtonian mechanics, by developing and applying the calculus of variations as it had been formulated by Leibniz. It is a lovely irony that this continental exploitation of Leibniz' mathematics—which was itself closely akin to Newton's version of calculus, which he called fluxions—was fundamental for the deepening establishment of the Newtonian theory to which Leibniz had objected because it reintroduced, according to Leibniz, occult forces into physics.

In order to attack the lunar theory, which also commanded attention as the most likely astronomical approach to the navigational problem of determining longitude at sea, Clairaut was forced to adopt methods of approximation, having derived general equations that neither he nor anyone else could integrate. Even so, Clairaut was unable to calculate from gravitational theory a value for the progression of the lunar apogee greater than 50 percent of the observed value; therefore, he supposed in 1747 (with Euler) that Newton's inverse-square law was but the first term of a series and, hence, an approximation not valid for distances as small as that between the Earth and Moon. This attempted refinement of Newtonian theory proved to be fruitless, however, and two years later Clairaut was able to obtain, by more detailed and elaborate calculations, the observed value from the simple inverse-square relation.

Certain of the three-body problems, most notably that of the secular acceleration of the Moon, defied early attempts at solution but finally yielded to the increasing power of the calculus of variations in the service of Newtonian theory. Thus it was that Laplace—in his five-volume *Traité de mécanique céleste* (1798–1827; *Celestial Mechanics*)—was able to comprehend the whole solar system as a dynamically stable, Newtonian gravitational system. The secular acceleration of the Moon reappeared as a theoretical problem in the middle of the 19th century, persisting into the 20th century and ultimately requiring that the effects of the tides be recognized in its solution.

Newtonian theory was also employed in much more dramatic discoveries that captivated the imagination of a broad and varied audience. Within 40 years of the discovery of Uranus in 1781 by the German-born British astronomer William Herschel, it was recognized that the planet's motion was somewhat anomalous. In the next 20 years the gravitational attraction of an unobserved planet was suspected to be the cause of Uranus' persisting deviations. In 1845 Urbain-Jean-Joseph Le Verrier of France and John Couch Adams of England independently calculated the position of this unseen body; the visual discovery (at the Berlin Observatory in 1846) of Neptune in just the position predicted constituted an immediately engaging and widely understood confirmation of Newtonian theory. In 1915 the American astronomer Percival Lowell published his prediction of yet another outer planet to account for further perturbations of Uranus not caused by Neptune. Pluto was finally discovered by sophisticated photographic techniques in 1930 and observed visually in 1950.

Discovery of Uranus, Neptune, and Pluto

In the second half of the 19th century, the innermost region of the solar system also received attention. In 1859 Le Verrier calculated the specifications of an intra-mercurial planet to account for a residual advance in the perihelion of Mercury's orbit (38 seconds of arc per century), an effect that was not gravitationally explicable in terms of known bodies. While a number of sightings of this predicted planet were reported between 1859 and 1878—the first of these resulting in Le Verrier's naming the new planet Vulcan—they were not confirmed by observations made either during subsequent solar eclipses or at the times of predicted transits of Vulcan across the Sun.

The theoretical comprehension of Mercury's residual motion involved the first successful departure from Newtonian gravitational theory. This came in the form of Einstein's formulation of the principle of the general of relativity, one consequence of which was a correction term for a body in motion. The correction involves the square of the ratio of the body's velocity to the enormous velocity of light and hence, among the planets, is appreciable only for the fastest: Mercury. Einstein's theory accounted for the residual effect, which by 1915 was calculated to be 43 seconds. This achievement, combined with the 1919 observation of the bending of a ray of light passing near a massive body (another consequence of general relativity theory), constitutes the main experimental verification of that theory.

Mercury's motion and the theory of relativity

Astronomy of the 18th, 19th, and early 20th centuries was not quite so completely Newtonian, however. Herschel's discovery of Uranus, for example, was not directly motivated by gravitational considerations. Nine years earlier, a German astronomer, Johann D. Titius, had announced a purely numerical sequence, subsequently refined by another German astronomer, Johann E. Bode, that related the mean radii of the planetary orbits—a relation entirely outside gravitational theory. The sequence, called Bode's law (or the Bode–Titius law), is given by $0+4=4$, $3+4=7$, $3\times2+4=10$, $3\times4+4=16$, and so on, yielding additional values of 28, 52, and 100. If the measured radius of the Earth's orbit is defined as being 10, then to a very good approximation that of Mercury is 4, Venus is 7, Mars is 15 plus, Jupiter is 52, and Saturn is 95 plus. The fit where it can be made is good and continues since the next number in the sequence is 196 and the measured radius for Uranus' orbit is 191, but no planet had been observed to correspond to the Bode–Titius law value of 28. Powerful computational methods finally verified the existence of the asteroid Ceres and confirmed the law. The Bode–Titius law subsequently provided simplifying assumptions for the calculations of the predicted positions, which led to the observations of Neptune and Pluto, while the novel properties of the asteroids (nearly 500 of which had been discovered by the end of the century) stimulated star charts of the zodiacal regions, provided the means for improved measurements of solar-system distances, and forced astronomers, by their very number and variety, to face the question of the allocation of resources.

Regularities in the structure of the solar system, such as the Bode–Titius law, and the fact that all planets move in the same direction around the Sun suggested that the system might originally have been formed by a simple mechanistic process. Laplace proposed that this process was driven by the cooling of the hot, extended, rotating atmosphere of the primitive Sun. As the atmosphere contracted, it would have to rotate faster (to conserve angular momentum), and when centrifugal force exceeded gravity at the outside, a ring of material would be detached, later to condense into a planet. The process would be repeated several times and might also produce satellites. After Herschel suggested that the nebulas he observed in the sky were condensing to stars, the Laplace theory became known as the "nebular hypothesis." It was the favoured theory of the origin of the solar system throughout the 19th century. During this period the associated idea that the Earth was originally a hot fluid ball that slowly cooled down while forming a solid outer crust dominated geologic speculation.

The nebular hypothesis

Attempts to detect the motion of the Earth caused investigators of the 18th and 19th centuries observational problems that were directly motivated by the Copernican theory. In 1728 the English astronomer James Bradley attributed annual changes that he observed in stellar positions to a slight tilting of the telescope with respect to the true direction of the star's light, a tilting that compensated for the Earth's motion. This effect, which depends also on the ratio of the Earth's velocity to the velocity of light, is the so-called aberration of light.

In 1838 the long-sought "stellar parallax" effect—the apparent motion of nearby stars due to the Earth's annual motion around the Sun—was discovered by the German astronomer Friedrich Wilhelm Bessel. While anticlimactic as a verification of the Copernican hypothesis, the measurement of parallax provided for the first time a direct quantitative estimate of the distances of a few stars.

While attention has been focused on the more positional aspects of astronomy, mention should be made of two other broad areas of investigation that in their 19th-century form derived largely from the work of William Herschel. These areas, dealing with more structural features of the heavens and with the physical character of the stars, developed in large measure with advancements in physics.

Optics. Since they provided the principal basis for subsequent investigations, Newton's optical views were subject to close consideration until well into the 19th century. From his researches into the phenomena of colour, Newton became convinced that dispersion necessarily accompanies refraction and that chromatic aberration (colour distortion) could therefore be eliminated by employing reflectors, rather than refractors, as telescopes. By the mid-18th century Euler and others had theoretical arguments against Newton, and Euler offered the human eye as an example of an achromatic lens system. Although he was virtually alone in this, Euler also rejected Newton's essentially corpuscular theory of the nature of light by explaining optical phenomena in terms of vibrations in a fluid ether. The dominance of Newton's theory throughout the 18th century was due partly to its successful direct application by Newton and his followers and partly to the comprehensiveness of Newton's thought. For example, Bradley's observations found an immediate and natural explanation in terms of the corpuscular theory that also was supported by the accelerating success of Newton's gravitational theory involving discrete particles of matter.

Criticism of Newton's optics

At the turn of the century, Thomas Young, an English physician studying the power of accommodation of the eye (*i.e.,* its focusing power), was led gradually to extensive investigations and discoveries in optics, including the effect of interference. By means of a wave theory of light, Young was able to explain both this effect, which in its most dramatic manifestation results in two rays of light canceling each other to produce darkness, and also the various colour phenomena observed by Newton. The wave theory of light was developed from 1815 onward in a series of brilliant mathematical and experimental memoirs of the physicist Augustin-Jean Fresnel but was countered by adherents of the corpuscular theory, most notably by a group of other French scientists, Pierre-Simon Laplace, Siméon-Denis Poisson, Étienne Malus, and Jean-Baptiste Biot, and most strikingly in connection with Malus' discovery (1808) of the polarization of light by reflection. Following Young's suggestion in 1817, Fresnel was able to render polarization effects comprehensible by means of a wave theory that considered light to be a transverse rather than a longitudinal wave, as the analogy with sound had suggested.

Wave theory of light

The propagation of a transverse wave, the velocity of which through various media and under a variety of conditions was measured terrestrially with increasing accuracy from mid-century onward, seemed to require an ether having the properties of a highly elastic solid (*e.g.,* such as steel), which, however, offered no resistance to the planetary motions. These bizarre properties stimulated a number of mechanical models of the ether, most notably those of the English physicist William Thomson, Lord Kelvin. In order to encompass the aberration of light by means of his wave theory, Fresnel had assumed that the motionless ether freely permeated the opaque Earth and thus remained unaffected by its motions. Furthermore, he derived as a theoretical consequence (verified experimentally in mid-century by Armand-Hippolyte-Louis Fizeau) that the ether was partially, and only partially, dragged along by a moving transparent substance depending on the index of refraction of the substance. However, all subsequent investigators (most notably the American scientists A.A. Michelson and Edward W. Morley, in 1887) failed in their attempts to measure the required ether drift. It was just to escape this difficulty of a necessary but undetected ether drift that George Francis FitzGerald of England and the Dutch theorist Hendrik Antoon Lorentz independently, at the close of the century, postulated the contraction of moving bodies in the direction of their motion through the ether. The Lorentz–FitzGerald contraction involves the square of the ratio of the velocity of the body to the velocity of light and ensures theoretically the experimental undetectability of the ether drift. It was the seeming necessity of arbitrary postulations of this kind that was eliminated by Einstein's formulation of relativity theory.

Electricity and magnetism. Until the end of the 18th century, investigations in electricity and magnetism exhibited more of the hypothetical and spontaneous character of Newton's *Opticks* than the axiomatic and somewhat forbidding tone of his *Principia.* Early in the century, in England Stephen Gray and in France Charles François de Cisternay DuFay studied the direct and induced electrification of various substances by the two kinds of electricity (then called vitreous and resinous and now known as positive and negative), as well as the capability of these substances to conduct the "effluvium" of electricity. By about mid-century, the use of Leyden jars (to collect charges) and the development of large static electricity machines brought the experimental science into the drawing room, while the theoretical aspects were being cast in various forms of the single-fluid theory (by the American Benjamin Franklin and the German-born physicist F.M.U.T.H. Aepinus, among others) and the two-fluid theory.

Theories of electricity

By the end of the 18th century, in England, Joseph Priestley had noted that no electric effect was exhibited inside an electrified hollow metal container and had brilliantly inferred from this similarity that the inverse-square law (of gravity) must hold for electricity as well. In a series of painstaking memoirs, the French physicist Charles-Augustin de Coulomb, using a torsion balance that Henry Cavendish had used in England to measure the gravitational force, demonstrated the inverse-square relation for electrical and magnetic attractions and repulsions. Coulomb went on to apply this law to calculate the surface distribution of the electrical fluid in such a fundamental manner as to provide the basis for the 19th-century extensions by Poisson and Lord Kelvin.

The discovery of galvanic electricity and the development of voltaic electricity opened whole new areas of investigation for the 19th century by providing convenient sources of sustained electrical current. The Danish physicist Hans Christian Ørsted's discovery, in 1820, of the magnetic effect accompanying an electric current led almost immediately to quantitative laws of electromagnetism and electrodynamics. By 1827, André-Marie Ampère had published a series of mathematical and experimental memoirs on his electrodynamic theory that not only rendered electromagnetism comprehensible but also ordinary magnetism, identifying both as the result of electrical currents. Ampère solidly established his electrodynamics by basing it on inverse-square forces (which, however, are directed at right angles to, rather than in, the line connecting the two interacting elements) and by demonstrating that the effects do not violate Newton's third law of motion, notwithstanding their transverse direction.

Relation of electric and magnetic effects

Michael Faraday's discovery in 1831 of electromagnetic induction (the inverse of the effect discovered by Ørsted), his experimental determination of the identity of the various forms of electricity (1833), his discovery of the rotation of the plane of polarization of light by magnetism (1845), in addition to certain findings of other investigators—*e.g.,* the discovery by James Prescott Joule in 1843 (and others) of the mechanical equivalent of heat (the conservation of energy)—all served to emphasize the essential unity of the forces of nature. Within electricity and magnetism attempts at theoretical unification were conceived in terms of either gravitational-type forces acting at a distance, as with Ampère, or, with Faraday, in terms of lines of force and the ambient medium in which they were thought to travel. The German physicists Wilhelm Eduard Weber and Rudolph Kohlrausch, in order to determine the

coefficients in his theory of the former kind, measured the ratio of the electromagnetic and electrostatic units of electrical charge to be equal to the velocity of light.

The Scottish physicist James Clerk Maxwell developed his profound mathematical electromagnetic theory from 1855 onward. He drew his conceptions from Faraday and thus relied fundamentally on the ether required by optical theory, while using ingenious mechanical models. One consequence of Maxwell's mature theory was that an electromagnetic wave must be propagated through the ether with a velocity equal to the ratio of the electromagnetic to electrostatic units. Combined with the earlier results of Weber and Kohlrausch, this result implied that light is an electromagnetic phenomenon. Moreover, it suggested that electromagnetic waves of wavelengths other than the narrow band corresponding to visible light should exist in nature or could be artificially generated.

The nature of light

Maxwell's theory received direct verification in 1886, when Heinrich Hertz of Germany detected the predicted electromagnetic waves. Their use in long-distance communication—"radio"—followed within two decades, and gradually physicists became acquainted with the entire electromagnetic spectrum.

Chemistry. Eighteenth-century chemistry was derived from and remained involved with questions of mechanics, light, and heat as well as with iatrochemistry and notions of medical therapy and the interaction between substances and the formation of new substances. Chemistry took many of its problems and much of its viewpoint from the *Opticks* and especially the "Queries" with which that work ends. Newton's suggestion of a hierarchy of clusters of unalterable particles formed by virtue of the specific attractions of its component particles led directly to comparative studies of interactions and thus to the tables of affinities of the physician Herman Boerhaave and others early in the century. This work culminated at the end of the century in the Swede Torbern Bergman's table that gave quantitative values of the affinity of substances both for reactions when "dry" and when in solution and that considered double as well as simple affinities.

Seventeenth-century investigations of "airs" or gases, combustion and calcination, and the nature and role of fire were incorporated by the chemists Johann Joachim Becher and Georg Ernst Stahl of Sweden into a theory of phlogiston. According to this theory, which was most influential after the middle of the 18th century, the fiery principle, phlogiston, was released into the air in the processes of combustion, calcination, and respiration. The theory explained that air was simply the receptacle for phlogiston, and any combustible or calcinable substance contained phlogiston as a principle or element and thus could not itself be elemental. Iron, in rusting, was considered to lose its compound nature and to assume its elemental state as the calx of iron by yielding its phlogiston into the ambient air.

Theory of phlogiston

Investigations that isolated and identified various gases in the second half of the 18th century, most notably the English chemist Joseph Black's quantitative manipulations of "fixed air" (carbon dioxide) and Joseph Priestley's discovery of "dephlogisticated air" (oxygen), were instrumental for the French chemist Antoine Lavoisier's formulation of his own oxygen theory of combustion and rejection of the phlogiston theory (*i.e.,* he explained combustion not as the result of the liberation of phlogiston, but rather as the result of the combination of the burning substance with oxygen). This transformation coupled with the reform in nomenclature at the end of the century (due to Lavoisier and others)—a reform that reflected the new conceptions of chemical elements, compounds, and processes—constituted the revolution in chemistry.

Lavoisier's revolution

Very early in the 19th century, another study of gases, this time in the form of a persisting Newtonian approach to certain meteorological problems by the British chemist John Dalton, led to the enunciation of a chemical atomic theory. From this theory, which was demonstrated to agree with the law of definite proportions and from which the law of multiple proportions was derived, Dalton was able to calculate definite atomic weights by assuming the simplest possible ratio for the numbers of combining atoms. For example, knowing from experiment that the ratio of the combining weights of hydrogen to oxygen in the formation of water is 1 to 8 and by assuming that one atom of hydrogen combined with one atom of oxygen, Dalton affirmed that the atomic weight of oxygen was eight, based on hydrogen as one. At the same time, however, in France, Joseph-Louis Gay-Lussac, from his volumetric investigations of combining gases, determined that two volumes of hydrogen combined with one of oxygen to produce water. While this suggested H_2O rather than Dalton's HO as the formula for water, with the result that the atomic weight of oxygen becomes 16, it did involve certain inconsistencies with Dalton's theory.

As early as 1811 the Italian physicist Amedeo Avogadro was able to reconcile Dalton's atomic theory with Gay-Lussac's volumetric law by postulating that Dalton's atoms were indeed compound atoms, or polyatomic. For a number of reasons, one of which involved the recent successes of electrochemistry, Avogadro's hypothesis was not accepted until it was reintroduced by the Italian chemist Stanislao Cannizzaro half a century later. From the turn of the century, the English scientist Humphry Davy and many others had employed the strong electric currents of voltaic piles for the analysis of compound substances and the discovery of new elements. From these results, it appeared obvious that chemical forces were essentially electrical in nature and that two hydrogen atoms, for example, having the same electrical charge, would repel each other and could not join to form the polyatomic molecule required by Avogadro's hypothesis. Until the development of a quantum-mechanical theory of the chemical bond, beginning in the 1920s, bonding was described by empirical "valence" rules but could not be satisfactorily explained in terms of purely electrical forces.

Avogadro's hypothesis

Between the presentation of Avogadro's hypothesis in 1811 and its general acceptance soon after 1860, several experimental techniques and theoretical laws were used by various investigators to yield different but self-consistent schemes of chemical formulas and atomic weights. After its acceptance, these schemes became unified. Within a few years of the development of another powerful technique, spectrum analysis, by the German physicists Gustav Kirchhoff and Robert Bunsen in 1859, the number of chemical elements whose atomic weights and other properties were known had approximately doubled since the time of Avogadro's announcement. By relying fundamentally but not slavishly upon the determined atomic weight values and by using his chemical insight and intuition, the Russian chemist Dmitry Ivanovich Mendeleyev provided a classification scheme that ordered much of this burgeoning information and was a culmination of earlier attempts to represent the periodic repetition of certain chemical and physical properties of the elements.

The significance of the atomic weights themselves remained unclear. In 1815 William Prout, an English chemist, had proposed that they might all be integer multiples of the weight of the hydrogen atom, implying that the other elements are simply compounds of hydrogen. More accurate determinations, however, showed that the atomic weights are significantly different from integers. They are not, of course, the actual weights of individual atoms, but by 1870 it was possible to estimate those weights (or rather masses) in grams by the kinetic theory of gases and other methods. Thus one could at least say that the atomic weight of an element is proportional to the mass of an atom of that element. (M.J.O./J.B.Sp./S.G.B.)

DEVELOPMENTS AND TRENDS OF THE 20TH CENTURY

Astronomy. Some of the most spectacular advances in modern astronomy have come from research on the large-scale structure and development of the universe. This research goes back to William Herschel's observations of nebulas at the end of the 18th century. Some astronomers considered them to be "island universes"—huge stellar systems outside of and comparable to the Milky Way Galaxy, to which the solar system belongs. Others, following Herschel's own speculations, thought of them simply as gaseous clouds—relatively small patches of diffuse matter within the Milky Way Galaxy, which might be in the

process of developing into stars and planetary systems, as described in Laplace's nebular hypothesis.

In 1912 Vesto Melvin Slipher began at the Lowell Observatory in Arizona an extensive program to measure the velocities of nebulas, using the Doppler shift of their spectral lines. (Doppler shift is the observed change in wavelength of the radiation from a source that results from the relative motion of the latter along the line of sight.) By 1925 he had studied about 40 nebulas, most of which were found to be moving away from the Earth according to the red shift (displacement toward longer wavelengths) of their spectra.

Although the nebulas were apparently so far away that their distances could not be measured directly by the stellar parallax method, an indirect approach was developed on the basis of a discovery made in 1908 by Henrietta Swan Leavitt at the Harvard College Observatory. Leavitt studied the magnitudes (apparent brightnesses) of a large number of variable stars, including the type known as Cepheid variables. Some of them were close enough to have measurable parallaxes so that their distances and thus their intrinsic brightnesses could be determined. She found a correlation between brightness and period of variation. Assuming that the same correlation holds for all stars of this kind, their observed magnitudes and periods could be used to estimate their distances.

In 1923 the American astronomer Edwin P. Hubble identified a Cepheid variable in the so-called Andromeda Nebula. Using Leavitt's period–brightness correlation, Hubble estimated its distance to be approximately 900,000 light-years. Since this was much greater than the size of the Milky Way system, it appeared that the Andromeda Nebula must be another galaxy (island universe) outside of our own.

In 1929 Hubble combined Slipher's measurements of the velocities of nebulas with further estimates of their distances and found that on the average such objects are moving away from the Earth with a velocity proportional to their distance. Hubble's velocity–distance relation suggested that the universe of galactic nebulas is expanding, starting from an initial state about 2,000,000,000 years ago in which all matter was contained in a fairly small volume. Revisions of the distance scale in the 1950s and later increased the "Hubble age" of the universe to more than 10,000,000,000 years.

The expanding universe

Calculations by Aleksandr A. Friedmann in the Soviet Union, Willem de Sitter in The Netherlands, and Georges Lemaître in Belgium, based on Einstein's general theory of relativity, showed that the expanding universe could be explained in terms of the evolution of space itself. According to Einstein's theory, space is described by the non-Euclidean geometry proposed in 1854 by the German mathematician G.F. Bernhard Riemann. Its departure from Euclidean space is measured by a "curvature" that depends on the density of matter. The universe may be finite, though unbounded, like the surface of a sphere. Thus the expansion of the universe refers not merely to the motion of extragalactic stellar systems within space but also to the expansion of the space itself.

The beginning of the expanding universe was linked to the formation of the chemical elements in a theory developed in the 1940s by the physicist George Gamow, a former student of Friedmann who had emigrated to the United States. Gamow proposed that the universe began in a state of extremely high temperature and density and exploded outward—the so-called big bang. Matter was originally in the form of neutrons, which quickly decayed into protons and electrons; these then combined to form hydrogen and heavier elements.

The big-bang theory

Gamow's students Ralph Alpher and Robert Herman estimated in 1948 that the radiation left over from the big bang should by now have cooled down to a temperature just a few degrees above absolute zero (0 K, or −459° F). In 1965 the predicted cosmic background radiation was discovered by Arno A. Penzias and Robert W. Wilson of the Bell Telephone Laboratories as part of an effort to build sensitive microwave-receiving stations for satellite communication. Their finding provided unexpected evidence for the idea that the universe was in a state of very high temperature and density sometime between 10,000,000,000 and 20,000,000,000 years ago.

Evolution of stars and formation of chemical elements. Just as the development of cosmology relied heavily on ideas from physics, especially Einstein's general theory of relativity, so did theories of stellar structure and evolution depend on discoveries in atomic physics. These theories also offered a fundamental basis for chemistry by showing how the elements could have been synthesized in stars.

The idea that stars are formed by the condensation of gaseous clouds was part of the 19th-century nebular hypothesis (see above). The gravitational energy released by this condensation could be transformed into heat, but calculations by Hermann von Helmholtz and Lord Kelvin indicated that this process would provide energy to keep the Sun shining for only about 20,000,000 years. Evidence from radiometric dating, starting with the work of the British physicist Ernest Rutherford in 1905, showed that the Earth is probably several billion years old. Astrophysicists were perplexed: what source of energy has kept the Sun shining for such a long time?

In 1925 Cecilia Payne, a graduate student from Britain at Harvard College Observatory, analyzed the spectra of stars using statistical atomic theories that related them to temperature, density, and composition. She found that hydrogen and helium are the most abundant elements in stars, though this conclusion was not generally accepted until it was confirmed four years later by the noted American astronomer Henry Norris Russell. By this time Prout's hypothesis that all the elements are compounds of hydrogen had been revived by physicists in a somewhat more elaborate form. The deviation of atomic weights from exact integer values (expressed as multiples of hydrogen) could be explained partly by the fact that some elements are mixtures of isotopes with different atomic weights and partly by Einstein's relation between mass and energy (taking account of the binding energy of the forces that hold together the atomic nucleus). The German physicist Werner Heisenberg proposed in 1932 that, whereas the hydrogen nucleus consists of just one proton, all heavier nuclei contain protons and neutrons. Since a proton can be changed into a neutron by fusing it with an electron, this meant that all the elements could be built up from protons and electrons—*i.e.*, from hydrogen atoms.

In 1938 the German-born physicist Hans Bethe proposed the first satisfactory theory of stellar energy generation based on the fusion of protons to form helium and heavier elements. He showed that once elements as heavy as carbon had been formed, a cycle of nuclear reactions could produce even heavier elements. Fusion of hydrogen into heavier elements would also provide enough energy to account for the Sun's energy generation over a period of billions of years. Although Bethe's theory, as extended by Fred Hoyle, Edwin E. Salpeter, and William A. Fowler, is the best one available, there is still some doubt about its accuracy because the neutrinos supposedly produced by the fusion reactions have not been observed in the amounts predicted.

Synthesis and dissemination of heavy elements

According to the theory of stellar evolution developed by the Indian-born American astrophysicist Subrahmanyan Chandrasekhar and others, a star will become unstable after it has converted most of its hydrogen to helium and may go through stages of rapid expansion and contraction. If the star is much more massive than the Sun, it will explode violently, giving rise to a supernova. The explosion will synthesize heavier elements and spread them throughout the surrounding interstellar medium, where they provide the raw material for the formation of new stars and eventually of planets and living organisms.

After a supernova explosion, the remaining core of the star may collapse further under its own gravitational attraction to form a dense star composed mainly of neutrons. This so-called neutron star, predicted theoretically in the 1930s by the astronomers Walter Baade and Fritz Zwicky, is apparently the same as the pulsar (a source of rapid, very regular pulses of radio waves), discovered in 1967 by Jocelyn Bell of the British radio astronomy group under Antony Hewish at Cambridge University.

More massive stars may undergo a further stage of evolu-

The black-hole phenomenon

tion beyond the neutron star: they may collapse to a black hole, in which the gravitational force is so strong that even light cannot escape. The black hole as a singularity in an idealized space-time universe was predicted from the general relativity theory by the German astronomer Karl Schwarzschild in 1916. Its role in stellar evolution was later described by the American physicists J. Robert Oppenheimer and John Wheeler. During the 1980s, possible black holes were thought to have been located in X-ray sources and at the centre of certain galaxies (see below *Study of the stars*).

Solar-system astronomy. This area of investigation, which lay relatively dormant through the first half of the 20th century, was revived in the 1960s under the stimulus of the Soviet and American space programs. Missions to the Moon and planets yielded a wealth of complex information which has yet to be completely assimilated by scientists. A single example of the resulting change in ideas about the history of the solar system will have to suffice here. Before the first manned lunar landing in 1969, there were three competing hypotheses about the origin of the Moon: (1) formation in its present orbit simultaneously with the Earth, as described in the nebular hypothesis; (2) formation elsewhere and subsequent capture by the Earth; and (3) ejection from the Earth by fission (popularly known theory that the Moon emanated from what is now the Pacific Ocean Basin). Following the analysis of lunar samples and theoretical criticism of these hypotheses, lunar scientists came to the conclusion that none of them was satisfactory. Photographs of the surface of Mercury taken by the U.S. Mariner 10 spacecraft in 1974, however, showed that it is heavily cratered like the Moon's surface. This finding, together with theoretical calculations by V.S. Safronov of the Soviet Union and George W. Wetherill of the United States on the formation of planets by accumulation (accretion or aggregation) of smaller solid bodies, suggested that the Earth was also probably subject to heavy bombardment soon after its formation. In line with this, a theory proposed by the American astronomers William K. Hartmann and A.G.W. Cameron has become the most popular. According to their theory, the Earth was struck by a Mars-sized object, and the force of the impact vaporized the outer parts of both bodies. The vapour thus produced remained in orbit around the Earth and eventually condensed to form the Moon. Like the hypothesis proposed by Luis Alvarez that attributes the extinction of the dinosaurs to an asteroid impact, the Hartmann–Cameron theory seemed so bizarre that it could not have been taken seriously until compelling evidence became available.

The Hartmann–Cameron theory

Physics. During the years 1896–1932 the foundations of physics changed so radically that many observers describe this period as a scientific revolution comparable in depth, if not in scope, to the one that took place during the 16th and 17th centuries. The 20th-century revolution changed many of the ideas about space, time, mass, energy, atoms, light, force, determinism, and causality that had apparently been firmly established by Newtonian physics during the 18th and 19th centuries. Moreover, according to some interpretations, the new theories demolished the basic metaphysical assumption of earlier science that the entire physical world has a real existence and objective properties independent of human observation.

The 20th-century scientific revolution

Closer examination of 19th-century physics shows that Newtonian ideas were already being undermined in many areas and that the program of mechanical explanation was openly challenged by several influential physicists toward the end of the century. Yet, there was no agreement as to what the foundations of a new physics might be. Modern textbook writers and popularizers often try to identify specific paradoxes or puzzling experimental results—*e.g.*, the failure to detect the Earth's absolute motion in the Michelson–Morley experiment—as anomalies that led physicists to propose new fundamental theories such as relativity. Historians of science have shown, however, that most of these anomalies did not directly cause the introduction of the theories that later resolved them. As with Copernicus' introduction of heliocentric astronomy, the motivation seems to have been a desire to satisfy aesthetic principles of theory structure rooted in earlier views of the world rather than a need to account for the latest experiment or calculation.

Radioactivity and the transmutation of elements. The discovery of radioactivity by the French physicist Henri Becquerel in 1896 is generally taken to mark the beginning of 20th-century physics. The successful isolation of radium and other intensely radioactive substances by Marie and Pierre Curie focused the attention of scientists and the public on this remarkable phenomenon and promoted a wide range of experiments.

Rutherford's investigations of radioactivity

Ernest Rutherford soon took the lead in studying the nature of radioactivity. He found that there are two distinct kinds of radiation emitted in radioactivity called alpha and beta rays. The alpha rays proved to be positively charged particles identical to ionized helium atoms. Beta rays are much less massive negatively charged particles; they were shown to be the same as the electrons discovered by J.J. Thomson in cathode rays in 1897. A third kind of ray, designated gamma, consists of high-frequency electromagnetic radiation.

Rutherford proposed that radioactivity involves a transmutation of one element into another. This proposal called into question one of the basic assumptions of 19th-century chemistry: that the elements consist of qualitatively different substances—92 of them by the end of the century. It implied a return to the ideas of Prout and the ancient atomists—namely, that everything in the world is composed of only one or a few basic substances.

Transmutation, according to Rutherford and his colleagues, was governed by certain empirical rules. For example, in alpha decay the atomic number of the "daughter" element is two less than that of the "mother" element, and its atomic weight is four less; this seems consistent with the fact that the alpha ray, identified as helium, has atomic number 2 and atomic weight 4, so that total atomic number and total atomic weight are conserved in the decay reaction.

Using these rules, Rutherford and his colleagues could determine the atomic numbers and atomic weights of many substances formed by radioactive decay, even though the substances decayed so quickly into others that these properties could not be measured directly. The atomic number of an element determines its place in Mendeleyev's periodic table (and thus its chemical properties; see above). It was found that substances of different atomic weight could have the same atomic number; such substances were called isotopes of an element.

Although the products of radioactive decay are determined by simple rules, the decay process itself seems to occur at random. All one can say is that there is a certain probability that an atom of a radioactive substance will decay during a certain time interval, or, equivalently, that half of the atoms of the sample will have decayed after a certain time—*i.e.*, the half life of the material.

The nucleus. At the University of Manchester (England), Rutherford led a group that rapidly developed new ideas about atomic structure. On the basis of an experiment conducted by Hans Geiger and Ernest Marsden in which alpha particles were scattered by a thin film of metal, Rutherford proposed a nuclear model of the atom (1911). In this model, the atom consists mostly of empty space, with a tiny, positively charged nucleus that contains most of the mass, surrounded by one or more negatively charged electrons. Henry G.J. Moseley, an English physicist, showed by an analysis of X-ray spectra that the electric charge on the nucleus is simply proportional to the atomic number of the element.

Discovery of the neutron

During the 1920s physicists thought that the nucleus was composed of two particles: the proton (the positively charged nucleus of hydrogen) and the electron. In 1932 the English physicist James Chadwick discovered the neutron, a particle with about the same mass as the proton but no electric charge. Since there were technical difficulties with the proton–electron model of the nucleus, physicists were willing to accept Heisenberg's hypothesis that it consists instead of protons and neutrons. The atomic number is then simply the number of protons in the nucleus, while the mass number, the integer closest to the atomic weight,

is equal to the total number of neutrons and protons. As mentioned above, this simple model of nuclear structure provided the basis for Hans Bethe's theory of the formation of elements from hydrogen in stars.

Splitting of the uranium atom

In 1938 the German physicists Otto Hahn and Fritz Strassmann found that, when uranium is bombarded by neutrons, lighter elements such as barium and krypton are produced. This phenomenon was interpreted by Lise Meitner and her nephew Otto Frisch as a breakup, or fission, of the uranium nucleus into smaller nuclei. Other physicists soon realized that since fission produces more neutrons, a chain reaction could result in a powerful explosion. World War II was about to begin, and physicists who had emigrated from Germany, Italy, and Hungary to the United States and Great Britain feared that Germany might develop an atomic bomb that could determine the outcome of the war. They persuaded the U.S. and British governments to undertake a major project to develop such a weapon first. The U.S. Manhattan Project did eventually produce atomic bombs based on the fission of uranium or of plutonium, a new artificially created element, and these were used against Japan in August 1945. Later, an even more powerful bomb based on the fusion of hydrogen atoms was developed and tested by both the United States and the Soviet Union. Thus nuclear physics began to play a major role in world history.

Einstein's 1905 trilogy. In a few months during the years 1665–66, Newton discovered the composite nature of light, analyzed the action of gravity, and invented the mathematical technique now known as calculus—or so he recalled in his old age. The only person who has ever matched Newton's amazing burst of scientific creativity—three revolutionary discoveries within a year—was Albert Einstein, who in 1905 published the special theory of relativity, the quantum theory of radiation, and a theory of Brownian movement that led directly to the final acceptance of the atomic structure of matter.

Relativity theory has already been mentioned several times in this article, an indication of its close connection with several areas of physical science. There is no room here to discuss the subtle line of reasoning that Einstein followed in arriving at his amazing conclusions; a brief summary of his starting point and some of the consequences will have to suffice.

In his 1905 paper on the electrodynamics of moving bodies, Einstein called attention to an apparent inconsistency in the usual presentation of Maxwell's electromagnetic theory as applied to the reciprocal action of a magnet and a conductor. The equations are different depending on which is "at rest" and which is "moving," yet the results must be the same. Einstein located the difficulty in the assumption that absolute space exists; he postulated instead

Relativity of space and time

that the laws of nature are the same for observers in any inertial frame of reference and that the speed of light is the same for all such observers.

From these postulates Einstein inferred: (1) an observer in one frame would find from his own measurements that lengths of objects in another frame are contracted by an amount given by the Lorentz–FitzGerald formula; (2) each observer would find that clocks in the other frame run more slowly; (3) there is no absolute time—events that are simultaneous in one frame of reference may not be so in another; and (4) the observable mass of any object increases as it goes faster.

Closely connected with the mass-increase effect is Einstein's famous formula $E = mc^2$: mass and energy are no longer conserved but can be interconverted. The explosive power of the atomic and hydrogen bombs derives from the conversion of mass to energy.

In a paper on the creation and conversion of light (usually called the "photoelectric effect paper"), published earlier in 1905, Einstein proposed the hypothesis that electromagnetic radiation consists of discrete energy quanta that can be absorbed or emitted only as a whole. Although this hypothesis would not replace the wave theory of light, which gives a perfectly satisfactory description of the phenomena of diffraction, reflection, refraction, and dispersion, it would supplement it by also ascribing particle properties to light.

Quantum theory of radiation

Until recently the invention of the quantum theory of radiation was generally credited to another German physicist, Max Planck, who in 1900 discussed the statistical distribution of radiation energy in connection with the theory of blackbody radiation (see below *The methodology of physics*). Although Planck did propose the basic hypothesis that the energy of a quantum of radiation is proportional to its frequency of vibration, it is not clear whether he used this hypothesis merely for mathematical convenience or intended it to have a broader physical significance. In any case, he did not explicitly advocate a particle theory of light before 1905. Historians of physics still disagree on whether Planck or Einstein should be considered the originator of the quantum theory.

Einstein's paper on Brownian movement seems less revolutionary than the other 1905 papers because most modern readers assume that the atomic structure of matter was well established at that time. Such was not the case, however. In spite of the development of the chemical atomic theory and of the kinetic theory of gases in the 19th century, which allowed quantitative estimates of such atomic properties as mass and diameter, it was still fashionable in 1900 to question the reality of atoms. This skepticism, which does not seem to have been particularly helpful to the progress of science, was promoted by the empiricist, or "positivist," philosophy advocated by Auguste Comte, Ernst Mach, Wilhelm Ostwald, Pierre Duhem, Henri Poincaré, and others. It was the French physicist Jean Perrin who, using Einstein's theory of Brownian movement, finally convinced the scientific community to accept the atom as a valid scientific concept.

Quantum mechanics. The Danish physicist Niels Bohr pioneered the use of the quantum hypothesis in developing a successful theory of atomic structure. Adopting Rutherford's nuclear model, he proposed in 1913 that the atom is like a miniature solar system, with the electrons moving in orbits around the nucleus just as the planets move around the Sun. Although the electrical attraction between the electrons and nucleus is mathematically similar to the gravitational attraction between the planets and the Sun, the quantum hypothesis is needed to restrict the electrons to certain orbits and to forbid them from radiating energy except when jumping from one orbit to another.

Bohr's model of atomic structure

Bohr's model provided a good description of the spectra and other properties of atoms containing only one electron—neutral hydrogen and singly ionized helium—but could not be satisfactorily extended to multi-electron atoms or molecules. It relied on an inconsistent mixture of old and new physical principles, hinting but not clearly specifying how a more adequate general theory might be constructed.

The nature of light was still puzzling to those who demanded that it should behave either like waves or like particles. Two experiments performed by American physicists seemed to favour the particle theory: Robert A. Millikan's confirmation of the quantum theory of the photoelectric effect proposed by Einstein; and Arthur H. Compton's experimental demonstration that X rays behave like particles when they collide with electrons. The findings of these experiments had to be considered along with the unquestioned fact that electromagnetic radiation also exhibits wave properties such as interference and diffraction.

Louis de Broglie, a French physicist, proposed a way out of the dilemma: accept the wave–particle dualism as a description not only of light but also of electrons and other entities previously assumed to be particles. In 1926 the Austrian physicist Erwin Schrödinger constructed a mathematical "wave mechanics" based on this proposal. His theory tells how to write down an equation for the wave function of any physical system in terms of the masses and charges of its components. From the wave function, one may compute the energy levels and other observable properties of the system.

Schrödinger's so-called wave mechanics

Schrödinger's equation, the most convenient form of a more general theory called quantum mechanics to which the German physicists Werner Heisenberg and Max Born also contributed, was brilliantly successful. Not only did it yield the properties of the hydrogen atom but it also allowed the use of simple approximating methods for more

complicated systems even though the equation could not be solved exactly. The application of quantum mechanics to the properties of atoms, molecules, and metals occupied physicists for the next several decades.

The founders of quantum mechanics did not agree on the philosophical significance of the new theory. Born proposed that the wave function determines only the probability distribution of the electron's position or path; it does not have a well-defined instantaneous position and velocity. Heisenberg made this view explicit in his indeterminacy principle: the more accurately one determines the position, the less accurately the velocity is fixed; the converse is also true. Heisenberg's principle is often called the uncertainty principle, but this is somewhat misleading. It tends to suggest incorrectly that the electron really has a definite position and velocity and that they simply have not been determined.

Heisenberg's indeterminacy principle

Einstein objected to the randomness implied by quantum mechanics in his famous statement that God "does not play dice." He also was disturbed by the apparent denial of the objective reality of the atomic world: Somehow the electron's position or velocity comes into existence only when it is measured. Niels Bohr expressed this aspect of the quantum worldview in his complementarity principle, building on de Broglie's resolution of the wave–particle dichotomy: A system can have such properties as wave or particle behaviour that would be considered incompatible in Newtonian physics but that are actually complementary; light exhibits either wave behaviour or particle behaviour, depending on whether one chooses to measure the one property or the other. To say that it is really one or the other, or to say that the electron really has both a definite position and momentum at the same time, is to go beyond the limits of science.

Bohr's viewpoint, which became known as the Copenhagen Interpretation of quantum mechanics, was that reality can be ascribed only to a measurement. Einstein argued that the physical world must have real properties whether or not one measures them; he and Schrödinger published a number of thought experiments designed to show that things can exist beyond what is described by quantum mechanics. During the 1970s and 1980s, advanced technology made it possible to actually perform some of these experiments, and quantum mechanics was vindicated in every case.

Chemistry. The long-standing problem of the nature of the force that holds atoms together in molecules was finally solved by the application of quantum mechanics. Although it is often stated that chemistry has been "reduced to physics" in this way, it should be pointed out that one of the most important postulates of quantum mechanics was introduced primarily for the purpose of explaining chemical facts and did not originally have any other physical justification. This was the so-called exclusion principle put forth by the Austrian physicist Wolfgang Pauli, which forbids more than one electron occupying a given quantum state in an atom. The state of an electron includes its spin, a property introduced by the Dutch-born American physicists George E. Uhlenbeck and Samuel A. Goudsmit. Using that principle and the assumption that the quantum states in a multi-electron atom are essentially the same as those in the hydrogen atom, one can postulate a series of "shells" of electrons and explain the chemical valence of an element in terms of the loss, gain, or sharing of electrons in the outer shell.

Some of the outstanding problems to be solved by quantum chemistry were: (1) The "saturation" of chemical forces. If attractive forces hold atoms together to form molecules, why is there a limit on how many atoms can stick together (generally only two of the same kind)? (2) Stereochemistry—the three-dimensional structure of molecules, in particular the spatial directionality of bonds as in the tetrahedral carbon atom. (3) Bond length—*i.e.*, there seems to be a well-defined equilibrium distance between atoms in a molecule that can be determined accurately by experiment. (4) Why some atoms (*e.g.*, helium) normally form no bonds with other atoms, while others form one or more. (These are the empirical rules of valence.)

Soon after J.J. Thomson's discovery of the electron in 1897, there were several attempts to develop theories of chemical bonds based on electrons. The most successful was that proposed in the United States by G.N. Lewis in 1916 and Irving Langmuir in 1919. They emphasized shared pairs of electrons and treated the atom as a static arrangement of charges. While the Lewis–Langmuir model as a whole was inconsistent with quantum theory, several of its specific features continued to be useful.

The key to the nature of the chemical bond was found to be the quantum-mechanical exchange effect, first described by Heisenberg in 1926–27. Resonance is related to the requirement that the wave function for two or more identical particles must have definite symmetry properties with respect to the coordinates of those particles—it must have plus or minus the same value (symmetric or antisymmetric, respectively) when those particles are interchanged. Particles such as electrons and protons, according to a hypothesis proposed by Enrico Fermi and P.A.M. Dirac, must have antisymmetric wave functions. Exchange may be imagined as a continual jumping back and forth or interchange of the electrons between two possible states. In 1927 the German physicists Walter Heitler and Fritz London used this idea to obtain an approximate wave function for two interacting hydrogen atoms. They found that with an antisymmetric wave function (including spin) there is an attractive force, while with a symmetric one there is a repulsive force. Thus two hydrogen atoms can form a molecule if their electron spins are opposite, but not if they are the same.

The quantum-mechanical exchange effect

The Heitler–London approach to the theory of chemical bonds was rapidly developed by John C. Slater and Linus C. Pauling in the United States. Slater proposed a simple general method for constructing multiple-electron wave functions that would automatically satisfy the Pauli exclusion principle. Pauling introduced a valence-bond method, picking out one electron in each of the two combining atoms and constructing a wave function representing a paired-electron bond between them. Pauling and Slater were able to explain the tetrahedral carbon structure in terms of a particular mixture of wave functions that has a lower energy than the original wave functions, so that the molecule tends to go into that state.

Pauling's valence-bond method

About the same time another American scientist, Robert S. Mulliken, was developing an alternative theory of molecular structure based on what he called molecular orbitals. (The idea had been used under a different name by John E. Lennard-Jones of England in 1929 and by Erich Hückel of Germany in 1931.) Here, the electron is not considered to be localized in a particular atom or two-atom bond, but rather it is treated as occupying a quantum state (an "orbital") that is spread over the entire molecule (see below *Studies of molecular structure*).

In treating the benzene molecule by the valence-bond method in 1933, Pauling and George W. Wheland constructed a wave function that was a linear combination of five possible structures—*i.e.*, five possible arrangements of double and single bonds. Two of them are the structures that had been proposed by the German chemist August Kekulé (later Kekule von Stradonitz) in 1865, with alternating single and double bonds between adjacent carbon atoms in the six-carbon ring. The other three (now called Dewar structures for the British chemist and physicist James Dewar, though they were first suggested by H. Wichelhaus in 1869) have one longer bond going across the ring. Pauling and Dewar described their model as involving resonance between the five structures. According to quantum mechanics, this does not mean that the molecule is sometimes "really" in one state and at other times in another, but rather that it is always in a composite state.

The valence-bond method, with its emphasis on resonance between different structures as a means of analyzing aromatic molecules, dominated quantum chemistry during the 1930s. The method was comprehensively presented and applied in Pauling's classic treatise *The Nature of the Chemical Bond* (1939), the most important work on theoretical chemistry in the 20th century. One reason for its popularity was that ideas similar to resonance had

been developed by organic chemists, notably F.G. Arndt in Germany and Christopher K. Ingold in England, independently of quantum theory during the late 1920s.

Increasing preference for the molecular-orbital method

After World War II there was a strong movement away from the valence-bond method toward the molecular-orbital method, led by Mulliken in the United States and by Charles Coulson, Lennard-Jones, H.C. Longuet-Higgins, and Michael J.S. Dewar in England. The advocates of the molecular-orbital method argued that their approach was simpler and easier to apply to complicated molecules, since it allowed one to visualize a definite charge distribution for each electron. (S.G.B.)

Physics

Physics is the basic physical science. Until rather recent times the terms physics and natural philosophy were used interchangeably for the science whose aim is the discovery and formulation of the fundamental laws of nature. As the modern sciences developed and became increasingly specialized, physics came to denote that part of physical science not included in astronomy, chemistry, geology, and engineering. Physics plays an important role in all the natural sciences, however, and all such fields have branches in which physical laws and measurements receive special emphasis, bearing such names as astrophysics, geophysics, biophysics, and even psychophysics. Physics can, at base, be defined as the science of matter, motion, and energy. Its laws are typically expressed with economy and precision in the language of mathematics.

Both experiment, the observation of phenomena under conditions that are controlled as precisely as possible, and theory, the formulation of a unified conceptual framework, play essential and complementary roles in the advancement of physics. Physical experiments result in measurements, which are compared with the outcome predicted by theory. A theory that reliably predicts the results of experiments to which it is applicable is said to embody a law of physics. However, a law is always subject to modification, replacement, or restriction to a more limited domain, if a later experiment makes it necessary.

Ultimate aim of physics

The ultimate aim of physics is to find a unified set of laws governing matter, motion, and energy at small (microscopic) subatomic distances, at the human (macroscopic) scale of everyday life, and out to the largest distances (*e.g.*, those on the extragalactic scale). This ambitious goal has been realized to a notable extent. Although a completely unified theory of physical phenomena has not yet been achieved (and possibly never will be), a remarkably small set of fundamental physical laws appears able to account for all known phenomena. The body of physics developed up to about the turn of the 20th century, known as classical physics, can largely account for the motions of macroscopic objects that move slowly with respect to the speed of light and for such phenomena as heat, sound, electricity, magnetism, and light. The modern developments of relativity and quantum theory modify these laws insofar as they apply to higher speeds, very massive objects, and to the tiny elementary constituents of matter, such as electrons, protons, and neutrons.

THE SCOPE OF PHYSICS

The traditionally organized branches or fields of classical and modern physics are delineated below.

Mechanics. Mechanics is generally taken to mean the study of the motion of objects (or their lack of motion) under the action of given forces. Classical mechanics is sometimes considered a branch of applied mathematics. It consists of kinematics, the description of motion, and dynamics, the study of the action of forces in producing either motion or static equilibrium (the latter constituting the science of statics). The 20th-century subjects of quantum mechanics, crucial to treating the structure of matter, subatomic particles, superfluidity, superconductivity, neutron stars, and other major phenomena, and relativistic mechanics, important when speeds approach that of light, are forms of mechanics that will be discussed later in this section.

In classical mechanics the laws are initially formulated for point particles in which the dimensions, shapes, and other intrinsic properties of bodies are ignored. Thus in the first approximation even objects as large as the Earth and the Sun are treated as pointlike—*e.g.*, in calculating planetary orbital motion. In rigid-body dynamics, the extension of bodies and their mass distributions are considered as well, but they are imagined to be incapable of deformation. The mechanics of deformable solids is elasticity; hydrostatics and hydrodynamics treat, respectively, fluids at rest and in motion.

Newton's laws of motion

The three laws of motion set forth by Isaac Newton form the foundation of classical mechanics, together with the recognition that forces are directed quantities (vectors) and combine accordingly. The first law, also called the law of inertia, states that, unless acted upon by an external force, an object at rest remains at rest, or if in motion, it continues to move in a straight line with constant speed. Uniform motion therefore does not require a cause. Accordingly, mechanics concentrates not on motion as such but on the change in the state of motion of an object that results from the net force acting upon it. Newton's second law equates the net force on an object to the rate of change of its momentum, the latter being the product of the mass of a body and its velocity. Newton's third law, that of action and reaction, states that when two particles interact, the forces each exerts on the other are equal in magnitude and opposite in direction. Taken together, these mechanical laws in principle permit the determination of the future motions of a set of particles, providing their state of motion is known at some instant, as well as the forces that act between them and upon them from the outside. From this deterministic character of the laws of classical mechanics, profound (and probably incorrect) philosophical conclusions have been drawn in the past and even applied to human history.

Conservation laws

Lying at the most basic level of physics, the laws of mechanics are characterized by certain symmetry properties, as exemplified in the aforementioned symmetry between action and reaction forces. Other symmetries, such as the invariance (*i.e.*, unchanging form) of the laws under reflections and rotations carried out in space, reversal of time, or transformation to a different part of space or to a different epoch of time, are present both in classical mechanics and in relativistic mechanics, and with certain restrictions, also in quantum mechanics. The symmetry properties of the theory can be shown to have as mathematical consequences basic principles known as conservation laws, which assert the constancy in time of the values of certain physical quantities under prescribed conditions. The conserved quantities are the most important ones in physics; included among them are mass and energy (in relativity theory, mass and energy are equivalent and are conserved together), momentum, angular momentum, and electric charge.

The study of gravitation. This field of inquiry has in the past been placed within classical mechanics for historical reasons, because both fields were brought to a high state of perfection by Newton and also because of its universal character. Newton's gravitational law states that every material particle in the universe attracts every other one with a force that acts along the line joining them and whose strength is directly proportional to the product of their masses and inversely proportional to the square of their separation. Newton's detailed accounting for the orbits of the planets and the Moon, as well as for such subtle gravitational effects as the tides and the precession of the equinoxes (a slow cyclical change in direction of the Earth's axis of rotation) through this fundamental force was the first triumph of classical mechanics. No further principles are required to understand the principal aspects of rocketry and space flight (although, of course, a formidable technology is needed to carry them out).

Modern theory of gravitation

The modern theory of gravitation was formulated by Albert Einstein and is called the general theory of relativity. From the long-known equality of the quantity "mass" in Newton's second law of motion and that in his gravitational law, Einstein was struck by the fact that acceleration can locally annul a gravitational force (as occurs in the so-called weightlessness of astronauts in an Earth-orbiting

spacecraft) and was led thereby to the concept of curved space-time. Completed in 1915, the theory was valued for many years mainly for its mathematical beauty and for correctly predicting a small number of phenomena, such as the gravitational bending of light around a massive object. Only in recent years, however, has it become a vital subject for both theoretical and experimental research. (Relativistic mechanics, discussed below, refers to Einstein's special theory of relativity, which is not a theory of gravitation.)

The study of heat, thermodynamics, and statistical mechanics. Heat is a form of internal energy associated with the random motion of the molecular constituents of matter or with radiation. Temperature is an average of a part of the internal energy present in a body (it does not include the energy of molecular binding or of molecular rotation). The lowest possible energy state of a substance is defined as the absolute zero of temperature. An isolated body eventually reaches uniform temperature, a state known as thermal equilibrium, as do two or more bodies placed in contact. The formal study of states of matter at (or near) thermal equilibrium is called thermodynamics; it is capable of analyzing a large variety of thermal systems without considering their detailed microstructures.

First law. The first law of thermodynamics is the energy conservation principle of mechanics (*i.e.*, for all changes in an isolated system, the energy remains constant) generalized to include heat.

Second law. The second law of thermodynamics asserts that heat will not flow from a place of lower temperature to one where it is higher without the intervention of an external device (*e.g.*, a refrigerator). The concept of entropy involves the measurement of the state of disorder of the particles making up a system. For example, if tossing a coin many times results in a random-appearing sequence of heads and tails, the result has a higher entropy than if heads and tails tend to appear in clusters. Another formulation of the second law is that the entropy of an isolated system never decreases with time.

Third law. The third law of thermodynamics states that the entropy at the absolute zero of temperature is zero, corresponding to the most ordered possible state.

Statistical mechanics. The science of statistical mechanics derives bulk properties of systems from the mechanical properties of their molecular constituents, assuming molecular chaos and applying the laws of probability. Regarding each possible configuration of the particles as equally likely, the chaotic state (the state of maximum entropy) is so enormously more likely than ordered states that an isolated system will evolve to it, as stated in the second law of thermodynamics. Such reasoning, placed in mathematically precise form, is typical of statistical mechanics, which is capable of deriving the laws of thermodynamics but goes beyond them in describing fluctuations (*i.e.*, temporary departures) from the thermodynamic laws that describe only average behaviour. An example of a fluctuation phenomenon is the random motion of small particles suspended in a fluid, known as Brownian motion.

Quantum statistical mechanics

Quantum statistical mechanics plays a major role in many other modern fields of science, as, for example, in plasma physics (the study of fully ionized gases), in solid-state physics, and in the study of stellar structure. From a microscopic point of view the laws of thermodynamics imply that, whereas the total quantity of energy of any isolated system is constant, what might be called the quality of this energy is degraded as the system moves inexorably, through the operation of the laws of chance, to states of increasing disorder until it finally reaches the state of maximum disorder (maximum entropy), in which all parts of the system are at the same temperature, and none of the state's energy may be usefully employed. When applied to the universe as a whole, considered as an isolated system, this ultimate chaotic condition has been called the "heat death."

The study of electricity and magnetism. Although conceived of as distinct phenomena until the 19th century, electricity and magnetism are now known to be components of the unified field of electromagnetism. Particles with electric charge interact by an electric force, while charged particles in motion produce and respond to magnetic forces as well. Many subatomic particles, including the electrically charged electron and proton and the electrically neutral neutron, behave like elementary magnets. On the other hand, in spite of systematic searches undertaken, no magnetic monopoles, which would be the magnetic analogues of electric charges, have ever been found.

Significance of the field concept

The field concept plays a central role in the classical formulation of electromagnetism, as well as in many other areas of classical and contemporary physics. Einstein's gravitational field, for example, replaces Newton's concept of gravitational action at a distance. The field describing the electric force between a pair of charged particles works in the following manner: each particle creates an electric field in the space surrounding it, and so also at the position occupied by the other particle; each particle responds to the force exerted upon it by the electric field at its own position.

Classical electromagnetism is summarized by the laws of action of electric and magnetic fields upon electric charges and upon magnets and by four remarkable equations formulated in the latter part of the 19th century by James Clerk Maxwell. The latter equations describe the manner in which electric charges and currents produce electric and magnetic fields, as well as the manner in which changing magnetic fields produce electric fields, and vice versa. From these relations Maxwell inferred the existence of electromagnetic waves—associated electric and magnetic fields in space, detached from the charges that created them, traveling at the speed of light, and endowed with such "mechanical" properties as energy, momentum, and angular momentum. The light to which the human eye is sensitive is but one small segment of an electromagnetic spectrum that extends from long-wavelength radio waves to short-wavelength gamma rays and includes X rays, microwaves, and infrared (or heat) radiation.

Optics. Because light consists of electromagnetic waves, the propagation of light can be regarded as merely a branch of electromagnetism. However, it is usually dealt with as a separate subject called optics: the part that deals with the tracing of light rays is known as geometrical optics, while the part that treats the distinctive wave phenomena of light is called physical optics. More recently, there has developed a new and vital branch, quantum optics, which is concerned with the theory and application of the laser, a device that produces an intense coherent beam of unidirectional radiation useful for many applications.

Quantum optics

The formation of images by lenses, microscopes, telescopes, and other optical devices is described by ray optics, which assumes that the passage of light can be represented by straight lines, that is, rays. The subtler effects attributable to the wave property of visible light, however, require the explanations of physical optics. One basic wave effect is interference, whereby two waves present in a region of space combine at certain points to yield an enhanced resultant effect (*e.g.*, the crests of the component waves adding together); at the other extreme, the two waves can annul each other, the crests of one wave filling in the troughs of the other. Another wave effect is diffraction, which causes light to spread into regions of the geometric shadow and causes the image produced by any optical device to be fuzzy to a degree dependent on the wavelength of the light. Optical instruments such as the interferometer and the diffraction grating can be used for measuring the wavelength of light precisely (about 500 micrometres) and for measuring distances to a small fraction of that length.

Atomic and chemical physics. One of the great achievements of the 20th century has been the establishment of the validity of the atomic hypothesis, first proposed in ancient times, that matter is made up of relatively few kinds of small, identical parts—namely, atoms. However, unlike the indivisible atom of Democritus and other ancients, the atom, as it is conceived today, can be separated into constituent electrons and nucleus. Atoms combine to form molecules, whose structure is studied by chemistry and chemical physics; they also form other types of compounds, such as crystals, studied in the field of condensed-matter physics. Such disciplines study the most important

attributes of matter (not excluding biologic matter) that are encountered in normal experience—namely, those that depend almost entirely on the outer parts of the electronic structure of atoms. Only the mass of the atomic nucleus and its charge, which is equal to the total charge of the electrons in the neutral atom, affect the chemical and physical properties of matter.

Although there are some analogies between the solar system and the atom due to the fact that the strengths of gravitational and electrostatic forces both fall off as the inverse square of the distance, the classical forms of electromagnetism and mechanics fail when applied to tiny, rapidly moving atomic constituents. Atomic structure is comprehensible only on the basis of quantum mechanics, and its finer details require as well the use of quantum electrodynamics (QED) (see below).

Spectroscopy

Atomic properties are inferred mostly by the use of indirect experiments. Of greatest importance has been spectroscopy, which is concerned with the measurement and interpretation of the electromagnetic radiations either emitted or absorbed by materials. These radiations have a distinctive character, which quantum mechanics relates quantitatively to the structures that produce and absorb them. It is truly remarkable that these structures are in principle, and often in practice, amenable to precise calculation in terms of a few basic physical constants: the mass and charge of the electron, the speed of light, and Planck's constant h, the fundamental constant of the quantum theory named for the German physicist Max Planck.

Condensed-matter physics. This field, which treats the thermal, elastic, electrical, magnetic, and optical properties of solid and liquid substances, has grown at an explosive rate in recent years and has scored numerous important scientific and technical achievements, including the transistor. Among solid materials, the greatest theoretical advances have been in the study of crystalline materials whose simple repetitive geometric arrays of atoms are multiple-particle systems that allow treatment by quantum mechanics. Because the atoms in a solid are coordinated with each other over large distances, the theory must go beyond that appropriate for atoms and molecules. Thus conductors, such as metals, contain some so-called free (or conduction) electrons, which are responsible for the electrical and most of the thermal conductivity of the material and which belong collectively to the whole solid rather than to individual atoms. Semiconductors and insulators, either crystalline or amorphous, are other materials studied in this field of physics.

Other aspects of condensed matter involve the properties of the ordinary liquid state, of liquid crystals, and, at temperatures near absolute zero, of the so-called quantum liquids. The latter exhibit a property known as superfluidity (completely frictionless flow), which is an example of macroscopic quantum phenomena. Such phenomena are also exemplified by superconductivity (completely resistance-less flow of electricity), a low-temperature property of certain metallic and ceramic materials. Besides their significance to technology, macroscopic liquid and solid quantum states are important in astrophysical theories of stellar structure in, for example, neutron stars.

Research on superfluidity and superconductivity

Nuclear physics. This branch of physics deals with the structure of the atomic nucleus and the radiation from unstable nuclei. About 10,000 times smaller than the atom, the constituent particles of the nucleus, protons and neutrons, attract one another so strongly by the nuclear forces that nuclear energies are approximately 1,000,000 times larger than typical atomic energies. Quantum theory is needed for understanding nuclear structure.

Like excited atoms, unstable radioactive nuclei (either naturally occurring or artificially produced) can emit electromagnetic radiation. The energetic nuclear photons are called gamma rays. Radioactive nuclei also emit other particles: negative and positive electrons (beta rays), accompanied by neutrinos, and helium nuclei (alpha rays).

A principal research tool of nuclear physics involves the use of beams of particles (*e.g.*, protons or electrons) directed as projectiles against nuclear targets. Recoiling particles and any resultant nuclear fragments are detected, and their directions and energies are analyzed to reveal details of nuclear structure and to learn more about the strong nuclear force. A much weaker nuclear force, the so-called weak interaction, is responsible for the emission of beta rays. Nuclear collision experiments use beams of higher-energy particles, including those of unstable particles called mesons produced by primary nuclear collisions in accelerators dubbed meson factories. Exchange of mesons between protons and neutrons is directly responsible for the strong nuclear force. (For the mechanism underlying mesons, see below *Fundamental forces and fields.*)

In radioactivity and in collisions leading to nuclear breakup, the chemical identity of the nuclear target is altered whenever there is a change in the nuclear charge. In fission and fusion nuclear reactions in which unstable nuclei are, respectively, split into smaller nuclei or amalgamated into larger ones, the energy release far exceeds that of any chemical reaction.

Particle physics. One of the most significant branches of contemporary physics is the study of the fundamental subatomic constituents of matter, the elementary particles. This field, also called high-energy physics, emerged in the 1930s out of the developing experimental areas of nuclear and cosmic-ray physics. Initially investigators studied cosmic rays, the very-high-energy extraterrestrial radiations that fall upon the Earth and interact in the atmosphere (see below *The methodology of physics*). However, after World War II, scientists gradually began using high-energy particle accelerators to provide subatomic particles for study. Quantum field theory, a generalization of QED to other types of force fields, is essential for the analysis of high-energy physics.

High-energy physics

During recent decades a coherent picture has evolved of the underlying strata of matter involving three types of particles called leptons, quarks, and field quanta, for whose existence evidence is good. (Other types of particles have been hypothesized but have not yet been detected.) Subatomic particles cannot be visualized as tiny analogues of ordinary material objects such as billiard balls, for they have properties that appear contradictory from the classical viewpoint. That is to say, while they possess charge, spin, mass, magnetism, and other complex characteristics, they are nonetheless regarded as pointlike. Leptons and quarks occur in pairs (*e.g.*, one lepton pair consists of the electron and the neutrino). Each quark and each lepton have an antiparticle with properties that mirror those of its partner (the antiparticle of the negatively charged electron is the positive electron, or positron; that of the neutrino is the antineutrino). In addition to their electric and magnetic properties, quarks have very strong nuclear forces and also participate in the weak nuclear interaction, while leptons take part in only the weak interaction.

Ordinary matter consists of electrons surrounding the nucleus, which is composed of neutrons and protons, each of which is believed to contain three quarks. Quarks have charges that are either positive two-thirds or negative one-third of the electron's charge, while antiquarks have the opposite charges. Mesons, responsible for the nuclear binding force, are composed of one quark and one antiquark. In addition to the particles in ordinary matter and their antiparticles, which are referred to as first-generation, there are probably two or more additional generations of quarks and leptons, more massive than the first. Evidence exists at present for the second generation and all but one quark of the third, namely the t (or top) quark, which may be so massive that a new higher-energy accelerator may be needed to produce it.

The quantum fields through which quarks and leptons interact with each other and with themselves consist of particle-like objects called quanta (from which quantum mechanics derives its name). The first known quanta were those of the electromagnetic field; they are also called photons because light consists of them. A modern unified theory of weak and electromagnetic interactions, known as the electroweak theory, proposes that the weak nuclear interaction involves the exchange of particles about 100 times as massive as protons. These massive quanta have been observed—namely, two charged particles, W^+ and W^-, and a neutral one, W^0.

In the theory of strong nuclear interactions known as

Quantum chromodynamics

quantum chromodynamics (QCD), eight quanta, called gluons, bind quarks to form protons and neutrons and also bind quarks to antiquarks to form mesons, the force itself being dubbed the "colour force." (This unusual use of the term colour is a somewhat forced analogue of ordinary colour mixing.) Quarks are said to come in three colours—red, blue, and green. (The opposites of these imaginary colours, minus-red, minus-blue, and minus-green, are ascribed to antiquarks.) Only certain colour combinations, namely colour-neutral, or "white" (*i.e.,* equal mixtures of the above colours cancel out one another, resulting in no net colour), are conjectured to exist in nature in an observable form. The gluons and quarks themselves, being coloured, are permanently confined (deeply bound within the particles of which they are a part), while the colour-neutral composites such as protons can be directly observed. One consequence of colour confinement is that the observable particles are either electrically neutral or have charges that are integral multiples of the charge of the electron. A number of specific predictions of QCD have been experimentally tested and found correct.

Quantum mechanics. Although the various branches of physics differ in their experimental methods and theoretical approaches, certain general principles apply to all of them. The forefront of contemporary advances in physics lies in the submicroscopic regime, whether it be in atomic, nuclear, condensed-matter, plasma, or particle physics, or in quantum optics, or even in the study of stellar structure. All are based upon quantum theory (*i.e.,* quantum mechanics and quantum field theory) and relativity, which together form the theoretical foundations of modern physics. Many physical quantities whose classical counterparts vary continuously over a range of possible values are in quantum theory constrained to have discontinuous, or discrete, values. The intrinsically deterministic character of classical physics is replaced in quantum theory by intrinsic uncertainty.

Differences from classical physics

According to quantum theory, electromagnetic radiation does not always consist of continuous waves; instead it must be viewed under some circumstances as a collection of particle-like photons, the energy and momentum of each being directly proportional to its frequency (or inversely proportional to its wavelength, the photons still possessing some wavelike characteristics). Conversely, electrons and other objects that appear as particles in classical physics are endowed by quantum theory with wavelike properties as well, such a particle's quantum wavelength being inversely proportional to its momentum. In both instances, the proportionality constant is the characteristic quantum of action (action being defined as energy × time)—that is to say, Planck's constant *h*.

In principle, all of atomic and molecular physics, including the structure of atoms and their dynamics, the periodic table of elements and their chemical behaviour, as well as the spectroscopic, electrical, and other physical properties of atoms, molecules, and condensed matter, can be accounted for by quantum mechanics. Roughly speaking, the electrons in the atom must fit around the nucleus as some sort of standing wave (the Schrödinger wave function; see above) analogous to the waves on a plucked violin or guitar string. As the fit determines the wavelength of the quantum wave, it necessarily determines its energy state. Consequently, atomic systems are restricted to certain discrete, or quantized, energies. When an atom undergoes a discontinuous transition, or quantum jump, its energy changes abruptly by a sharply defined amount, and a photon of that energy is emitted when the energy of the atom decreases, or is absorbed in the opposite case.

Although atomic energies can be sharply defined, the positions of the electrons within the atom cannot be, quantum mechanics giving only the probability for the electrons to have certain locations. This is a consequence of the feature that distinguishes quantum theory from all other approaches to physics, the indeterminacy (or uncertainty) principle of Werner Heisenberg. As was explained earlier in the article, this principle holds that measuring a particle's position with increasing precision necessarily increases the uncertainty as to the particle's momentum, and conversely. The ultimate degree of uncertainty is controlled by the magnitude of Planck's constant, which is so small as to have no apparent effects except in the world of microstructures. In the latter case, however, because both a particle's position and its velocity or momentum must be known precisely at some instant in order to predict its future history, quantum theory precludes such certain prediction and thus escapes determinism.

The indeterminacy principle as the distinguishing feature of quantum theory

The complementary wave and particle aspects, or wave–particle duality, of electromagnetic radiation and of material particles furnish another illustration of the uncertainty principle. When an electron exhibits wavelike behaviour, as in the phenomenon of electron diffraction, this excludes its exhibiting particle-like behaviour in the same observation. Similarly, when electromagnetic radiation in the form of photons interacts with matter, as in the Compton effect in which X-ray photons collide with electrons, the result resembles a particle-like collision and the wave nature of electromagnetic radiation is precluded. The principle of complementarity, asserted by Niels Bohr, who pioneered the theory of atomic structure, states that the physical world presents itself in the form of various complementary pictures, no one of which is by itself complete, all of these pictures being essential for our total understanding. Thus both wave and particle pictures are needed for understanding either the electron or the photon.

Although it deals with probabilities and uncertainties, the quantum theory has been spectacularly successful in explaining otherwise inaccessible atomic phenomena and in thus far meeting every experimental test. Its predictions, especially those of QED, are the most precise and the best checked of any in physics; some of them have been tested and found accurate to better than one part per billion.

Relativistic mechanics. In classical physics, space is conceived as having the absolute character of an empty stage in which events in nature unfold as time flows onward independently; events occurring simultaneously for one observer are presumed to be simultaneous for any other; mass is taken as impossible to create or destroy; and a particle given sufficient energy acquires a velocity that can increase without limit. The special theory of relativity, developed principally by Einstein in 1905 and now so adequately confirmed by experiment as to have the status of physical law, shows that all these, as well as other apparently obvious assumptions, are false.

The special theory of relativity and its ramifications

Specific and unusual relativistic effects flow directly from Einstein's two basic postulates, which are formulated in terms of so-called inertial reference frames. These are reference systems that move in such a way that in them Newton's first law, the law of inertia, is valid. The set of inertial frames consists of all those that move with constant velocity with respect to each other (accelerating frames therefore being excluded). Einstein's postulates are: (1) All observers, whatever their state of motion relative to a light source, measure the same speed for light; and (2) The laws of physics are the same in all inertial frames.

The first postulate, the constancy of the speed of light, is an experimental fact from which follow the distinctive relativistic phenomena of space contraction, time dilation, and the relativity of simultaneity: as measured by an observer assumed to be at rest, an object in motion is contracted along the direction of its motion, and moving clocks run slow; two spatially separated events that are simultaneous for a stationary observer occur sequentially for a moving observer. As a consequence, space intervals in three-dimensional space are related to time intervals, thus forming so-called four-dimensional space-time.

The second postulate is called the principle of relativity. It is equally valid in classical mechanics (but not in classical electrodynamics until Einstein reinterpreted it). This postulate implies, for example, that table tennis played on a train moving with constant velocity is just like table tennis played with the train at rest, the states of rest and motion being physically indistinguishable. In relativity theory, mechanical quantities such as momentum and energy have forms that are different from their classical counterparts but give the same values for speeds that are small compared to the speed of light, the maximum permissible speed in nature (about 300,000 kilometres per second, or 186,000 miles per second). According to rela-

Equivalence of mass and energy

tivity, mass and energy are equivalent and interchangeable quantities, the equivalence being expressed by Einstein's famous equation $E = mc^2$, where m is an object's mass and c is the speed of light.

The general theory of relativity, as discussed above, is Einstein's theory of gravitation, which uses the principle of the equivalence of gravitation and locally accelerating frames of reference. Einstein's theory has special mathematical beauty; it generalizes the "flat" space-time concept of special relativity to one of curvature. It forms the background of all modern cosmological theories (see COSMOS: *Relativistic cosmologies*). In contrast to some vulgarized popular notions of it, which confuse it with moral and other forms of relativism, Einstein's theory does not argue that "all is relative." On the contrary, it is largely a theory based upon those physical attributes that do not change, or, in the language of the theory, that are invariant.

Conservation laws and symmetry. Since the early period of modern physics, there have been conservation laws, which state that certain physical quantities, such as the total electric charge of an isolated system of bodies, do not change in the course of time. In the 20th century it has been proved mathematically that such laws follow from the symmetry properties of nature, as expressed in the laws of physics. The conservation of mass-energy of an isolated system, for example, follows from the assumption that the laws of physics may depend upon time intervals but not upon the specific time at which the laws are applied. The symmetries and the conservation laws that follow from them are regarded by modern physicists as being even more fundamental than the laws themselves, since they are able to limit the possible forms of laws that may be proposed in the future.

Conservation laws are valid in classical, relativistic, and quantum theory for mass-energy, momentum, angular momentum, and electric charge. (In nonrelativistic physics, mass and energy are separately conserved.) Momentum, a directed quantity equal to the mass of a body multiplied by its velocity or to the total mass of two or more bodies multiplied by the velocity of their centre of mass, is conserved when, and only when, no external force acts. Similarly angular momentum, which is related to spinning motions, is conserved in a system upon which no net turning force, called torque, acts. External forces and torques break the symmetry conditions from upon which the respective conservation laws follow.

In quantum theory, and especially in the theory of elementary particles, there are additional symmetries and conservation laws, some exact and others only approximately valid, which play no significant role in classical physics. Among these are the conservation of so-called quantum numbers related to left-right reflection symmetry of space (called parity) and to the reversal symmetry of motion (called time reversal). These quantum numbers are conserved in all processes other than the weak nuclear interaction.

Internal symmetries

Other symmetry properties not obviously related to space and time (and referred to as internal symmetries) characterize the different families of elementary particles and, by extension, their composites. Quarks, for example, have a property called baryon number, as do protons, neutrons, nuclei, and unstable quark composites. All of these except the quarks are known as baryons. A failure of baryon-number conservation would exhibit itself, for instance, by a proton decaying into lighter non-baryonic particles. Indeed, intensive search for such proton decay has been conducted, but so far it has been fruitless. Similar symmetries and conservation laws hold for an analogously defined lepton number, and they also appear, as does the law of baryon conservation, to hold absolutely.

Fundamental forces and fields. The four basic forces of nature, in order of increasing strength, are thought to be: (1) the gravitational force between particles with mass; (2) the electromagnetic force between particles with charge or magnetism or both; (3) the colour force between quarks; and (4) the weak nuclear interaction by which, for example, quarks can change their type, so that a neutron decays into a proton, an electron, and an antineutrino. The strong nuclear interaction that binds protons and neutrons into nuclei and is responsible for fission, fusion, and other nuclear reactions is in principle derived from the colour force. Nuclear physics is thus related to QCD as chemistry is to atomic physics.

Underlying principles of quantum field theory

According to quantum field theory, each of the four fundamental interactions is mediated by the exchange of quanta, called vector gauge bosons, which share certain common characteristics. All have an intrinsic spin of one unit, measured in terms of Planck's constant h. (Leptons and quarks each have one-half unit of spin.) The term gauge refers to a special type of symmetry they possess. This symmetry was first seen in the equations for electromagnetic potentials, quantities from which electromagnetic fields can be derived. It is possessed in pure form by the eight massless gluons of QCD, but in the electroweak theory, the unified theory of electromagnetic and weak interactions, gauge symmetry is partially broken, so that only the photon remains massless, with the other gauge bosons (W^+, W^-, and Z) acquiring large masses. At present, theoretical physicists are trying to produce a further unification of QCD with the electroweak theory and, more ambitiously still, to unify them with a quantum version of gravity in which the force would be transmitted by massless quanta of two units of spin called gravitons.

THE METHODOLOGY OF PHYSICS

Physics has evolved and continues to evolve without any single strategy. Essentially an experimental science, refined measurements can reveal unexpected behaviour. On the other hand, mathematical extrapolation of existing theories into new theoretical areas, critical reexamination of apparently obvious but untested assumptions, argument by symmetry or analogy, aesthetic judgment, pure accident, and hunch—each of these plays a role (as in all of science). Thus, for example, the quantum hypothesis proposed by Planck was based on observed departures of the character of blackbody radiation (radiation emitted by a heated body that absorbs all radiant energy incident upon it) from that predicted by classical electromagnetism. P.A.M. Dirac predicted the existence of the positron in making a relativistic extension of the quantum theory of the electron. The elusive neutrino, without mass or charge, was hypothesized by Wolfgang Pauli as an alternative to abandoning the conservation laws in the beta-decay process. Maxwell conjectured that if changing magnetic fields create electric fields (which was known to be so), then changing electric fields might create magnetic fields, leading him to the electromagnetic theory of light. Einstein's special theory of relativity was based on a critical reexamination of the meaning of simultaneity, while his general theory rests on the equivalence of inertial and gravitational mass.

Although the tactics may vary from problem to problem, the physicist invariably tries to make unsolved problems more tractable by constructing a series of idealized models, with each successive model being a more realistic representation of the actual physical situation. Thus, in the theory of gases, the molecules are at first imagined to be particles that are as structureless as billiard balls with vanishingly small dimensions. This ideal picture is then improved on step by step.

Correspondence principle

The correspondence principle, a useful guiding principle for extending theoretical interpretations, was formulated by Niels Bohr in the context of the quantum theory. It asserts that when a valid theory is generalized to a broader arena, the new theory's predictions must agree with the old one in the overlapping region in which both are applicable. For example, the more comprehensive theory of physical optics must yield the same result as the more restrictive theory of ray optics whenever wave effects proportional to the wavelength of light are negligible on account of the smallness of that wavelength. Similarly, quantum mechanics must yield the same results as classical mechanics in circumstances when Planck's constant can be considered as negligibly small. Likewise, for speeds small compared to the speed of light (as for baseballs in play), relativistic mechanics must coincide with Newtonian classical mechanics.

Some ways in which experimental and theoretical physi-

cists attack their problems are illustrated by the following examples.

The modern experimental study of elementary particles began with the detection of new types of unstable particles produced in the atmosphere by primary radiation, the latter consisting mainly of high-energy protons arriving from space. The new particles were detected in Geiger counters and identified by the tracks they left in instruments called cloud chambers and in photographic plates. After World War II, particle physics, then known as high-energy nuclear physics, became a major field of science. Today's high-energy particle accelerators can be several kilometres in length, cost hundreds (or even thousands) of millions of dollars, and accelerate particles to enormous energies (trillions of electron volts). Experimental teams, such as those that discovered the W^+, W^-, and Z quanta of the weak force at the European Laboratory for Particle Physics (CERN) in Geneva can have 100 or more physicists from many countries, along with a larger number of technical workers serving as support personnel. A variety of visual and electronic techniques are used to interpret and sort the huge amounts of data produced by their efforts, and particle-physics laboratories are major users of the most advanced technology, be it superconductive magnets or supercomputers.

Reliance on advanced technology

Theoretical physicists use mathematics both as a logical tool for the development of theory and for calculating predictions of the theory to be compared with experiment. Newton, for one, invented integral calculus to solve the following problem, which was essential to his formulation of the law of universal gravitation: Assuming that the attractive force between any pair of point particles is inversely proportional to the square of the distance separating them, how does a spherical distribution of particles, such as the Earth, attract another nearby object? Integral calculus, a procedure for summing many small contributions, yields the simple solution that the Earth itself acts as a point particle with all its mass concentrated at the centre. In modern physics, Dirac predicted the existence of the then-unknown positive electron (or positron) by finding an equation for the electron that would combine quantum mechanics and the special theory of relativity.

Mathematics and electronic computers as invaluable tools

RELATIONS BETWEEN PHYSICS AND OTHER DISCIPLINES AND SOCIETY

Influence of physics on related disciplines. Because physics elucidates the simplest fundamental questions in nature on which there can be a consensus, it is hardly surprising that it has had a profound impact on other fields of science, on philosophy, on the worldview of the developed world, and, of course, on technology.

Indeed, whenever a branch of physics has reached such a degree of maturity that its basic elements are comprehended in general principles, it has moved from basic to applied physics and thence to technology. Thus almost all current activity in classical physics consists of applied physics, and its contents form the core of many branches of engineering. Discoveries in modern physics are converted with increasing rapidity into technical innovations and analytical tools for associated disciplines. There are, for example, such nascent fields as nuclear and biomedical engineering, quantum chemistry and quantum optics, and radio, X-ray, and gamma-ray astronomy, as well as such analytic tools as radioisotopes, spectroscopy, and lasers, which all stem directly from basic physics.

Applied physics

Apart from its specific applications, physics—especially Newtonian mechanics—has become the prototype of the scientific method, its experimental and analytic methods sometimes being imitated (and sometimes inappropriately so) in fields far from the related physical sciences. Some of the organizational aspects of physics, based partly on the successes of the radar and atomic-bomb projects of World War II, also have been imitated in large-scale scientific projects, as, for example, in astronomy and space research.

The impact of physics on philosophy

The great influence of physics on the branches of philosophy concerned with the conceptual basis of human perceptions and understanding of nature, such as epistemology, is evidenced by the earlier designation of physics itself as natural philosophy. Present-day philosophy of science deals largely, though not exclusively, with the foundations of physics. Determinism, the philosophical doctrine that the universe is a vast machine operating with strict causality whose future is determined in all detail by its present state, is rooted in Newtonian mechanics, which obeys that principle. Moreover, the schools of materialism, naturalism, and empiricism have in large degree considered physics to be a model for philosophical inquiry. An extreme position is taken by the logical positivists, whose radical distrust of the reality of anything not directly observable leads them to demand that all significant statements must be formulated in the language of physics.

The uncertainty principle of quantum theory has prompted a reexamination of the question of determinism, and its other philosophical implications remain in doubt. Particularly problematic is the matter of the meaning of measurement, for which recent theories and experiments confirm some apparently noncausal predictions of standard quantum theory. It is fair to say that though physicists agree that quantum theory works, they still differ as to what it means.

Influence of related disciplines on physics. The relationship of physics to its bordering disciplines is a reciprocal one. Just as technology feeds on fundamental science for new practical innovations, so physics appropriates the techniques and instrumentation of modern technology for advancing itself. Thus experimental physicists utilize increasingly refined and precise electronic devices. Moreover, they work closely with engineers in designing basic scientific equipment, such as high-energy particle accelerators. Mathematics has always been the primary tool of the theoretical physicist, and even abstruse fields of mathematics such as group theory and differential geometry have become invaluable to the theoretician classifying subatomic particles or investigating the symmetry characteristics of atoms and molecules. Much of contemporary research in physics depends on the high-speed computer. It allows the theoretician to perform computations that are too lengthy or complicated to be done with paper and pencil. Also, it allows experimentalists to incorporate the computer into their apparatus, so that the results of measurements can be provided nearly instantaneously on-line as summarized data while an experiment is in progress.

The physicist in society. Because of the remoteness of much of contemporary physics from ordinary experience and its reliance on advanced mathematics, physicists have sometimes seemed to the public to be initiates in a latter-day secular priesthood who speak an arcane language and can communicate their findings to laymen only with great difficulty. Yet, the physicist has come to play an increasingly significant role in society, particularly since World War II. Governments have supplied substantial funds for research at academic institutions and at government laboratories through such agencies as the National Science Foundation and the Department of Energy in the United States, which has also established a number of national laboratories, including the Fermi National Accelerator Laboratory in Batavia, Ill., with the world's largest particle accelerator. CERN, mentioned above, is composed of 14 European countries and operates a large accelerator at the Swiss–French border. Physics research is supported in the Federal Republic of Germany by the Max Planck Society for the Advancement of Science and in Japan by the Japan Society for the Promotion of Science. In Trieste, Italy, there is the International Center for Theoretical Physics, which has strong ties to developing countries. These are only a few examples of the widespread international interest in fundamental physics.

Basic research in physics is obviously dependent on public support and funding, and with this development has come, albeit slowly, a growing recognition within the physics community of the social responsibility of scientists for the consequences of their work and for the more general problems of science and society. (R.T.We./L.M.B.)

Chemistry

Chemistry is the study of the material substances that occur on Earth and elsewhere in the universe. The great

Principal concerns

challenge in chemistry is the development of a coherent explanation of the complex behaviour of materials, why they appear as they do, what gives them their enduring properties, and how interactions among different substances can bring about the formation of new substances and the destruction of old ones. From the earliest attempts to understand the material world in rational terms, chemists have struggled to develop theories of matter that satisfactorily explain both permanence and change. The ordered assembly of indestructible atoms into small and large molecules, or extended networks of intermingled atoms, is generally accepted as the basis of permanence, while the reorganization of atoms or molecules into different arrangements lies behind theories of change. Thus chemistry involves the study of the atomic composition and structural architecture of substances, as well as the varied interactions among substances that can lead to sudden, often violent reactions.

Practical applications

Chemistry also is concerned with the utilization of natural substances and the creation of artificial ones. Cooking, fermentation, glass making, and metallurgy are all chemical processes that date from the beginnings of civilization. Today, vinyl, Teflon, liquid crystals, semiconductors, and superconductors represent the fruits of chemical technology. The 20th century has seen dramatic advances in the comprehension of the marvelous and complex chemistry of living organisms, and a molecular interpretation of health and disease holds great promise. Modern chemistry, aided by increasingly sophisticated instruments, studies materials as small as single atoms and as large and complex as DNA (deoxyribonucleic acid), which contains millions of atoms. New substances can even be designed to bear desired characteristics and then synthesized. The rate at which chemical knowledge continues to accumulate is remarkable. Over time more than 8,000,000 different chemical substances, both natural and artificial, have been characterized and produced. The number was less than 500,000 as recently as 1965.

Intimately interconnected with the intellectual challenges of chemistry are those associated with industry. In the mid-19th century the German chemist Justus von Liebig commented that the wealth of a nation could be gauged by the amount of sulfuric acid it produced. This acid, essential to many manufacturing processes, remains today the leading chemical product of industrialized countries. As Liebig recognized, a country that produces large amounts of sulfuric acid is one with a strong chemical industry and a strong economy as a whole. The production, distribution, and utilization of a wide range of chemical products is common to all highly developed nations. In fact, one can say that the "iron age" of civilization is being replaced by a "polymer age," for in some countries the total volume of polymers now produced exceeds that of iron.

THE SCOPE OF CHEMISTRY

The days are long past when one person could hope to have a detailed knowledge of all areas of chemistry. Those pursuing their interests into specific areas of chemistry communicate with others who share the same interests. Over time a group of chemists with specialized research interests become the founding members of an area of specialization. The areas of specialization that emerged early in the history of chemistry, such as organic, inorganic, physical, analytical, and industrial chemistry, along with biochemistry, remain of greatest general interest. There has been, however, much growth in the areas of polymer, environmental, and medicinal chemistry during the 20th century. Moreover, new specialities continue to appear, as, for example, pesticide, forensic, and computer chemistry.

Analytical chemistry. Most of the materials that occur on Earth, such as wood, coal, minerals, or air, are mixtures of many different and distinct chemical substances. Each pure chemical substance (*e.g.*, oxygen, iron, or water) has a characteristic set of properties that gives it its chemical identity. Iron, for example, is a common silver-white metal that melts at 1,535° C, is very malleable, and readily combines with oxygen to form the common substances hematite and magnetite. The detection of iron in a mixture of metals, or in a compound such as magnetite, is a branch of analytical chemistry called qualitative analysis. Measurement of the actual amount of a certain substance in a compound or mixture is termed quantitative analysis. Quantitative analytic measurement has determined, for instance, that iron makes up 72.3 percent, by mass, of magnetite, the mineral commonly seen as black sand along beaches and stream banks. Over the years, chemists have discovered chemical reactions that indicate the presence of such elemental substances by the production of easily visible and identifiable products. Iron can be detected by chemical means if it is present in a sample to an amount of 1 part per million or greater. Some very simple qualitative tests reveal the presence of specific chemical elements in even smaller amounts. The yellow colour imparted to a flame by sodium is visible if the sample being ignited has as little as one-billionth of a gram of sodium. Such analytic tests have allowed chemists to identify the types and amounts of impurities in various substances and to determine the properties of very pure materials. Substances used in common laboratory experiments generally have impurity levels of less than 0.1 percent. For special applications, one can purchase chemicals that have impurities totaling less than 0.001 percent. The identification of pure substances and the analysis of chemical mixtures enable all other chemical disciplines to flourish.

Qualitative analysis and quantitative analysis

The importance of analytical chemistry has never been greater than it is today. The demand in modern societies for a variety of safe foods, affordable consumer goods, abundant energy, and labour-saving technologies places a great burden on the environment. All chemical manufacturing produces waste products in addition to the desired substances, and waste disposal has not always been carried out carefully. Disruption of the environment has occurred since the dawn of civilization, and pollution problems have increased with the growth of global population. The techniques of analytical chemistry are relied on heavily to maintain a benign environment. The undesirable substances in water, air, soil, and food must be identified, their point of origin fixed, and safe, economical methods for their removal or neutralization developed. Once the amount of a pollutant deemed to be hazardous has been assessed, it becomes important to detect harmful substances at concentrations well below the danger level. Analytical chemists seek to develop increasingly accurate and sensitive techniques and instruments.

Sophisticated analytic instruments, often coupled with computers, have improved the accuracy with which chemists can identify substances and have lowered detection limits. An analytic technique in general use is gas chromatography, which separates the different components of a gaseous mixture by passing the mixture through a long, narrow column of absorbent but porous material. The different gases interact differently with this absorbent material and pass through the column at different rates. As the separate gases flow out of the column, they can be passed into another analytic instrument called a mass spectrometer, which separates substances according to the mass of their constituent ions. A combined gas chromatograph–mass spectrometer can rapidly identify the individual components of a chemical mixture whose concentrations may be no greater than a few parts per billion. Similar or even greater sensitivities can be obtained under favourable conditions using techniques such as atomic absorption, polarography, and neutron activation. The rate of instrumental innovation is such that analytic instruments often become obsolete within 10 years of their introduction. Newer instruments are more accurate and faster and are employed widely in the areas of environmental and medicinal chemistry.

Gas chromatography and mass spectrometry

Inorganic chemistry. Modern chemistry, which dates more or less from the acceptance of the law of conservation of mass in the late 18th century, focused initially on those substances that were not associated with living organisms. Study of such substances, which normally have little or no carbon, constitutes the discipline of inorganic chemistry. Early work sought to identify the simple substances—namely, the elements—that are the constituents of all more complex substances. Some elements, such as gold and carbon, have been known since antiquity, and

many others were discovered and studied throughout the 19th and early 20th centuries. Today, more than 100 are known. The study of such simple inorganic compounds as sodium chloride (common salt) has led to some of the fundamental concepts of modern chemistry, the law of definite proportions providing one notable example. This law states that for most pure chemical substances the constituent elements are always present in fixed proportions by mass (*e.g.*, every 100 grams of salt contains 39.3 grams of sodium and 60.7 grams of chlorine). The crystalline form of salt, known as halite, consists of intermingled sodium and chlorine atoms, one sodium atom for each one of chlorine. Such a compound, formed solely by the combination of two elements, is known as a binary compound. Binary compounds are very common in inorganic chemistry, and they exhibit little structural variety. For this reason, the number of inorganic compounds is limited in spite of the large number of elements that may react with each other. If three or more elements are combined in a substance, the structural possibilities become greater.

After a period of quiescence in the early part of the 20th century, inorganic chemistry has again become an exciting area of research. Compounds of boron and hydrogen, known as boranes, have unique structural features that forced a change in thinking about the architecture of inorganic molecules. Some inorganic substances have structural features long believed to occur only in carbon compounds, and a few inorganic polymers have even been produced. Ceramics are materials composed of inorganic elements combined with oxygen. For centuries ceramic objects have been made by strongly heating a vessel formed from a paste of powdered minerals. Although ceramics are quite hard and stable at very high temperatures, they are usually brittle. Currently, new ceramics strong enough to be used as turbine blades in jet engines are being manufactured. There is hope that ceramics will one day replace steel in components of internal-combustion engines. In 1987 a ceramic containing yttrium, barium, copper, and oxygen, with the approximate formula $YBa_2Cu_3O_7$, was found to be a superconductor at a temperature of about 100 K. A superconductor offers no resistance to the passage of an electrical current, and this new type of ceramic could very well find wide use in electrical and magnetic applications. A superconducting ceramic is so simple to make that it can be prepared in a high school laboratory. Its discovery illustrates the unpredictability of chemistry, for fundamental discoveries can still be made with simple equipment and inexpensive materials.

Development of new ceramics

Many of the most interesting developments in inorganic chemistry bridge the gap with other disciplines. Organometallic chemistry investigates compounds that contain inorganic elements combined with carbon-rich units. Many organometallic compounds play an important role in industrial chemistry as catalysts, which are substances that are able to accelerate the rate of a reaction even when present in only very small amounts. Some success has been achieved in the use of such catalysts for converting natural gas to related but more useful chemical substances. Chemists also have created large inorganic molecules that contain a core of metal atoms, such as platinum, surrounded by a shell of different chemical units. Some of these compounds, referred to as metal clusters, have characteristics of metals, while others react in ways similar to biologic systems. Trace amounts of metals in biologic systems are essential for processes such as respiration, nerve function, and cell metabolism. Processes of this kind form the object of study of bioinorganic chemistry. Although organic molecules were once thought to be the distinguishing chemical feature of living creatures, it is now known that inorganic chemistry plays a vital role as well.

Organometallic chemistry

Bioinorganic chemistry

Organic chemistry. Organic compounds are based on the chemistry of carbon. Carbon is unique in the variety and extent of structures that can result from the three-dimensional connections of its atoms. The process of photosynthesis converts carbon dioxide and water to oxygen and compounds known as carbohydrates. Both cellulose, the substance that gives structural rigidity to plants, and starch, the energy storage product of plants, are polymeric carbohydrates. Simple carbohydrates produced by photosynthesis form the raw material for the myriad organic compounds found in the plant and animal kingdoms. When combined with variable amounts of hydrogen, oxygen, nitrogen, sulfur, phosphorus, and other elements, the structural possibilities of carbon compounds become limitless, and their number far exceeds the total of all nonorganic compounds. A major focus of organic chemistry is the isolation, purification, and structural study of these naturally occurring substances. Many natural products are simple molecules. Examples include formic acid (HCO_2H) in ants, ethyl alcohol (C_2H_5OH) in fermenting fruit, and oxalic acid ($C_2H_2O_4$) in rhubarb leaves. Other natural products, such as penicillin, vitamin B_{12}, proteins, and nucleic acids, are exceedingly complex. The isolation of pure natural products from their host organism is made difficult by the low concentrations in which they may be present. Once they are isolated in pure form, however, modern instrumental techniques can reveal structural details for amounts weighing as little as one-millionth of a gram. The correlation of the physical and chemical properties of compounds with their structural features is the domain of physical organic chemistry. Once the properties endowed upon a substance by specific structural units termed functional groups are known, it becomes possible to design novel molecules that may exhibit desired properties. The preparation, under controlled laboratory conditions, of specific compounds is known as synthetic chemistry. Some products are easier to synthesize than to collect and purify from their natural sources. Tons of vitamin C, for example, are synthesized annually. Many synthetic substances have novel properties that make them especially useful. Plastics are a prime example, as are many drugs and agricultural chemicals. A continuing challenge for synthetic chemists is the structural complexity of most organic substances. To synthesize a desired substance, the atoms must be pieced together in the correct order and with the proper three-dimensional relationships. Just as a given pile of lumber and bricks can be assembled in many ways to build houses of several different designs, so too can a fixed number of atoms be connected together in various ways to give different molecules. Only one structural arrangement out of the many possibilities will be identical with a naturally occurring molecule. The antibiotic erythromycin, for example, contains 37 carbon, 67 hydrogen, and 13 oxygen atoms, along with one nitrogen atom. Even when joined together in the proper order, these 118 atoms can give rise to 262,144 different structures, only one of which has the characteristics of natural erythromycin. The great abundance of organic compounds, their fundamental role in the chemistry of life, and their structural diversity have made their study especially challenging and exciting. Organic chemistry is the largest area of specialization among the various fields of chemistry.

Isolation, purification, and structural study of material substances

Synthetic chemistry

Biochemistry. As understanding of inanimate chemistry grew during the 19th century, attempts to interpret the physiological processes of living organisms in terms of molecular structure and reactivity gave rise to the discipline of biochemistry. Biochemists employ the techniques and theories of chemistry to probe the molecular basis of life. An organism is investigated on the premise that its physiological processes are the consequence of many thousands of chemical reactions occurring in a highly integrated manner. Biochemists have established, among other things, the principles that underlie energy transfer in cells, the chemical structure of cell membranes, the coding and transmission of hereditary information, muscular and nerve function, and biosynthetic pathways. In fact, related biomolecules have been found to fulfill similar roles in organisms as different as bacteria and human beings. The study of biomolecules, however, presents many difficulties. Such molecules are often very large and exhibit great structural complexity; moreover, the chemical reactions they undergo are usually exceedingly fast. The separation of the two strands of DNA, for instance, occurs in one-millionth of a second. Such rapid rates of reaction are possible only through the intermediary action of biomolecules called enzymes. Enzymes are proteins that owe their remarkable rate-accelerating abilities to their three-dimen-

Probing the molecular basis of life

sional chemical structure. Not surprisingly, biochemical discoveries have had a great impact on the understanding and treatment of disease. Many ailments due to inborn errors of metabolism have been traced to specific genetic defects. Other diseases result from disruptions in normal biochemical pathways.

Concern with therapeutic agents

Frequently, symptoms can be alleviated by drugs, and the discovery, mode of action, and degradation of therapeutic agents is another of the major areas of study in biochemistry. Bacterial infections can be treated with sulfonamides, penicillins, and tetracyclines, and research into viral infections has revealed the effectiveness of acyclovir against the herpes virus. There is much current interest in the details of carcinogenesis and cancer chemotherapy. It is known, for example, that cancer can result when cancer-causing molecules, or carcinogens as they are called, react with nucleic acids and proteins and interfere with their normal modes of action. Researchers have developed tests that can identify molecules likely to be carcinogenic. The hope, of course, is that progress in the prevention and treatment of cancer will accelerate once the biochemical basis of the disease is more fully understood.

Methods of gene synthesis

The molecular basis of biologic processes is an essential feature of the fast-growing disciplines of molecular biology and biotechnology. Chemistry has developed methods for rapidly and accurately determining the structure of proteins and DNA. In addition, efficient laboratory methods for the synthesis of genes are being devised. Ultimately, the correction of genetic diseases by replacement of defective genes with normal ones may become possible.

Polymer chemistry. The simple substance ethylene is a gas composed of molecules with the formula CH_2CH_2. Under certain conditions, many ethylene molecules will join together to form a long chain called polyethylene, with the formula $(CH_2CH_2)_n$, where n is a variable but large number. Polyethylene is a tough, durable solid material quite different from ethylene. It is an example of a polymer, which is a large molecule made up of many smaller molecules (monomers), usually joined together in a linear fashion. Many naturally occurring substances, including cellulose, starch, cotton, wool, rubber, leather, proteins, and DNA, are polymers. Polyethylene, nylon, and acrylics are examples of synthetic polymers. The study of such materials lies within the domain of polymer chemistry, a specialty that has flourished in the 20th century. The investigation of natural polymers overlaps considerably with biochemistry, but the synthesis of new polymers, the investigation of polymerization processes, and the characterization of the structure and properties of polymeric materials all pose unique problems for polymer chemists.

Polymer chemists have designed and synthesized polymers that vary in hardness, flexibility, softening temperature, solubility in water, and biodegradability. They have produced polymeric materials that are as strong as steel yet lighter and more resistant to corrosion. Oil, natural gas, and water pipelines are now routinely constructed of plastic pipe. In recent years, automakers have increased their use of plastic components to build lighter vehicles that consume less fuel. Other industries such as those involved in the manufacture of textiles, rubber, paper, and packaging materials are built upon polymer chemistry.

Developing special catalysts

Besides producing new kinds of polymeric materials, researchers are concerned with developing special catalysts that are required by the large-scale industrial synthesis of commercial polymers. Without such catalysts, the polymerization process would be very slow in certain cases.

Physical chemistry. Many chemical disciplines, such as those already discussed, focus on certain classes of materials that share common structural and chemical features. Other specialties may be centred not on a class of substances but rather on their interactions and transformations. The oldest of these fields is physical chemistry, which seeks to measure, correlate, and explain the quantitative aspects of chemical processes. The Anglo-Irish chemist Robert Boyle, for example, discovered in the 17th century that at room temperature the volume of a fixed quantity of gas decreases proportionally as the pressure on it increases. Thus, for a gas at constant temperature, the product of its volume V and pressure P equals a constant number—*i.e.*, $PV =$ constant. Such a simple arithmetic relationship is valid for nearly all gases at room temperature and at pressures equal to or less than one atmosphere. Subsequent work has shown that the relationship loses its validity at higher pressures, but more complicated expressions that more accurately match experimental results can be derived. The discovery and investigation of such chemical regularities, often called laws of nature, lie within the realm of physical chemistry. For much of the 18th century the source of mathematical regularity in chemical systems was assumed to be the continuum of forces and fields that surround the atoms making up chemical elements and compounds. Developments in the 20th century, however, have shown that chemical behaviour is best interpreted by a quantum mechanical model of atomic and molecular structure (see above *Quantum mechanics*). The branch of physical chemistry that is largely devoted to this subject is theoretical chemistry. Theoretical chemists make extensive use of computers to help them solve complicated mathematical equations. Other branches of physical chemistry include chemical thermodynamics, which deals with the relationship between heat and other forms of chemical energy, and chemical kinetics, which seeks to measure and understand the rates of chemical reactions.

Electrochemistry

Electrochemistry investigates the interrelationship of electric current and chemical change. The passage of an electric current through a chemical solution causes changes in the constituent substances that are often reversible—*i.e.*, under different conditions the altered substances themselves will yield an electric current. Common batteries contain chemical substances that, when placed in contact with each other by closing an electrical circuit, will deliver current at a constant voltage until the substances are consumed. At present there is much interest in devices that can use the energy in sunlight to drive chemical reactions whose products are capable of storing the energy. The discovery of such devices would make possible the widespread utilization of solar energy.

There are many other disciplines within physical chemistry that are concerned more with the general properties of substances and the interactions among substances than with the substances themselves. Photochemistry is a specialty that investigates the interaction of light with matter. Chemical reactions initiated by the absorption of light can be very different from those that occur by other means. Vitamin D, for example, is formed in the human body when the steroid ergosterol absorbs solar radiation; ergosterol does not change to vitamin D in the dark.

Surface chemistry

A rapidly developing subdiscipline of physical chemistry is surface chemistry. It examines the properties of chemical surfaces, relying heavily on instruments that can provide a chemical profile of such surfaces. Whenever a solid is exposed to a liquid or a gas, a reaction occurs initially on the surface of the solid, and its properties can change dramatically as a result. Aluminum is a case in point: it is resistant to corrosion precisely because the surface of the pure metal reacts with oxygen to form a layer of aluminum oxide, which serves to protect the interior of the metal from further oxidation. Numerous reaction catalysts perform their function by providing a reactive surface on which substances can react.

Industrial chemistry. The manufacture, sale, and distribution of chemical products is one of the cornerstones of a developed country. Chemists play an important role in the manufacture, inspection, and safe handling of chemical products, as well as in product development and general management. The manufacture of basic chemicals such as oxygen, chlorine, ammonia, and sulfuric acid provides the raw materials for industries producing textiles, agricultural products, metals, paints, and pulp and paper. Specialty chemicals are produced in smaller amounts for industries involved with such products as pharmaceuticals, foodstuffs, packaging, detergents, flavours, and fragrances. To a large extent, the chemical industry takes the products and reactions common to "bench-top" chemical processes and scales them up to industrial quantities.

The monitoring and control of bulk chemical processes, especially with regard to heat transfer, pose problems usually tackled by chemists and chemical engineers. The

disposal of by-products also is a major problem for bulk chemical producers. These and other challenges of industrial chemistry set it apart from the more purely intellectual disciplines of chemistry discussed above. Yet, within the chemical industry, there is a considerable amount of fundamental research undertaken within traditional specialties. Most large chemical companies have research-and-development capability. Pharmaceutical firms, for example, operate large research laboratories in which chemists test molecules for pharmacological activity. The new products and processes that are discovered in such laboratories are often patented and become a source of profit for the company funding the research. A great deal of the research conducted in the chemical industry can be termed applied research because its goals are closely tied to the products and processes of the company concerned. New technologies often require much chemical expertise. The fabrication of, say, electronic microcircuits involves close to 100 separate chemical steps from start to finish. Thus, the chemical industry evolves with the technological advances of the modern world and at the same time often contributes to the rate of progress.

Applied research

THE METHODOLOGY OF CHEMISTRY

Chemistry is to a large extent a cumulative science. Over time the number and extent of observations and phenomena studied increase. Not all hypotheses and discoveries endure unchallenged, however. Some of them are discarded as new observations or more satisfying explanations appear. Nonetheless, chemistry has a broad spectrum of explanatory models for chemical phenomena that have endured and been extended over time. These now have the status of theories, interconnected sets of explanatory devices that correlate well with observed phenomena. As new discoveries are made, they are incorporated into existing theory whenever possible. However, as the discovery of high-temperature superconductors in 1986 illustrates, accepted theory is never sufficient to predict the course of future discovery. Serendipity, or chance discovery, will continue to play as much a role in the future as will theoretical sophistication.

Studies of molecular structure. The chemical properties of a substance are a function of its structure, and the techniques of X-ray crystallography now enable chemists to determine the precise atomic arrangement of complex molecules. A molecule is an ordered assembly of atoms. Each atom in a molecule is connected to one or more neighbouring atoms by a chemical bond. The length of bonds and the angles between adjacent bonds are all important in describing molecular structure, and a comprehensive theory of chemical bonding is one of the major achievements of modern chemistry. Fundamental to bonding theory is the atomic–molecular concept.

Basic concepts of modern bonding theory

Atoms and elements. As far as general chemistry is concerned, atoms are composed of the three fundamental particles: the proton, the neutron, and the electron. Although the proton and the neutron are themselves composed of smaller units, their substructure has little impact on chemical transformation. As was explained in an earlier section, the proton carries a charge of +1, and the number of protons in an atomic nucleus distinguishes one type of chemical atom from another. The simplest atom of all, hydrogen, has a nucleus composed of a single proton. The neutron has very nearly the same mass as the proton, but it has no charge. Neutrons are contained with protons in the nucleus of all atoms other than hydrogen. The atom with one proton and one neutron in its nucleus is called deuterium. Because it has only one proton, deuterium exhibits the same chemical properties as hydrogen but has a different mass. Hydrogen and deuterium are examples of related atoms called isotopes. The third atomic particle, the electron, has a charge of −1, but its mass is 1,836 times smaller than that of a proton. The electron occupies a region of space outside the nucleus termed an orbital. Some orbitals are spherical with the nucleus at the centre. Because electrons have so little mass and move about at speeds close to half that of light, they exhibit the same wave–particle duality as photons of light. This means that some of the properties of an electron are best described by considering the electron to be a particle, while other properties are consistent with the behaviour of a standing wave. The energy of a standing wave, such as a vibrating string, is distributed over the region of space defined by the two fixed ends and the up-and-down extremes of vibration. Such a wave does not exist in a fixed region of space as does a particle. Early models of atomic structure envisioned the electron as a particle orbiting the nucleus, but electron orbitals are now interpreted as the regions of space occupied by standing waves called wave functions (see above *Quantum mechanics*). These wave functions represent the regions of space around the nucleus in which the probability of finding an electron is high. They play an important role in bonding theory, as will be discussed later.

Concept of wave functions

Each proton in an atomic nucleus requires an electron for electrical neutrality. Thus, as the number of protons in a nucleus increases, so too does the number of electrons. The electrons, alone or in pairs, occupy orbitals increasingly distant from the nucleus. Electrons farther from the nucleus are attracted less strongly by the protons in the nucleus, and they can be removed more easily from the atom. The energy required to move an electron from one orbital to another, or from one orbital to free space, gives a measure of the energy level of the orbitals. These energies have been found to have distinct, fixed values; they are said to be quantized. The energy differences between orbitals give rise to the characteristic patterns of light absorption or emission that are unique to each chemical atom.

A new chemical atom—that is, an element—results each time another proton is added to an atomic nucleus. Consecutive addition of protons generates the whole range of elements known to exist in the universe. Compounds are formed when two or more different elements combine through atomic bonding. Such bond formation is a consequence of electron pairing and constitutes the foundation of all structural chemistry.

Ionic and covalent bonding. When two different atoms approach each other, the electrons in their outer orbitals can respond in two distinct ways. An electron in the outermost atomic orbital of atom A may move completely to an outer but stabler orbital of atom B. The charged atoms that result, A^+ and B^-, are called ions, and the electrostatic force of attraction between them gives rise to what is termed an ionic bond. Most elements can form ionic bonds, and the substances that result commonly exist as three-dimensional arrays of positive and negative ions. Ionic compounds are frequently crystalline solids that have high melting points (*e.g.*, table salt).

The second way in which the two outer electrons of atoms A and B can respond to the approach of A and B is to pair up to form a covalent bond. In the simple view known as the valence-bond model, in which electrons are treated strictly as particles, the two paired electrons are assumed to lie between the two nuclei and are shared equally by atoms A and B, resulting in a covalent bond. Atoms joined together by one or more covalent bonds constitute molecules. Hydrogen gas is composed of hydrogen molecules, which consist in turn of two hydrogen atoms linked by a covalent bond. The notation H_2 for hydrogen gas is referred to as a molecular formula. Molecular formulas indicate the number and type of atoms that make up a molecule. The molecule H_2 is responsible for the properties generally associated with hydrogen gas. Most substances on Earth have covalently bonded molecules as their fundamental chemical unit, and their molecular properties are completely different from those of the constituent elements. The physical and chemical properties of carbon dioxide, for example, are quite distinct from those of pure carbon and pure oxygen.

Valence-bond model

The interpretation of a covalent bond as a localized electron pair is an oversimplification of the bonding situation. A more comprehensive description of bonding that considers the wave properties of electrons is the molecular-orbital theory. According to this theory, electrons in a molecule, rather than being localized between atoms, are distributed over all the atoms in the molecule in a spatial distribution described by a molecular orbital. Such orbitals result when the atomic orbitals of bonded atoms combine with each other. The total number of molecular

The molecular-orbital theory

orbitals present in a molecule is equal to the sum of all atomic orbitals in the constituent atoms prior to bonding. Thus, for the simple combination of atoms A and B to form the molecule AB, two atomic orbitals combine to generate two molecular orbitals. One of these, the so-called bonding molecular orbital, represents a region of space enveloping both the A and B atoms, while the other, the anti-bonding molecular orbital, has two lobes, neither of which occupies the space between the two atoms. The bonding molecular orbital is at a lower energy level than are the two atomic orbitals, while the anti-bonding orbital is at a higher energy level. The two paired electrons that constitute the covalent bond between A and B occupy the bonding molecular orbital. For this reason, there is a high probability of finding the electrons between A and B, but they can be found elsewhere in the orbital as well. Because only two electrons are involved in bond formation and both can be accommodated in the lower energy orbital, the anti-bonding orbital remains unpopulated. This theory of bonding predicts that bonding between A and B will occur because the energy of the paired electrons after bonding is less than that of the two electrons in their atomic orbitals prior to bonding. The formation of a covalent bond is thus energetically favoured. The system goes from a state of higher energy to one of lower energy.

Another feature of this bonding picture is that it is able to predict the energy required to move an electron from the bonding molecular orbital to the anti-bonding one. The energy required for such an electronic excitation can be provided by visible light, for example, and the wavelength of the light absorbed determines the colour displayed by the absorbing molecule (*e.g.*, violets are blue because the pigments in the flower absorb the red rays of natural light and reflect more of the blue). As the number of atoms in a molecule increases, so too does the number of molecular orbitals. Calculation of molecular orbitals for large molecules is mathematically difficult, but computers have made it possible to determine the wave equations for several large molecules. Molecular properties predicted by such calculations correlate well with experimental results.

Use of computers to determine wave equations

Isomerism. Many elements can form two or more covalent bonds, but only a few are able to form extended chains of covalent bonds. The outstanding example is carbon, which can form as many as four covalent bonds and can bond to itself indefinitely. Carbon has six electrons in total, two of which are paired in an atomic orbital closest to the nucleus. The remaining four are farther from the nucleus and are available for covalent bonding. When there is sufficient hydrogen present, carbon will react to form methane, CH_4. When all four electron pairs occupy the four molecular orbitals of lowest energy, the molecule assumes the shape of a tetrahedron, with carbon at the centre and the four hydrogen atoms at the apexes. The C–H bond length is 110 picometres (1 picometre = 10^{-12} metre), and the angle between adjacent C–H bonds is close to 110°. Such tetrahedral symmetry is common to many carbon compounds and results in interesting structural possibilities. If two carbon atoms are joined together, with three hydrogen atoms bonded to each carbon atom, the molecule ethane is obtained. When four carbon atoms are joined together, two different structures are possible: a linear structure designated *n*-butane and a branched structure called iso-butane. These two structures have the same molecular formula, C_4H_{10}, but a different order of attachment of their constituent atoms. The two molecules are termed structural isomers. Each of them has unique chemical and physical properties, and they are different compounds. The number of possible isomers increases rapidly as the number of carbon atoms increases. There are five isomers for C_6H_{14}, 75 for $C_{10}H_{22}$, and 6.2×10^{13} for $C_{80}H_{82}$. When carbon forms bonds to atoms other than hydrogen, such as oxygen, nitrogen, and sulfur, the structural possibilities become even greater. It is this great potential for structural diversity that makes carbon compounds essential to living organisms.

Structural isomers

Even when the bonding sequence of carbon compounds is fixed, further structural variation is still possible. When two carbon atoms are joined together by two bonding pairs of electrons, a double bond is formed. A double bond forces the two carbon atoms and attached groups into a rigid, planar structure. As a result, a molecule such as CHCl=CHCl can exist in two nonidentical forms called geometric isomers. Structural rigidity also occurs in ring structures, and attached groups can be on the same side of a ring or on different sides. Yet another opportunity for isomerism arises when a carbon atom is bonded to four different groups. These can be attached in two different ways, one of which is the mirror image of the other. This type of isomerism is called optical isomerism, because the two isomers affect plane-polarized light differently. Two optical isomers are possible for every carbon atom that is bonded to four different groups. For a molecule bearing 10 such carbon atoms, the total number of possible isomers will be $2^{10} = 1{,}024$. Large biomolecules often have 10 or more carbon atoms for which such optical isomers are possible. Only one of all the possible isomers will be identical to the natural molecule. For this reason, the laboratory synthesis of large organic molecules is exceedingly difficult. Only in the last few decades of the 20th century have chemists succeeded in developing reagents and processes that yield specific optical isomers. They expect that new synthetic methods will make possible the synthesis of ever more complex natural products.

Optical isomers

Investigations of chemical transformations. *Basic factors.* The structure of ionic substances and covalently bonded molecules largely determines their function. As noted above, the properties of a substance depend on the number and type of atoms it contains and on the bonding patterns present. Its bulk properties also depend, however, on the interactions among individual atoms, ions, or molecules. The force of attraction between the fundamental units of a substance dictate whether, at a given temperature and pressure, that substance will exist in the solid, liquid, or gas phase. At room temperature and pressure, for example, the strong forces of attraction between the positive ions of sodium (Na^+) and the negative ions of chlorine (Cl^-) draw them into a compact solid structure. The weaker forces of attraction among neighbouring water molecules allow the looser packing characteristic of a liquid. Finally, the very weak attractive forces acting among adjacent oxygen molecules are exceeded by the dispersive forces of heat; oxygen, consequently, is a gas. Interparticle forces thus affect the chemical and physical behaviour of substances, but they also determine to a large extent how a particle will respond to the approach of a different particle. If the two particles react with each other to form new particles, a chemical reaction has occurred. Notwithstanding the unlimited structural diversity allowed by molecular bonding, the world would be devoid of life if substances were incapable of change. The study of chemical transformation, which complements the study of molecular structure, is built on the concepts of energy and entropy.

Energy and the first law of thermodynamics. The concept of energy is a fundamental and familiar one in all the sciences. In simple terms, the energy of a body represents its ability to do work, and work itself is a force acting over a distance.

Chemical systems can have both kinetic energy (energy of motion) and potential energy (stored energy). The kinetic energy possessed by any collection of molecules in a solid, liquid, or gas is known as its thermal energy. Since liquids expand when they have more thermal energy, a liquid column of mercury, for example, will rise higher in an evacuated tube as it becomes warmer. In this way a thermometer can be used to measure the thermal energy, or temperature, of a system. The temperature at which all molecular motion comes to a halt is known as absolute zero.

The kinetic and potential energy of chemical systems

Energy also may be stored in atoms or molecules as potential energy. When protons and neutrons combine to form the nucleus of a certain element, the reduction in potential energy is matched by the production of a huge quantity of kinetic energy. Consider, for instance, the formation of the deuterium nucleus from one proton and one neutron. The fundamental mass unit of the chemist is the mole, which represents the mass, in grams, of 6.02×10^{23} individual particles, whether they be atoms or molecules.

One mole of protons has a mass of 1.007825 grams and one mole of neutrons has a mass of 1.008665 grams. By simple addition the mass of one mole of deuterium atoms (ignoring the negligible mass of one mole of electrons) should be 2.016490 grams. The measured mass is 0.00239 gram less than this. The missing mass is known as the binding energy of the nucleus and represents the mass equivalent of the energy released by nucleus formation. By using Einstein's formula for the conversion of mass to energy ($E = mc^2$), one can calculate the energy equivalent of 0.00239 gram as 2.15×10^8 kilojoules. This is approximately 240,000 times greater than the energy released by the combustion of one mole of methane. Such studies of the energetics of atom formation and interconversion are part of a specialty known as nuclear chemistry.

Concerns of nuclear chemistry

The energy released by the combustion of methane is about 900 kilojoules per mole. Although much less than the energy released by nuclear reactions, the energy given off by a chemical process such as combustion is great enough to be perceived as heat and light. Energy is released in so-called exothermic reactions because the chemical bonds in the product molecules, carbon dioxide and water, are stronger and stabler than those in the reactant molecules, methane and oxygen. The chemical potential energy of the system has decreased, and most of the released energy appears as heat, while some appears as radiant energy, or light. The heat produced by such a combustion reaction will raise the temperature of the surrounding air and, at constant pressure, increase its volume. This expansion of air results in work being done. In the cylinder of an internal-combustion engine, for example, the combustion of gasoline results in hot gases that expand against a moving piston. The motion of the piston turns a crankshaft, which then propels the vehicle. In this case, chemical potential energy has been converted to thermal energy, some of which produces useful work. This process illustrates a statement of the conservation of energy known as the first law of thermodynamics. This law states that, for an exothermic reaction, the energy released by the chemical system is equal to the heat gained by the surroundings plus the work performed. By measuring the heat and work quantities that accompany chemical reactions, it is possible to ascertain the energy differences between the reactants and the products of various reactions. In this manner, the potential energy stored in a variety of molecules can be determined, and the energy changes that accompany chemical reactions can be calculated.

Measuring the heat and work quantities that accompany reactions

Entropy and the second law of thermodynamics. Some chemical processes occur even though there is no net energy change. Consider a vessel containing a gas, connected to an evacuated vessel via a channel wherein a barrier obstructs passage of the gas. If the barrier is removed, the gas will expand into the evacuated vessel. This expansion is consistent with the observation that a gas always expands to fill the volume available. When the temperature of both vessels is the same, the energy of the gas before and after the expansion is the same. The reverse reaction does not occur, however. The spontaneous reaction is the one that yields a state of greater disorder. In the expanded volume, the individual gas molecules have greater freedom of movement and thus are more disordered. The measure of the disorder of a system is a quantity termed entropy. At a temperature of absolute zero, all movement of atoms and molecules ceases, and the disorder—and entropy—of such perfectly compacted substances is zero. (Zero entropy at zero temperature is in accord with the third law of thermodynamics.) All substances above absolute zero will have a positive entropy value that increases with temperature. When a hot body cools down, the thermal energy it loses passes to the surrounding air, which is at a lower temperature. As the entropy of the cooling body decreases, the entropy of the surrounding air increases. In fact, the increase in entropy of the air is greater than the decrease in entropy of the cooling body. This is consistent with the second law, which states that the total entropy of a system and its surroundings always increases in a spontaneous reaction. Thus the first and second laws of thermodynamics indicate that, for all processes of chemical change throughout the universe, energy is conserved but entropy increases.

Application of the laws of thermodynamics to chemical systems allows chemists to predict the behaviour of chemical reactions. When energy and entropy considerations favour the formation of product molecules, reagent molecules will act to form products until an equilibrium is established between products and reagents. The ratio of products to reagents is specified by a quantity known as an equilibrium constant, which is a function of the energy and entropy differences between the two. What thermodynamics cannot predict, however, is the rate at which chemical reactions occur. For fast reactions an equilibrium mixture of products and reagents can be established in one millisecond or less; for slow reactions the time required could be hundreds of years.

Predicting the behaviour of chemical reactions

Rates of reaction. When the specific rates of chemical reactions are measured experimentally, they are found to be dependent on the concentrations of reacting species, temperature, and a quantity called activation energy. Chemists explain this phenomenon by recourse to the collision theory of reaction rates. This theory builds on the premise that a reaction between two or more chemicals requires, at the molecular level, a collision between two rapidly moving molecules. If the two molecules collide in the right way and with enough kinetic energy, one of the molecules may acquire enough energy to initiate the bond-breaking process. As this occurs, new bonds may begin to form, and ultimately reagent molecules are converted into product molecules. The point of highest energy during bond breaking and bond formation is called the transition state of the molecular process. The difference between the energy of the transition state and that of the reacting molecules is the activation energy that must be exceeded for a reaction to occur. Reaction rates increase with temperature because the colliding molecules have greater energies, and more of them will have energies that exceed the activation energy of reaction. The modern study of the molecular basis of chemical change has been greatly aided by lasers and computers. It is now possible to study short-lived collision products and to better determine the molecular mechanisms that fix the rate of chemical reactions. This knowledge is useful in designing new catalysts that can accelerate the rate of reaction by lowering the activation energy. Catalysts are important for many biochemical and industrial processes because they speed up reactions that ordinarily occur too slowly to be useful. Moreover, they often do so with increased control over the structural features of the product molecules. A rhodium phosphine catalyst, for example, has enabled chemists to obtain 96 percent of the correct optical isomer in a key step in the synthesis of L-dopa, a drug used for treating Parkinson's disease.

Collision theory of reaction rates

Studying the molecular basis of chemical change with lasers and computers

CHEMISTRY AND SOCIETY

For the first two-thirds of the 20th century, chemistry was seen by many as the science of the future. The potential of chemical products for enriching society appeared to be unlimited. Increasingly, however, and especially in the public mind, the negative aspects of chemistry have come to the fore. Disposal of chemical by-products at waste-disposal sites of limited capacity has resulted in environmental and health problems of enormous concern. The legitimate use of drugs for the medically supervised treatment of diseases has been tainted by the growing misuse of mood-altering drugs. The very word chemicals has come to be used all too frequently in a pejorative sense. There is, as a result, a danger that the pursuit and application of chemical knowledge may be seen as bearing risks that outweigh the benefits.

It is easy to underestimate the central role of chemistry in modern society, but chemical products are essential if the world's population is to be clothed, housed, and fed. The world's reserves of fossil fuels (*e.g.*, oil, natural gas, and coal) will eventually be exhausted, some as soon as the 21st century, and new chemical processes and materials will provide a crucial alternative energy source. The conversion of solar energy to more concentrated, useful forms, for example, will rely heavily on discoveries in chemistry.

Practical contributions of chemistry

Long-term, environmentally acceptable solutions to pollution problems are not attainable without chemical knowledge. There is much truth in the aphorism that "chemical problems require chemical solutions." Chemical inquiry will lead to a better understanding of the behaviour of both natural and synthetic materials and to the discovery of new substances that will help future generations better supply their needs and deal with their problems.

Progress in chemistry can no longer be measured only in terms of economics and utility. The discovery and manufacture of new chemical goods must continue to be economically feasible but must be environmentally acceptable as well. The impact of new substances on the environment can now be assessed before large-scale production begins, and environmental compatibility has become a valued property of new materials. For example, compounds consisting of carbon fully bonded to chlorine and fluorine, called chlorofluorocarbons (or Freons), were believed to be ideal for their intended use when they were first discovered. They are nontoxic, nonflammable gases and volatile liquids that are very stable. These properties led to their widespread use as solvents, refrigerants, and propellants in aerosol containers. Time has shown, however, that these compounds decompose in the upper regions of the atmosphere and that the decomposition products act to destroy stratospheric ozone. Limits have now been placed on the use of chlorofluorocarbons, but it is impossible to recover the amounts already dispersed into the atmosphere.

The chlorofluorocarbon problem illustrates how difficult it is to anticipate the overall impact that new materials can have on the environment. Chemists are working to develop methods of assessment, and prevailing chemical theory provides the working tools. Once a substance has been identified as hazardous to the existing ecological balance, it is the responsibility of chemists to locate that substance and neutralize it, limiting the damage it can do or removing it from the environment entirely. The last years of the 20th century will see many new, exciting discoveries in the processes and products of chemistry. Inevitably, the harmful effects of some substances will outweigh their benefits, and their use will have to be limited. Yet, the positive impact of chemistry on society as a whole seems beyond doubt. (M.C.U.)

Astronomy

Astronomy comprises the study of all extraterrestrial objects. Until the invention of the telescope and the discovery of the laws of motion and gravity in the 17th century, astronomy was primarily concerned with noting and predicting the positions of the Sun, Moon, and planets, initially for calendrical and astrological purposes and later for navigational applications and scientific interest. The catalog of objects now studied is much broader and includes, in order of increasing distances, the solar system, the stars that make up the Milky Way Galaxy, and other more distant stellar objects and galaxies. With the advent of scientific space probes, the Earth also has come to be studied as one of the planets, though its more detailed investigation remains the domain of the geologic sciences.

THE SCOPE OF ASTRONOMY

Astrophysics and cosmology as major fields of study

During the 20th century astronomy has expanded to include astrophysics, the application of physical and chemical knowledge to an understanding of the nature of celestial objects and the physical processes that control their formation, evolution, and emission of radiation. Study of the nuclear reactions that provide the energy radiated by stars has shown how the diversity of atoms found in nature can be derived from a universe that originally consisted exclusively of hydrogen. Concerned with phenomena on the largest scale is cosmology, the study of the evolution of the universe. Astrophysics has transformed cosmology from an almost purely speculative activity to a modern science capable of predictions that can be tested.

Its great advances notwithstanding, astronomy is still subject to a major constraint; it is inherently an observational rather than an experimental science. Almost all measurements must be performed at great distances from the objects of interest, with no control over such quantities as their temperature, pressure, or chemical composition. There are a few exceptions to this limitation—namely, meteorites, rock and soil samples brought back from the Moon, and interplanetary dust particles collected in or above the stratosphere. These can be examined with laboratory techniques to provide information that cannot be obtained in any other way. Future space missions may yield samples of comet-tail dust or surface materials from Mars, but much of astronomy appears otherwise confined to Earth-based observations.

Determining astronomical distances. A central problem in astronomy is the determination of distances. Without a knowledge of distances, sizes would remain nothing more than angular diameters, and stellar brightness could not be converted into true radiated power or luminosity. Astronomical distance measurement began with a knowledge of the Earth's diameter, which provided a base for triangulation. Within the inner solar system, some distances can now be better determined through the timing of radar reflections. For the outer planets, triangulation is still used. Beyond the solar system, distances to the closest stars are determined through triangulation, with the diameter of the Earth's orbit serving as the baseline and shifts in stellar parallax being the measured quantities. Stellar distances are commonly expressed by astronomers in parsecs (pc; 1 pc = 3.086×10^{18} centimetres) or in kiloparsecs. They can be measured out to about 50 parsecs by trigonometric parallax (see STARS AND STAR CLUSTERS: *Determining stellar distances*). Less direct measurements must be used for more distant stars and galaxies.

Use of radar and triangulation

Only two general methods for galactic distances are described here. In the first, a clearly identifiable type of star is used as a reference standard because its luminosity (total radiated power) has been well determined. Such a star is termed a "standard candle." The distance to a standard candle can be calculated from its known luminosity and its measured intensity. Identification of the standard candle can be made through its spectrum or regular variations in brightness. (Corrections may have to be made for the absorption of starlight over great distances.) This method forms the basis of measurements of distances to the closest galaxies. It has been found that such distances generally correlate with the speeds of recession (as determined from Doppler-shifted wavelengths; see above). This correlation is expressed in the Hubble law, velocity = $H \times$ distance, with H denoting Hubble's constant, the best value of which is thought to lie between 50 and 100 kilometres per second per megaparsec (km/sec/Mpc). It is currently used to determine distances to remote galaxies in which standard candles have not been found. Application of the Hubble law, however, has been questioned as it relates to some quasars (energetic galactic nuclei), and the value of Hubble's constant cited above is still a subject of disagreement.

Methods for determining distances to galaxies

Study of the solar system. The solar system took shape 4,600,000,000 years ago, when it condensed within a large cloud of gas and dust. Gravitational attraction holds the planets in their elliptical orbits around the Sun. Besides the Earth, five major planets (Mercury, Venus, Mars, Jupiter, and Saturn) have been known from ancient times. Since then, only three others have been discovered: Uranus by accident in 1781, and Neptune and Pluto in 1846 and 1930, respectively, after deliberate searches.

The average Earth–Sun distance was originally defined as the astronomical unit (a.u.) and provides a convenient measure for distances within the solar system. The astronomical unit is now defined dynamically (using Kepler's third law), and it has the value $1.49597870 \times 10^{13}$ centimetres. The semimajor axis of the Earth's orbit is $1 + (3.1 \times 10^{-8})$ a.u. Mercury, at 0.39 a.u., is the closest planet to the Sun, while Pluto, at 39.5 a.u., is the farthest. The planes of the planetary orbits (other than that of Pluto) are all within a few degrees of the ecliptic, the plane that contains the Earth's orbit around the Sun. As viewed from far above the Earth's North Pole, all planets move in the same (counterclockwise) direction in their orbits.

All of the planets, apart from those closest to the

Sun (Mercury and Venus), have satellites very diverse in appearance, size, and structure, as revealed through closeup observations from long-range space probes. Four planets—Jupiter, Saturn, Neptune, and Uranus—have rings consisting of small rocks and particles that are confined to disklike systems as they orbit their parent planets.

Most of the mass of the solar system is concentrated in the Sun, with its 1.99×10^{33} grams. Together, all of the planets amount to about 2.7×10^{30} grams, with Jupiter alone accounting for 71 percent of this. The solar system also contains a very large number of much smaller objects. In order of decreasing size, these are the asteroids (also called minor planets), comets, meteoroids, and dust particles.

Planetary data

The four terrestrial planets, Mercury, Venus, Earth, and Mars, along with the Moon, have average densities in the range 3.9–5.5 grams per cubic centimetre (g/cm³), setting them apart from the outer planets, whose densities are all close to 1 g/cm³, the density of water. The compositions of these two groups of planets must therefore be significantly different. This is probably attributable to the conditions that prevailed during the early development of the solar system (see below *Theories of origin*). Planetary temperatures now range from about 500° C on Mercury's surface through the typical 20° C on Earth to −135° C on Jupiter and down to −230° C on Pluto.

The surfaces of the terrestrial planets and many satellites show extensive cratering produced by high-speed impacts. On Earth, with its large quantities of water and an active atmosphere, many of these cosmic footprints have eroded, but remnants of very large craters can be seen in satellite and aerial photographs of the terrestrial surface. On Mercury, Mars, and the Moon, the absence of water and any significant atmosphere has left the craters unchanged for billions of years, apart from disturbances produced by infrequent later impacts. Cratering on the largest scale seems to have ceased about 3,000,000,000 years ago, but there is clear evidence for a continued cosmic drizzle of small particles, with the larger objects churning (gardening) the lunar surface and the smallest producing microscopic impact pits in crystals in the lunar rocks.

Lunar exploration. During the U.S. Apollo missions, a total sample weight of 381 kilograms was collected; 300 grams of lunar material also was returned by three unmanned Soviet Luna space vehicles. Less than 10 percent of the samples has so far been distributed for analysis, but planetary science has been revolutionized by these expeditions. A wide range of laboratory techniques has been employed on the lunar samples. The results of the analysis have enabled investigators to determine the composition and age of the lunar surface. Seismic techniques have made it possible to probe the lunar interior. In addition, a retroreflector left on the Moon's surface by Apollo astronauts returns a high-power laser beam emitted from the Earth, enabling researchers to monitor on a regular basis the Earth–Moon distance to an accuracy of a few centimetres. This experiment provides data that can be used in calculations of the dynamics of the Earth–Moon system.

Significance of the lunar missions

Planetary studies. Mercury is too hot to retain an atmosphere, but Venus' brilliant white appearance is the result of its being completely enveloped by thick clouds of carbon dioxide. Below the upper clouds it has a hostile atmosphere containing clouds of sulfuric acid droplets. The cloud cover shields the planet's surface from direct sunlight, but the energy that does filter through warms the surface, which then radiates at infrared wavelengths. The long waves of infrared radiation are trapped by the dense clouds, resulting in a very high surface temperature of almost 480° C. Radar can penetrate the thick Venusian clouds and has been used to map the planet's surface. The Martian atmosphere is very thin, only about 0.006 that of the Earth, and composed mostly of carbon dioxide (95 percent), with very little water vapour. The outer planets have atmospheres composed largely of light gases. For example, hydrogen and helium, along with some methane and ammonia, have been detected on Jupiter.

Each of the planets rotates on its axis, and nearly all of them rotate in the same (counterclockwise) direction, as viewed from above the ecliptic. The two exceptions are Venus, which rotates in the clockwise direction beneath its cloud cover, and Uranus, which has its rotational axis very nearly in the plane of the ecliptic.

Discovery of magnetic fields

Some of the planets have magnetic fields. The Earth's field extends outward until it is disturbed by the solar wind, an outward flow of protons and electrons from the Sun that carries a magnetic field along with it. Through processes not yet fully understood, protons and electrons from the solar wind and cosmic rays populate two doughnut-shaped regions called the Van Allen radiation belts, the inner of which extends from about 1,000 to 5,000 kilometres above the Earth's surface and the outer from roughly 15,000 to 25,000 kilometres. In these belts, trapped particles spiral along paths that take them around the Earth while bouncing back and forth between the Northern and Southern hemispheres. During periods of increased solar activity, these regions of trapped particles are disturbed, and some of the particles move down into the Earth's atmosphere where they collide with atoms and molecules to produce auroras.

Jupiter has a stronger magnetic field than the Earth's and many more trapped electrons, whose synchrotron radiation (electromagnetic radiation of the kind produced by electrons in accelerators called synchrotrons) is detectable from Earth. Saturn has a magnetic field that is not quite as strong as Jupiter's, but it, too, has a region of trapped particles. Mercury has a weak magnetic field that is only about 1 percent as strong as the Earth's and contains no discernible trapped particles. No magnetic field has been detected around any of the other planets.

Investigations of the minor bodies. Approximately 3,500 asteroids have now been identified. Most have orbits close to the ecliptic and move in the asteroid belt located between 2.3 and 3.3 a.u. from the Sun. Only about 250 of these objects are larger than 100 kilometres, and their total mass is thought to be roughly $\frac{1}{2,000}$ that of the Earth.

Asteroids, comets, and meteoroids

Comets are considered to come from a vast reservoir, the Oort cloud, which orbits the Sun at distances of 30,000 to 100,000 a.u. More than 600 comets have so far been discovered. Most make only a single pass through the inner solar system, but some are deflected by Jupiter or Saturn into orbits that allow them to return at predictable times. Comet Halley is the best known of these periodic comets, with its next return predicted for AD 2060. About 30 comets have periods of less than 100 years. Comet masses have not been well determined, but most are thought to be less than 10^{18} grams, or 1,000,000,000 times smaller than the Earth.

Even smaller than comets are the meteoroids, lumps of stony material. Meteoroids vary in size from small rocks to large boulders weighing a ton or more. A few have orbits that bring them into the Earth's atmosphere and down to the ground as meteorites. These are classified into three broad groups: stony or chondrites (about 93 percent), iron (5.7 percent), and stony-iron (1.5 percent). Smaller meteoroids that enter the atmosphere may heat up sufficiently to vaporize and appear as meteors. Many, perhaps most, of the meteors occur in showers and follow orbits that seem to be identical with those of certain comets, thus pointing to a cometary origin. For example, each May the Earth crosses the orbit of Comet Halley, and the Eta Aquarid meteor shower becomes visible. Micrometeorites, the smallest meteoroidal particles, can be detected from Earth-orbiting satellites or be sufficiently slowed by atmospheric friction to be collected by specially equipped aircraft flying in the stratosphere and returned for laboratory inspection.

Measurement of radioisotope abundance

Determinations of age and chemical composition. The age of the solar system, about 4,500,000,000 years, has been derived from measurements of radioactivity in meteorites, lunar samples, and the Earth's crust. Abundances of the isotopes of uranium, thorium, and rubidium and their decay products, lead and strontium, are the measured quantities.

Assessment of the chemical composition of the solar system is based on data from the Earth, Moon, and meteorites, as well as on the spectral analysis of light from the Sun and planets. In broad outline, the solar system abundances of the elements decrease with increasing atomic

weight. Hydrogen atoms are by far the most abundant, with 93 percent; helium is next, with 6.7 percent; and all other types of atoms together amount to only 2.3 percent.

Theories of origin. The origin of the Earth, Moon, and solar system as a whole is a problem that has not yet been settled in detail. The Sun probably formed by condensation of the central region of a large cloud of gas and dust, with the planets and other solar-system bodies forming soon after, their composition strongly influenced by the temperature and density gradients in the evolving solar nebula. Less-volatile materials could condense into solids relatively close to the Sun to form the terrestrial planets. The abundant, volatile lighter elements could condense only at much greater distances.

The origin of the planetary satellites is not settled. There is still the question as to the origin of the Moon, and professional opinion has been oscillating between theories that see its origin and condensation simultaneous with the formation of the Earth, to an explanation in terms of a large impact on the Earth resulting in the expulsion of material that subsequently formed the Moon. For the outer planets with their multiple satellites, many very small and quite unlike one another, the picture is even less clear. Some of the objects have icy surfaces, while others are heavily cratered, and at least one, Jupiter's Io, is volcanic. Some of the satellites may have formed along with their parent planets, and others may have formed elsewhere and been captured. (For an in-depth treatment of the solar system and its components, see SOLAR SYSTEM.)

Study of the stars. *Measuring observable stellar properties.* The measurable quantities in stellar astrophyics include the externally observable features of the stars: distance, temperature spectrum and luminosity, composition (of the outer layers), diameter, mass, and variability in any of these. Theoretical astrophysicists use these observations to model the structure of stars and to devise theories for their formation and evolution.

Basis for theoretical work

In a system dating back at least to the Greek astronomer-mathematician Hipparchus in the 2nd century BC, apparent stellar brightness (m) is measured in magnitudes. Magnitudes are now defined so that a first-magnitude star is 100 times brighter than a star of sixth magnitude. The human eye cannot see stars fainter than about sixth magnitude, but modern instruments used with large telescopes can record stars as faint as about 26th magnitude. By convention, the absolute magnitude (M) is the magnitude that a star would appear to have if it were located at a standard distance of 10 parsecs. These quantities are related through the expression: $m - M = 5 \log_{10} r - 5$, where r is the star's distance in parsecs.

The magnitude scale is anchored on a group of standard stars. An absolute measure of radiant power is luminosity, usually expressed in ergs per second. This can be calculated when m and r are known. Correction might be necessary for the interstellar absorption of starlight.

Methods of measuring stellar diameter

There are several methods of measuring a star's diameter. From the brightness and distance, the luminosity (L) can be calculated; and from observations of the brightness at different wavelengths, the temperature (T) can be calculated. As the radiation from many stars can be well approximated by a Planck blackbody spectrum, these measured quantities can be related through the expression $L = 4\pi R^2 \sigma T^4$, thus providing a means of calculating R, the star's radius. In this expression, σ is Stefan's constant, 5.67×10^{-5} erg/cm^{2}deg^{4}sec. (The radius R refers to the star's photosphere, the region where the star becomes effectively opaque to outside observation.) Stellar angular diameters can be measured through interference effects. Alternatively, it is possible to monitor the intensity of the starlight during occultation by the Moon, which produces diffraction fringes whose pattern depends on the angular diameter of the star. Stellar angular diameters of several milliarcseconds can be measured, but so far only for relatively bright and close stars.

Many stars occur in binary systems, with the two partners in orbits around their mutual centre of mass. Such a system provides the best measurement of stellar masses. The period (P) of a binary system is related to the masses of the two stars ($m_1 + m_2$), and the orbital semimajor axis a via Kepler's third law: $P^2 = 4\pi^2 a^3 / G[m_1 + m_2]$. ($G$ is, as before, the universal gravitational constant.) From diameters and masses, average values of the stellar density can be calculated, and thence the central pressure. With the assumption of an equation of state, the central temperature can then be calculated. Thus, for example, in the Sun the central density is 158 g/cm^3, and the pressure is calculated to be more than 1,000,000,000 atmospheres and the temperature about 15,000,000 K. At this temperature, all atoms are ionized, and so the solar interior consists of a plasma, an ionized gas, with hydrogen and helium nuclei and electrons as major constituents. A small fraction of the hydrogen nuclei possess such high speeds that, upon colliding, their electrostatic repulsion is overcome, resulting in the formation by fusion of helium nuclei and a release of energy. Some of this energy is carried away by neutrinos, but most of it is carried by photons to the surface of the Sun to maintain its luminosity.

Determining stellar mass

Other stars, both more and less massive than the Sun, have broadly similar structures, but the size, central pressure and temperature, and fusion rate are functions of a star's mass and composition. Stars and their internal fusion (and resultant luminosity) are held stable against collapse through a delicate balance between the inward pressure produced by gravitational attraction and the outward pressure supplied by the photons produced in the fusion reactions.

Use of the Hertzsprung–Russell diagram

Stars that are in this condition of hydrostatic equilibrium are termed main-sequence stars, and they occupy a well-defined band on the Hertzsprung–Russell (H–R) diagram, in which luminosity is plotted against colour index or temperature. Temperature is deduced from broad-band spectral measurements in several standard wavelength intervals. Measurement of apparent magnitudes in the B and V bands (centred on 4350 and 5550 angstroms [Å], respectively) permits calculation of the colour index, $CI = m_B - m_V$, from which the temperature can be calculated.

For a given temperature, there are stars that have luminosity much greater than main-sequence stars. Given the R^2T^4 dependence of luminosity, greater luminosity implies a larger radius, and such stars are called giants or supergiants. Conversely, stars with a luminosity much less than that of main-sequence stars of the same temperature must be smaller and are termed dwarfs. White dwarfs are stars with temperatures that typically range from 10,000 to 12,000 K, and they appear visually as white or blue-white.

Spectral classification

Spectral classification, based initially on the colour index, has major spectral types O, B, A, F, G, K, and M, each subdivided into 10 parts. The strength of spectral lines of the more abundant elements in a star's atmosphere allows additional subdivisions within a class. Thus the Sun, a main-sequence star, is classified as G2 V, where the V denotes main sequence. Betelgeuse, a red supergiant with a surface temperature about half that of the Sun but with a luminosity of about 10,000 solar units, is classified as M2 Iab. (For more specific information about stellar spectra and bulk stellar properties, see STARS AND STAR CLUSTERS.)

Stellar formation and evolution. The range of physically allowable masses for stars is very narrow. If the star's mass is too small, the central temperature will be too low to sustain fusion reactions. The theoretical minimum stellar mass is about 0.08 solar mass. An upper theoretical limit of approximately 100 solar masses has been suggested, but this value is not firmly defined. Stars as massive as this will have luminosities about 1,000,000 times greater than that of the Sun.

Prevailing conception of star formation and evolution

A general model of star formation and evolution has been developed, and the major features seem to be established. This model shows that a large cloud of gas and dust can contract under its own gravitational attraction if its temperature is sufficiently low. Gravitational energy is released, and the contracting central material heats up until a point is reached where the outward radiation pressure balances the inward gravitational pressure and contraction ceases. Fusion reactions take over as the star's primary source of energy, and the star is then on the main sequence. The time required to pass through these

formative stages and onto the main sequence is less than 100,000,000 years for a star with as much mass as the Sun. It takes longer for less massive stars and a much shorter time for those much more massive.

Once a star has reached its main-sequence stage, it evolves relatively slowly, with the fusion of the hydrogen nuclei in its core to form helium nuclei. Continued fusion not only releases the energy that is radiated but also involves nucleosynthesis, the production of heavier nuclei.

Computer modeling of stellar evolution

Stellar evolution has been followed through computer modeling because the time scales for most stages are generally too extended for measurable changes to be seen even over a period of many years. One exception is the supernova, where major changes have been seen on a scale of hours, and computations have shown that some changes take place on a scale of milliseconds. Calculation of nucleosynthesis during stable and explosive stellar stages has yielded abundances of nuclei in general agreement with observed abundances.

The nature of the terminal products of stellar evolution depend on stellar mass. Some stars pass through an unstable stage in which their dimensions, temperature, and luminosity change cyclically over periods of hours or days. These so-called Cepheid variables serve as standard candles for distance measurements. Some stars may blow off their outer layers to produce planetary nebulas. The expanding material can be seen glowing in a thin shell as it disperses into the interstellar medium, while the remnant core, initially with a surface temperature of as high as 100,000 K, cools to become a white dwarf. The maximum stellar mass that can exist as a white dwarf is about 1.4 solar masses and is called the Chandrasekhar limit. More massive stars may end up as either neutron stars or black holes (see below).

The average density of a white dwarf is calculated to exceed 1,000,000 g/cm^3. Further compression is limited by a quantum condition called degeneracy, in which only certain energies are allowed for electrons. Under sufficiently great pressure, the electrons combine with protons to form neutrons. The resulting neutron star will have a density in the range 10^{14}–10^{15} g/cm^3, which is comparable to the density within atomic nuclei. The behaviour of large masses having nuclear densities is not yet sufficiently understood to be able to set a limit on the maximum size of a neutron star, but it is thought to be in the region of five solar masses.

Still more massive remnants of stellar evolution would have smaller dimensions and would be even denser than neutron stars. Such remnants are conceived to be black holes, objects so compact that no radiation can escape from their gravitational force within a characteristic distance called the Schwarzschild radius. This critical dimension is defined by $R_S = 2GM/c^2$. (Here, R_S is the Schwarzschild radius, G is the gravitational constant, M is the object's mass, and c is the speed of light.) For an object of three solar masses, the R_S would be about three kilometres.

Although no light can be detected from a black hole, the presence of such an object may be manifested in the effects of its gravitational field, as, for example, in a binary star system. If a black hole is a member of a double star, it may pull matter from its companion—a normal visible star—toward itself. This matter is accelerated as it approaches the black hole and becomes so intensely heated that it radiates large amounts of X rays from the periphery of the black hole outside the Schwarzschild radius.

Search for black holes

A few candidates for black holes have been found (*e.g.*, the X-ray source Cygnus X-1). Each of them has an estimated mass clearly exceeding that allowable for a neutron star, a factor crucial in the identification of possible black holes. (For further information on black holes, as well as on stellar evolution, see STARS AND STAR CLUSTERS; COSMOS: *The end states of stars: Black-hole model for active galactic nuclei.*)

Whereas the detection of black holes still needs to be confirmed, the existence of neutron stars was proved in 1968 when they were identified with the then newly discovered pulsars, objects characterized by the emission of radiation at short and extremely regular intervals generally between one and 1,000 pulses per second and stable to better than one part per billion. Pulsars are regarded as rotating neutron stars, the remnants of some supernovas.

Study of the Milky Way Galaxy. *General properties.* Stars are not distributed randomly throughout space. Many stars occur in systems consisting of two or three members with separations of less than 1,000 a.u. On a larger scale, clusters may contain many thousands of stars. Galaxies are much larger systems of stars and usually include clouds of gas and dust.

The solar system is located within the Milky Way Galaxy close to the equatorial plane and about 8.7 kiloparsecs (kpc) from the galactic centre. The galactic diameter is about 30 kiloparsecs, as indicated by luminous matter. There is evidence, however, for nonluminous matter—the so-called dark matter—extending out nearly twice this distance. The entire galactic system is rotating so that, at the position of the Sun, the orbital speed is about 220 kilometres per second, and a complete circuit will take roughly 240,000,000 years. Application of Kepler's third law leads to an estimate of the galactic mass of about 100,000,000,000 solar masses. The rotational velocity can be measured from the Doppler shifts observed in the 21-centimetre emission line of neutral hydrogen and the lines of millimetre wavelengths from various molecules, especially carbon monoxide. At great distances from the galactic centre, the rotational velocity does not drop off as expected but rather increases slightly. This condition appears to require a much larger galactic mass than can be accounted for by the known (luminous) matter. Additional evidence for the presence of dark matter comes from a variety of other observations. The nature and extent of the dark matter (or missing mass) constitutes one of today's major astronomical puzzles.

The problem of dark matter

There are about 100,000,000,000 stars in the Milky Way Galaxy. Star concentrations within the Galaxy fall into three types: open clusters, globular clusters, and associations. Open clusters lie primarily in the disk of the Galaxy and contain from 50 to 1,000 stars within a region of no more than 10 parsecs. Associations tend to have somewhat fewer stars; moreover, the constituent stars are not as closely grouped as those in the clusters and are for the most part hotter. Globular clusters may extend up to about 100 parsecs in size and may have as many as 100,000 stars. The importance of the globular clusters lies in their use as indicators of the age of the Galaxy. Because massive stars evolve more rapidly than do smaller stars, the age of a cluster can be estimated from its H–R diagram. In a young cluster the main sequence will be well populated, but in an old cluster the heavier stars will have evolved away from the main sequence. The extent of the depopulation of the main sequence provides an index of age. In this way, the oldest globular clusters have been found to be about 15,000,000,000 years old, which must therefore be the minimum age of the Galaxy.

Determining the age of the Milky Way with the aid of globular clusters

Investigations of interstellar matter. The interstellar medium (ISM) occupies the regions that lie between the stars. On average, it contains less than one atom per cubic centimetre, with about 1 percent as much mass in the form of minute dust grains. The gas, mostly hydrogen, has been mapped through its 21-centimetre emission line. The gas also contains numerous molecules. Some of these have been detected by the visible-wavelength absorption lines they impose on the spectra of more distant stars, while others have been identified by their own emission lines at millimetre wavelengths. Many of the interstellar molecules occur in giant molecular clouds, wherein complex organic molecules have been discovered.

In the vicinity of a very hot O- or B-type star, the intensity of ultraviolet radiation is sufficiently high to ionize the surrounding hydrogen out to a distance as large as 100 parsecs to produce an HII region known as a Strömgren sphere. Such regions are strong and characteristic emitters of radiation at radio wavelengths, and their dimensions are well calibrated in terms of the luminosity of the central star. Using radio interferometers, astronomers are able to measure the angular diameters of HII regions even in some external galaxies and can thereby deduce the great distances to those remote systems. This method can be used for distances up to about 30 megaparsecs.

Measuring the angular diameters of HII regions

Interstellar dust grains scatter and absorb starlight, with the effect being roughly inversely proportional to wavelength from the infrared to the near ultraviolet. As a result, stellar spectra tend to be reddened. Absorption amounts typically to about one magnitude per kiloparsec but varies considerably in different directions. Some dusty regions contain silicate materials, identified by a broad absorption feature around 10-micrometre wavelengths. Another prominent feature around 2200 angstroms has sometimes but not conclusively been attributed to graphite grains.

Starlight often shows a small degree of polarization (a few percent), with the effect increasing with stellar distance. This is attributed to the scattering of the starlight from dust grains that have been partially aligned in a weak interstellar magnetic field. The strength of this field is estimated to be a few microgauss, very close to the strength inferred from observations of nonthermal cosmic radio noise. This radio background has been identified as synchrotron radiation, emitted by highly relativistic cosmic-ray electrons (*i.e.,* those traveling at nearly the speed of light) that pass through the interstellar magnetic field. The spectrum of the cosmic radio noise is close to what is calculated on the basis of measurements of the cosmic rays near the Earth.

Observations of nonthermal cosmic radio noise

The cosmic rays constitute another component of the interstellar medium. Cosmic rays that are detected in the vicinity of the Earth consist of high-speed nuclei and electrons. Individual particle energies, expressed in electron volts (eV; 1 eV = 1.6×10^{-12} erg), range with decreasing numbers from about 10^6 to more than 10^{20} eV. Among the nuclei, hydrogen (protons) are the most plentiful at 89 percent, helium next with 9 percent, and all other nuclei at about 1 percent. Electrons are roughly 2 percent as abundant as the nuclear component.

A minority of the cosmic rays are produced in the Sun, especially at times of increased solar activity (as indicated by sunspots and flares). The origin of the galactic cosmic rays has not yet been conclusively identified, but they are thought to be produced in stellar processes such as supernova explosions, perhaps with additional acceleration occurring in the interstellar regions.

Observations of the galactic centre. The central region of the Milky Way Galaxy is so heavily obscured by large quantities of dust that direct observation has become possible only with the development of astronomy for wavelengths that penetrate to or relatively near the Earth's surface—namely, radio, infrared, and more recently X rays and gamma rays. Together, these observations have revealed a nuclear region of intense activity, with a large number of separate sources and a great deal of dust. Detection of gamma-ray emission at a line energy of 511,000 eV, which corresponds to the annihilation of electrons and positrons, along with radio mapping of a region no more than 20 a.u. across, points to a very compact and energetic source at the centre of the Galaxy. Whether this is powered by a black hole or some very close and hot stars remains to be determined. (For more specific information on this matter, see COSMOS: *Galaxies.*)

Study of other galaxies and related phenomena. Galaxies are generally classified into three principal types according to their appearance: spiral, elliptical, and irregular. Galactic dimensions are typically in the tens of kiloparsecs and the distances between galaxies typically in megaparsecs.

Classification of galaxies

Spiral galaxies (of which the Milky Way system is a characteristic example) tend to be flattened, roughly circular systems, with their constituent stars strongly concentrated along spiral arms. These arms are thought to be produced by traveling density waves, which compress and expand the galactic material. (For an explanation of the density-wave theory, see COSMOS: *Dynamics of ellipticals and spirals.*) Between the spiral arms exists a diffuse interstellar medium of gas and dust, mostly at very low temperatures (below 100 K). Spiral galaxies are typically a few kiloparsecs in thickness; they have a central bulge and taper gradually toward the outer edges.

Ellipticals show none of the spiral features. They are more densely packed stellar systems, ranging from the nearly spherical (type E0) to the very flattened (E7), and they contain little interstellar matter. Irregular galaxies account for only a few percent of all stellar systems and exhibit none of the regular features associated with either spirals or ellipticals.

Properties vary considerably between the different types of galaxies. Spirals generally have masses in the range 10^9–10^{12} solar masses, with ellipticals having values from 10 times smaller to 10 times larger and the irregulars, generally 10 times larger. The visual galactic luminosities show similar spreads, but the irregulars tend to be less luminous. In contrast, at radio wavelengths the maximum luminosity of spirals is usually 100,000 times less than for ellipticals or irregulars.

Quasars are objects whose spectra display very large redshifts, implying that they lie at the greatest distances (see above). They were discovered in 1963, but they remained enigmatic for many years. Quasars are now considered to be the exceedingly luminous cores of distant galaxies.

Identification of quasars

The Milky Way Galaxy is a member of the Local Group of galaxies, which contains about two dozen members and extends out to a distance of about one megaparsec. Two of the closest members are the Large and Small Magellanic Clouds, irregular galaxies about 50 kiloparsecs away. At 680 kiloparsecs, the Andromeda Galaxy is one of the most distant in the Local Group. Some members of the Local Group are moving toward the Milky Way system, while others are traveling away from it. At greater distances, all galaxies are moving away from the Milky Way Galaxy. Their speeds (as determined from the redshifted wavelengths in their spectra) are proportional to their distances. As explained earlier, the Hubble law relates these two quantities. In the absence of any other method, the Hubble law continues to be used for distance determinations to the farthest objects—that is, galaxies and quasars for which redshifts can be measured. (For a detailed treatment of the properties, structures, and distribution of galaxies, see GALAXIES.)

Cosmology. The currently accepted cosmological model is the big bang. In this picture, the expansion of the universe began with an intense explosion roughly 15,000,000,000 years ago. In this primordial fireball, the temperature exceeded 10^{12} K and most of the energy was in the form of radiation. As the expansion proceeded (accompanied by cooling), the role of the radiation became diminished and other physical processes dominated in turn. Thus, after about one minute, the temperature had dropped to the 1,000,000,000 K range, making it possible for nuclear reactions to take place and produce deuterons and helium nuclei. (At the higher temperatures that prevailed earlier, these nuclei would have been promptly disrupted by high-energy photons.) With further expansion, the time between nuclear collisions would have increased and the proportion of deuterons and helium nuclei would have stabilized. After a few hundred thousand years, the temperature must have dropped sufficiently for electrons to remain attached to nuclei to constitute atoms. Galaxies are thought to have begun forming after a few million years, but this stage is still very poorly understood. Star formation probably started much later, after some billions of years, and the process continues today.

The big-bang model and observational support

The observational support for this general model comes from several independent directions. The expansion has been documented by the galactic redshifted spectra. The radiation left over from the original fireball would have cooled with the expansion. One of the most striking discoveries of the 20th century came in 1965 with the observation of a widespread cosmic radiation corresponding to a temperature of close to 3 K. The spectrum of this cosmic radio noise peaks at approximately one-millimetre wavelength in the far infrared, a difficult region to observe; however, the spectrum has been well mapped at many wavelengths from that point through the radio region. Additional support for the big-bang theory comes from the observed cosmic abundances of deuterium and helium. Normal stellar nucleosynthesis cannot produce the measured quantities, which fit well with calculations of production during the early stages of the big bang.

Surveys of the cosmic 3 K radiation have indicated that it is extremely uniform in all directions. Calculations have shown that it is difficult to achieve this degree of isotropy

unless there was a very early and rapid inflationary period before the expansion settled into its present mode. However, the isotropy poses problems for models of galaxy formation. It has been conjectured that galaxies originate from turbulent conditions that produce local fluctuations of density toward which matter would then be gravitationally attracted. Such density variations are difficult to reconcile with the isotropy required by observations of the 3 K background radiation.

The very earliest stages of the big bang are less well understood. The conditions of temperature and density that prevailed prior to the first microsecond require the introduction of theoretical ideas of subatomic particles. Such particles are usually studied in laboratories with giant accelerators, but the region of particle energies of potential significance to the question at hand lies beyond the range of accelerators currently available. Fortunately, some important conclusions can be drawn from the observed cosmic helium abundance, which is dependent on conditions during the early big bang. This observed helium abundance serves to set a limit to the number of families of certain types of subatomic particles that can exist.

Finally, there is the question of the future behaviour of the universe: Is it open? That is to say, will the expansion continue indefinitely? Or is it closed, so that the expansion will slow down and eventually reverse, resulting in contraction? The outcome hinges on the total mass of the universe. For this reason, the extent of the dark matter (the missing mass) may be decisive. The amount of luminous mass currently known to exist is insufficient to close the universe, but if dark matter is universally present in the quantities suggested by galactic rotation or galactic-cluster dynamics, then the universe will be closed. An additional factor might be the mass of neutrinos. Theoretically postulated to have zero mass, neutrinos are now experimentally known to have a mass less than $^1/_{10,000}$ that of the electron. Because there are so many neutrinos in the universe, neutrino masses even as small as this might render the universe closed. The issue remains unresolved, although the detection of neutrinos (albeit only 19 of them) associated with a supernova observed in 1987 suggests that neutrino mass is probably less than the critical value postulated. (For a full discussion of the big-bang theory, including conceptions about the ultimate fate of the universe, see COSMOS: *Cosmological models.*)

THE METHODOLOGY OF ASTRONOMY

Astronomical observations involve a sequence of stages, each of which may impose constraints on the type of information attainable. Radiant energy is collected with telescopes and brought to a focus on a detector, which is calibrated so that its sensitivity and spectral response are known. Accurate pointing is required to permit the correlation of observations made with different systems in different wavelength intervals. The radiation must be spectrally analyzed so that the emission processes can be identified.

Telescopic observations. Before Galileo's use of telescopes for astronomical investigations in 1609, all observations were made by naked eye, with corresponding limits on the faintness and degree of detail that could be seen. Since that time, telescopes have become central to astronomy. With apertures much larger than the pupil of the human eye, telescopes permit the study of faint and distant objects. In addition, sufficient radiant energy can be collected in short time intervals to permit rapid fluctuations in intensity to be detected.

Optical telescopes

Optical telescopes are either refractors or reflectors that use lenses or mirrors, respectively, for their main light-collecting elements (objectives). Refractors are effectively limited to apertures with a diameter of 102 centimetres (40 inches) or less because of problems inherent in the use of large lenses. These distort under their own weight and can be supported only around the perimeter; an appreciable amount of light is lost due to absorption in the glass. Large-aperture refractors are extremely long and require large and expensive domes. The largest modern telescopes are all reflectors. They are not subject to the chromatic problems of refractors, can be better supported mechanically, and can be housed in smaller domes because of their more compact dimensions.

The angular resolving power of a single telescope is limited by the wave nature of light. For a telescope having an objective with diameter D and operating at a wavelength λ, the angular resolution (in radians) can be approximately described by the ratio (λ/D). Optical telescopes can have very high intrinsic resolving powers; in practice, however, these are not attained. Atmospheric effects limit the resolution to about one arc second.

The atmosphere does not transmit radiation of all wavelengths equally well. This restricts ground-based astronomy to the visible and radio regions of the spectrum, with some relatively narrow "windows" in the nearer infrared. Longer infrared wavelengths are heavily absorbed by atmospheric water vapour and carbon dioxide. Atmospheric effects can be reduced by careful site selection and by carrying out observations at high altitudes. Most major optical observatories are now located on high mountains, well away from cities and their reflected lights. Infrared telescopes have been constructed atop Mauna Kea in Hawaii and in the Canary Islands where atmospheric humidity is very low. Airborne telescopes designed mainly for infrared observations, such as the one aboard the Kuiper Airborne Observatory, operate at an altitude of about 12,000 metres (40,000 feet), with flight durations limited to a few hours. Telescopes for infrared-, X-, and gamma-ray observations have been carried to altitudes of more than 30,000 metres by balloons. Higher altitudes can be attained during short-duration rocket flights for ultraviolet observations. Telescopes for all wavelengths from infrared to gamma rays have been carried by remote-controlled Earth-orbiting satellites as well as by some manned space vehicles.

Observations from aircraft, balloons, rockets, and Earth-orbiting spacecraft

Angular resolution better than one milliarcsecond has been achieved at radio wavelengths by using several telescopes in an array. In such an arrangement, the effective aperture (D) becomes the greatest distance between component telescopes. For example, in the Very Large Array (VLA), operated near Socorro, N.M., by the National Radio Astronomy Observatory, 27 movable radio dishes are set out along tracks that extend for nearly 21 kilometres. In a technique called very long baseline interferometry (VLBI), simultaneous observations are made with radio telescopes thousands of kilometres apart.

Very long baseline interferometry

The Earth is a moving platform for astronomical observations. It is important that the specification of precise celestial coordinates be made in ways that correct for telescope location, the position of the Earth in its orbit around the Sun, and the epoch of observation, since the Earth's axis of rotation moves slowly over the years. Time measurements are now based on atomic clocks rather than on the Earth's rotation, and telescopes can be driven continuously to compensate for the planet's rotation so as to permit tracking of a given astronomical object.

Use of radiation detectors. While the human eye remains an important astronomical tool, detectors capable of greater sensitivity and more rapid response are needed at visible wavelengths and especially to extend observations beyond that region of the electromagnetic spectrum. Photography has been used for more than 100 years and continues to be an essential tool. Long time exposures may be needed to reveal faint objects. This integrative property of photography smooths out rapid variations in intensity, for which electronic methods must otherwise be used. Furthermore, photography provides an archival record. A photograph of a particular celestial object may include the images of many other objects that were not of interest when the picture was taken but that become the focus of study years later. When quasars were discovered in 1963, for example, photographic plates dating to before 1900, held by the Harvard College Observatory, were examined to trace possible changes in the position or intensity of quasar 3C273. Also, major photographic surveys, such as those of the National Geographic Society and Palomar Observatory, can provide a historic base for long-term studies.

Photography in astronomical research

Photographic film converts only a few percent of the incident photons into its images, whereas efficiencies of better than 50 percent can be achieved by several electronic meth-

ods of detection. The greater sensitivity and intrinsically rapid response of such methods are exploited for tracking exceedingly rapid intensity variations. For example, pulsars that emit their radiation at millisecond intervals can be followed and their pulse shapes monitored. The arrival of individual photons can be recorded with phototubes or more recent devices, such as the charge-coupled device (CCD). Special photographic materials can be employed for the shortest infrared wavelengths, but, for wavelengths longer than a few micrometres, semiconductor detectors that operate at temperatures of liquid nitrogen (77 K) or liquid helium (below 4 K) have to be used. In detectors of this kind, photons are absorbed in small semiconducting crystals to produce a minute temperature increase or a change in electrical resistance. Individual photons are not recorded. Reception of radio waves is based on the production of a small voltage in an antenna rather than on photon counting. Individual X-ray and gamma-ray photons possess sufficient energy to be detected through the ionization they produce.

Spectral analysis

Spectral analysis involves measuring intensity as a function of wavelength or frequency. In some detectors, such as those for X and gamma rays, the energy of each photon can be measured directly. Photographic film is sensitive to photons over a wide range of wavelengths. For low-resolution spectroscopy, broad-band filters suffice. Greater resolution can be obtained with prisms, gratings, and interferometers.

As a departure from the traditional astronomical approach of remote observing, certain more recent lines of research involve the analysis of actual samples under laboratory conditions. These include studies of meteorites, rock samples returned from the Moon, and interplanetary dust particles collected by aircraft in the stratosphere or by Earth-orbiting spacecraft. In all such cases, a wide range of highly sensitive laboratory techniques can be adapted for the often microscopic samples. It is possible to supplement chemical analysis with mass spectroscopy, so that the isotopic composition can be determined. Radioactivity and collisions between cosmic-ray particles can produce minute quantities of gas that remain trapped in crystals. Heating under laboratory conditions releases this gas, which then is analyzed in a mass spectrometer. X-ray spectrometers, electron microscopes, and microprobes are employed to determine crystal structure and composition, from which temperature and pressure conditions at the time of formation can be inferred.

Theoretical approaches. Theory is just as important as observation in astronomy. It is required for the interpretation of observational data, for the construction of models of celestial objects and their properties, and for guiding further observations. Theoretical astrophysics is based on the laws of physics that have been validated with great precision through controlled experiments. Application of these laws to specific astrophysical problems, however, may yield equations too complex for direct solution. Two general approaches are then available. In the traditional method, an idealized description of the problem is formulated, incorporating only the major physical components, to provide equations that can be either solved directly or evaluated numerically. Alternatively, a computer program that will explore the problem numerically can be devised. Computational science is taking its place as a major division alongside theory and experiment.

Theoretical astrophysics

IMPACT OF ASTRONOMY

No area of science is totally self-contained. Discoveries in one field find applications in others, often unpredictably. There are various notable examples of this involving astronomical studies. Newton's laws of motion and gravity emerged from the analysis of planetary and lunar orbits. Observations during the 1919 solar eclipse provided dramatic confirmation of Einstein's general theory of relativity, which has gained further support with the recent discovery and tracking of the binary pulsar designated PSR 1913 + 16. The behaviour of nuclear matter and of some elementary particles is now better understood as a result of measurements of neutron stars and the cosmological helium abundance, respectively. Study of the theory of synchrotron radiation was greatly stimulated by its detection in radiation emitted by the Crab Nebula, and machines are now being used to produce synchrotron radiation to probe the structure of materials.

Astronomical knowledge also has had impact outside of science. The earliest calendars were based on astronomical observations of the cycles of repeated solar and lunar positions. Also, for centuries, familiarity with the positions and apparent motions of the stars enabled sea voyagers to navigate with reasonable accuracy. Perhaps the single greatest effect that astronomical studies have had on modern society has been in molding the attitude of its members. The development of what is known as the scientific method was strongly influenced by astronomical observations. The power of science to provide the basis for accurate predictions of such phenomena as eclipses and the positions of the planets and later, so dramatically, of comets has generated an attitude toward science that remains an important social force today. (M.W.Fr.)

BIBLIOGRAPHY

History of physical sciences: A. RUPERT HALL and MARIE BOAS HALL, *A Brief History of Science* (1964, reprinted 1988), provides a good introduction to the subject; A.E.E. MCKENZIE, *The Major Achievements of Science,* 2 vol. (1960, reprinted 1988), concentrates on developments from the 16th century, with brief extracts from original sources; and CECIL J. SCHNEER, *The Search for Order: The Development of the Major Ideas in the Physical Sciences from the Earliest Times to the Present* (1960, reissued 1984 as *The Evolution of Physical Science*), accounts for the developments from the 17th through the 19th centuries. Comprehensive surveys include STEPHEN F. MASON, *A History of the Sciences,* new rev. ed. (1962); and STEPHEN TOULMIN and JUNE GOODFIELD, *The Architecture of Matter* (1962, reissued 1982). THOMAS S. KUHN, *The Structure of Scientific Revolutions,* 2nd enl. ed. (1970), presents the "paradigm" theory of science, based on historical examples; GERALD HOLTON, *Thematic Origins of Scientific Thought: Kepler to Einstein,* rev. ed. (1988), offers a new interpretation of the history of science, with case studies of the work of Einstein and others; and I. BERNARD COHEN, *Revolution in Science* (1985), is a comparative study. See also STEPHEN G. BRUSH, *The History of Modern Science: A Guide to the Second Scientific Revolution, 1800–1950* (1988). CHARLES COULSTON GILLISPIE (ed.), *Dictionary of Scientific Biography,* 16 vol. (1970–80), is an excellent source for authoritative biographical data. For references to the scholarly literature in the history of science, consult the "Critical Bibliography of the History of Science and Its Cultural Influences," an annual feature in *Isis,* an international review of the history of science. Cumulations of this bibliography appeared as MAGDA WHITROW (ed.), *Isis Cumulative Bibliography: 1913–65,* 6 vol. (1971–84); and JOHN NEU (ed.), *Isis Cumulative Bibliography 1966–1975,* 2 vol. (1980–85).

Physics: I. BERNARD COHEN, *The Birth of a New Physics,* rev. ed. (1985), is an account of the work of Galileo, Newton, and other 17th-century scientists. See also EMILIO SEGRÈ, *From Falling Bodies to Radio Waves: Classical Physicists and Their Discoveries* (1984), and *From X-Rays to Quarks: Modern Physicists and Their Discoveries* (1980; originally published in Italian, 1976). HENRY A. BOORSE and LLOYD MOTZ (eds.), *The World of the Atom,* 2 vol. (1966), is a comprehensive anthology of historical sources on 19th- and 20th-century developments in atomic physics. For recent history, see STEPHEN G. BRUSH, "Resource Letter HP-1: History of Physics," *American Journal of Physics,* 55:683-691 (August 1987). An interesting collection of writings by physicists is presented in JEFFERSON HANE WEAVER (ed.), *The World of Physics: A Small Library of the Literature of Physics from Antiquity to the Present,* 3 vol. (1987).

Reference works surveying the scope and methodology of physics include ROBERT M. BESANÇON (ed.), *The Encyclopedia of Physics,* 3rd ed. (1985); and CESARE EMILIANI, *Dictionary of the Physical Sciences: Terms, Formulas, Data* (1987). Other works include DAVID HALLIDAY and ROBERT RESNICK, *Fundamentals of Physics,* 3rd extended ed., 2 vol. (1988), a good standard text; GERALD HOLTON, *Introduction to Concepts and Theories in Physical Science,* 2nd ed., rev. by STEPHEN G. BRUSH (1973, reprinted 1985), analyzing the physical theories from a historical standpoint; RICHARD P. FEYNMAN, ROBERT B. LEIGHTON, and MATTHEW SANDS, *The Feynman Lectures on Physics,* 3 vol. (1963–65), and RICHARD P. FEYNMAN, *QED: The Strange Theory of Light and Matter* (1985), works by a modern master; FRANK CLOSE, MICHAEL MARTEN, and CHRISTINE SUTTON, *The Particle Explosion* (1987), a discussion of the latest developments in fundamental physics, written for the general reader; STEVEN WEINBERG, *The Discovery of Subatomic Particles* (1983), and *The First Three Minutes: A Modern View of the Origin*

of the Universe, updated ed. (1988); P.C.W. DAVIES, *Space and Time in the Modern Universe* (1977); and PETER G. BERGMANN, *The Riddle of Gravitation,* rev. ed. (1987). An unusual social history of the U.S. scientific community is presented in DANIEL J. KEVLES, *The Physicists* (1978, reprinted 1987).

Chemistry: Historical developments in chemistry through the 17th century are explored in ROBERT P. MULTHAUF, *The Origins of Chemistry* (1966). CECIL J. SCHNEER, *Mind and Matter: Man's Changing Concepts of the Material World* (1970, reprinted 1988), gives an interesting account of the history of chemistry in relation to the structure of matter; AARON J. IHDE, *The Development of Modern Chemistry* (1964, reprinted 1984), is a comprehensive history covering the 18th to the middle of the 20th century; and FREDERIC LAWRENCE HOLMES, *Lavoisier and the Chemistry of Life: An Exploration of Scientific Creativity* (1985), is a study of chemical experimentation.

Concise explanations of chemical terms can be found in DOUGLAS M. CONSIDINE and GLENN D. CONSIDINE (eds.), *Van Nostrand Reinhold Encyclopedia of Chemistry,* 4th ed. (1984), a specialized reference work. Comprehensive treatment of chemical theories and reactivity is presented in DONALD A. MCQUARRIE and PETER A. ROCK, *General Chemistry,* 2nd ed. (1987); and in JOHN C. KOTZ and KEITH F. PURCELL, *Chemistry & Chemical Reactivity* (1987). Studies of common applications of chemistry, intended for the general reader, include WILLIAM R. STINE *et al., Applied Chemistry,* 2nd ed. (1981); and JOHN W. HILL, *Chemistry for Changing Times,* 5th ed. (1988). An overview of modern chemistry and discussion of its prospects is found in GEORGE C. PIMENTEL and JANICE A. COONROD, *Opportunities in Chemistry: Today and Tomorrow* (1987). P.W. ATKINS, *Molecules* (1987), is a pictorial examination of chemical structure. LIONEL SALEM, *Marvels of the Molecule,* trans. from French (1987), presents the molecular orbital theory of chemical bonding in simple terms. The extent to which molecular structure has become central to biochemistry is demonstrated in the articles in *The Molecules of Life: Readings from Scientific American* (1985). The fundamental principles governing chemical change and the laws of thermodynamics are presented, with a minimum of mathematics, in P.W. ATKINS, *The Second Law* (1984); and JOHN B. FENN, *Engines, Energy, and Entropy: A Thermodynamics Primer* (1982). Social aspects of developments in chemistry, especially the environmental costs of use of chemical products, are studied in LUCIANO CAGLIOTI, *The Two Faces of Chemistry,* trans. from Italian (1983).

Astronomy: ANTONIE PANNEKOEK, *A History of Astronomy* (1961, reissued 1969; originally published in Dutch, 1951); and STEPHEN TOULMIN and JUNE GOODFIELD, *The Fabric of the Heavens* (1961), are comprehensive surveys. THOMAS S. KUHN, *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought* (1957), discusses the ancient geocentric theory of the universe and the heliocentric system that replaced it. CHARLES A. WHITNEY, *The Discovery of Our Galaxy* (1971, reprinted 1988), explains theories and observations in stellar astronomy from the 18th through the 20th centuries. OTTO STRUVE and VELTA ZEBERGS, *Astronomy of the 20th Century* (1962), studies the developments and achievements of the first 50 years of the century. The same period is covered in a later study from the ongoing series "The General History of Astronomy": OWEN GINGERICH (ed.), *Astrophysics and Twentieth-Century Astronomy to 1950* (1984).

Literature on current knowledge in astronomy is voluminous. C.W. ALLEN, *Astrophysical Quantities,* 3rd ed. (1973, reprinted 1983), provides a compact compilation of data; and KENNETH R. LANG, *Astrophysical Formulae: A Compendium for the Physicist and Astrophysicist,* 2nd rev. ed. (1980), is a comprehensive reference source, with extensive formulas and background data. ROBERT BURNHAM, JR., *Burnham's Celestial Handbook: An Observer's Guide to the Universe Beyond the Solar System,* rev. ed., 3 vol. (1978), is a guide to stars, nebulas, and galaxies. ARTHUR P. NORTON, *Norton's Star Atlas and Reference Handbook (Epoch 1950.0),* 17th ed., edited by GILBERT E. SATTERTHWAITE (1978, reprinted 1986), is a popular atlas. Introductory texts include GEORGE O. ABELL, DAVID MORRISON, and SIDNEY C. WOLFF, *Exploration of the Universe,* 5th ed. (1987), a comprehensive classic; and A.E. ROY and D. CLARKE, *Astronomy: Principles and Practice,* 3rd ed. (1988), a survey of experimental basics.

The scope and methodology of astronomy are explored in specialized works: J. KELLY BEATTY, BRIAN O'LEARY, and ANDREW CHAIKIN, *The New Solar System,* 2nd ed. (1982); PETER H. CADOGAN, *The Moon: Our Sister Planet* (1981); JOHN C. BRANDT and ROBERT D. CHAPMAN, *Introduction to Comets* (1981); GORDON WALKER, *Astronomical Observations: An Optical Perspective* (1987); R.J. TAYLER, *The Stars: Their Structure and Evolution* (1970, reissued 1981); CARL E. FICHTEL and JACOB I. TROMBKA, *Gamma Ray Astrophysics: New Insights into the Universe* (1981); CLAUS E. ROLFS and WILLIAM S. RODNEY, *Cauldrons in the Cosmos: Nuclear Astrophysics* (1988); DIMITRI MIHALAS and JAMES BINNEY, *Galactic Astronomy: Structure and Kinematics,* 2nd ed. (1981); JOHN D. KRAUS *et al., Radio Astronomy,* 2nd ed. (1986); M.S. LONGAIR, *High Energy Astrophysics: An Informal Introduction for Students of Physics and Astronomy* (1981); MICHAEL ROWAN-ROBINSON, *The Cosmological Distance Ladder: Distance and Time in the Universe* (1985); and EDWARD R. HARRISON, *Cosmology, the Science of the Universe* (1981). For up-to-date reviews of specialized subjects and for current listings of observational information, consult the *Annual Review of Astronomy and Astrophysics, Annual Review of Earth and Planetary Sciences, The Astronomical Almanac* (annual), and *The Observer's Handbook* (annual).

(S.G.B./L.M.B./M.C.U./M.W.Fr.)

Picasso

The enormous body of Picasso's work, including painting, sculpture, works on paper, ceramics, and poetry, remains; and the legend lives on—a tribute to the vitality of the "disquieting" Spaniard with the "sombre . . . piercing" eyes who superstitiously believed that work would keep him alive. For nearly 80 of his 91 years Picasso devoted himself to an artistic production that contributed significantly to and paralleled the whole development of modern art in the 20th century.

LIFE AND CAREER

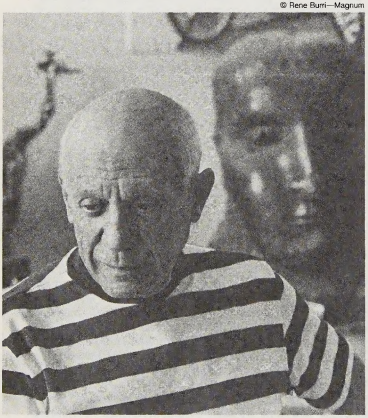
Early years. Pablo Picasso, son of José Ruiz Blasco, a professor of drawing, and Maria Picasso López, was born in Málaga, Spain, on Oct. 25, 1881. His unusual adeptness for drawing began to manifest itself early, around the age of 10, when he became his father's pupil in La Coruña, where the family moved in 1891. From that point his ability to experiment with what he learned and to develop new expressive means quickly allowed him to surpass his father's abilities. In La Coruña his father shifted his own ambitions to those of his son, providing him with models and support for his first exhibition there at the age of 13.

La Llotja

The family moved to Barcelona in the autumn of 1895, and Pablo entered the local art academy (La Llotja), where his father had assumed his last post as professor of drawing. The family hoped that their son would achieve success as an academic painter, and in 1897 his eventual fame in Spain seemed assured; in that year his painting "Science and Charity," for which his father modeled for the doctor, was awarded an honorable mention in Madrid at the Fine Arts Exhibition.

The Spanish capital was the obvious next stop for the young artist intent on gaining recognition and fulfilling family expectations. Pablo Ruiz duly set off for Madrid in the autumn of 1897 and entered the Royal Academy of San Fernando. But finding the teaching there stupid, he increasingly spent his time recording life around him, in the cafés, on the streets, in the brothels, and in the Prado, where he discovered Spanish painting. He wrote: "The Museum of paintings is beautiful. Velázquez first class; from El Greco some magnificent heads, Murillo does not convince me in every one of his pictures." Works by these and other artists would capture Picasso's imagination at different times during his long career. Goya, for instance, was an artist whose works Picasso copied in the Prado in 1898 (a portrait of the bullfighter Pepe Illo and the drawing for one of the Caprichos, "Bien tirada está," which shows a Celestina [procuress] checking a young maja's stockings). These same characters reappear in his late work—Pepe Illo in a series of engravings (1957) and Celestina as a kind of voyeuristic self-portrait, especially in the series of etchings and engravings known as "Suite 347" (1968).

© Rene Burri—Magnum



Picasso.

Picasso fell ill in the spring of 1898 and spent most of the remaining year convalescing in the Catalan village of Horta de Ebro in the company of his Barcelona friend Manuel Pallarès. When Picasso returned to Barcelona in early 1899, he was a changed man: he had put on weight, he had learned to live on his own in the open countryside, he spoke Catalan, and most importantly he had made the decision to break with his art school training and to reject his family's plans for his future. He even began to show a decided preference for his mother's surname, and more often than not he signed his works P.R. Picasso (by late 1901 he had dropped the Ruiz altogether).

Break with art school training

In Barcelona Picasso moved among a circle of Catalan artists and writers whose eyes were turned toward Paris. These were his friends at the café Els Quatre Gats ("The Four Cats," styled after the Chat Noir ["Black Cat"] in Paris), where Picasso had his first Barcelona exhibition in February 1900, and they were the subjects of more than 50 portraits (in mixed media) in the show. In addition, there was a dark, moody "modernista" painting, "Last Moments" (later painted over), showing the visit of a priest to the bedside of a dying woman, a work that was accepted for the Spanish section of the Exposition Universelle in Paris in that year. Eager to see his own work in place and to experience Paris firsthand, Picasso set off in the company of his studio-mate Carles Casagemas ("Portrait of Carles Casagemas," 1899) to conquer, if not Paris, at least a corner of Montmartre.

Discovery of Paris. One of Picasso's principal artistic discoveries on that trip (October–December) was colour—not the drab colours of the Spanish palette, the black of the shawls of Spanish women, or the ochres and browns of the Spanish landscape, but brilliant colour—the colour of Van Gogh, of new fashion, of a city celebrating a world's fair. Using charcoal, pastels, watercolours, and oils, Picasso recorded life in the French capital ("Lovers in the Street," 1900). In "Moulin de la Galette" (1900) he paid tribute to French artists such as Toulouse-Lautrec and Steinlen as well as his Catalan compatriot Ramon Casas.

Casagemas

After just two months Picasso returned to Spain with Casagemas, who had become despondent about a failed love affair. Having tried unsuccessfully to amuse his friend in Málaga, Picasso took off for Madrid, where he worked as an art editor for a new journal, *Arte Joven.* Casagemas returned to Paris and attempted to shoot the woman he loved, then turned the gun on himself and died. The impact on Picasso was deep: it was not just that he had lost his loyal friend and perhaps felt a sense of guilt for having abandoned him; more importantly, he had gained the emotional experience and the material that would stimulate the powerful expressiveness of the works of the so-called Blue Period. Picasso made two death portraits of Casagemas several months later in 1901 as well as two funeral scenes ("Mourners" and "Evocation"), and in 1903 Casagemas appeared as the artist in the enigmatic painting "La Vie."

Blue Period. Between 1901 and mid-1904, when blue was the predominant colour in his paintings, Picasso

moved back and forth between Barcelona and Paris, taking material for his work from one place to the other. For example, his visits to the Women's Prison of Saint-Lazare in Paris in 1901–02, which provided him with free models and compelling subject matter ("The Soup," 1902), were reflected in his depictions of Barcelona street people—blind or lonely beggars and castaways in 1902–03 ("Crouching Woman," 1902; "Blind Man's Meal," 1903; "Old Jew and a Boy," 1903). The subject of maternity (women were allowed to keep nursing children with them at the prison) also preoccupied Picasso at a time when he was searching for material that would best express traditional art-historical subjects in 20th-century terms.

The move to Paris. Picasso finally made the decision to move permanently to Paris in the spring of 1904, and his work reflects a change of spirit and especially a change of intellectual and artistic currents. The traveling circus and saltimbanques became a subject he shared with a new and important friend, Guillaume Apollinaire. To both the poet and the painter these rootless wandering performers ("Girl Balancing on a Ball," 1905; "The Actor," 1905) became a kind of evocation of the artist's position in modern society. Picasso specifically made this identification in "Family of Saltimbanques" (1905), where he assumes the role of Harlequin and Apollinaire is the strongman (according to their mutual friend, the writer André Salmon).

Friendship with Apollinaire

Picasso's personal circumstances also changed at the end of 1904, when Fernande Olivier became his mistress. Her presence inspired many works during the years leading up to Cubism, especially on their trip to Gosol in 1906 ("Woman with Loaves"), including the sculpture "Head of a Woman" (1909) and several paintings related to it ("Woman with Pears," 1909).

Colour never came easily to Picasso, and he reverted to a generally more Spanish (*i.e.*, monochromatic) palette. The tones of the Blue Period were replaced from late 1904 to 1906 in the so-called Rose Period by those of pottery, of flesh, and of the earth itself ("The Harem," 1906). Picasso seems to have been working with colour in an attempt to come closer to sculptural form, especially in 1906 ("Two Nudes"; "La Toilette"). His "Portrait of Gertrude Stein" (1906) and a "Self-Portrait with Palette" (1906) show this development as well as the influence of his discovery of primitive Iberian sculpture.

Les Demoiselles d'Avignon. Toward the end of 1906 Picasso began work on a large composition that came to be called "Les Demoiselles d'Avignon" (1907). His violent treatment of the female body and masklike painting of the faces (influenced by a study of African art) have made this work controversial. Yet the work was firmly based upon art-historical tradition: a renewed interest in El Greco contributed to the fracturing of the space and the gestures of the figures, while the overall composition owed much to Cézanne's "Bathers" as well as to Ingres's harem scenes. The "Demoiselles," however, named by Picasso's friend Max Jacob (to refer to Avignon Street in Barcelona, where sailors found popular brothels), was perceived as a shocking and direct assault: these women were not conventional images of beauty but prostitutes who challenged the very tradition from which they were born. Although he had his collectors by this date (Leo and Gertrude Stein, the Russian merchant Sergey Shchukin) and a dealer (Daniel-Henry Kahnweiler), Picasso chose to roll up the canvas of the "Demoiselles" and to keep it out of sight for several years.

In 1908 the African-influenced striations and masklike heads were superseded by a technique that incorporated elements he and his new friend Georges Braque found in the work of Cézanne, whose shallow space and characteristic planar brushwork are especially evident in Picasso's work of 1909. Still lifes, inspired by Cézanne, also became an important subject for the first time in Picasso's career.

Cubism. Picasso and Braque worked together closely during the next few years (1909–12)—the only time Picasso ever worked with another painter in this way—and they developed what came to be known as Analytical Cubism. Early Cubist paintings were often misunderstood by critics and viewers because they were thought to be merely geometric art. Yet the painters themselves believed they were presenting a new kind of reality that broke away from Renaissance tradition, especially from the use of perspective and illusion. For example, they showed multiple views of an object on the same canvas to convey more information than could be contained in a single, limited illusionistic view.

As Kahnweiler saw it, Cubism signified the opening up of closed form by the "re-presentation" of the form of objects and their position in space instead of their imitation through illusionistic means; and the analytic process of fracturing objects and space, light and shadow, and even colour was likened by Apollinaire to the way in which the surgeon dissects a cadaver. This type of analysis is characteristic of Picasso's work beginning in 1909, especially in the landscapes he made on a trip to Spain that summer ("Factory at Horta de Ebro"). These were followed in 1910 with a series of hermetic portraits ("Ambroise Vollard"; "Daniel-Henry Kahnweiler"); and in his 1911–12 paintings of seated figures, often playing musical instruments ("The Accordionist," 1911), Picasso merged figures, objects, and space on a kind of grid. The palette was once again limited to monochromatic ochres, browns, and grays.

Neither Braque nor Picasso desired to move into the realm of total abstraction in their Cubist works, although they implicitly accepted inconsistencies such as different points of view, different axes, and different light sources in the same picture. Furthermore, the inclusion of abstract and representational elements on the same picture plane led both artists to reexamine what two-dimensional elements, such as newspaper lettering, signified. A song title, "Ma Jolie," for instance, could point to events outside the painting; it could refer narratively to Picasso's new mistress, Eva (Marcelle Humbert). But it could also point to compositional elements within the painting, to the function of flat pictorial elements that play off other flat planes or curvilinear motifs. The inclusion of lettering also produced the powerful suggestion that Cubist pictures could be read coming forward from the picture plane rather than receding (in traditional perspective) into it. And the Cubists' manipulation of the picture shape—their use of the oval, for example—redefined the edge of the work in a way that underlined the fact that in a Cubist picture the canvas provides the real space.

Abstract elements in Cubism

Collage. By 1912 Picasso and Braque were gluing real paper (papier collé) and other materials (collage) onto their canvases, taking a stage further the Cubist conception of a work as a self-contained, constructed object. This Synthetic phase (1912–14) saw the reintroduction of colour, while the actual materials often had an industrial reference (*e.g.*, sand or printed wallpaper). Still lifes and, occasionally, heads were the principal subjects for both artists. And in Picasso's works the multiple references inherent in his Synthetic compositions—curves that refer to guitars and at the same time to ears, for instance—introduce an element of play that is characteristic of so much of his work ("Student with a Pipe," 1913) and lead to the suggestion that one thing becomes transformed into another. "Absinthe glass" (1914; six versions), for example, is in part sculpture (cast bronze), in part collage (a real silver sugar strainer is welded onto the top), and in part painting (Neo-Impressionist brushstrokes cover planes of white paint). But the work is neither sculpture, nor collage, nor painting; planes refer to two-dimensionality, while the object indeed possesses three dimensions. The work of art thus hovers between reality and illusion.

By 1915 Picasso's life had changed and so, in a sense, had the direction of his art. At the end of that year his beloved Eva died, and the painting he had worked on during her illness ("Harlequin," 1915; Museum of Modern Art, New York City) gives testimony to his grief—a half-Harlequin, half-Pierrot artist before an easel holds an unfinished canvas against a black background.

Parade. World War I dispersed Picasso's circle; Apollinaire, Braque, and others left for the front, while most of Picasso's Spanish compatriots returned to their neutral homeland. Picasso stayed in France, and from 1916 his friendship with the composer Erik Satie took him into a new avant-garde circle that remained active during the

war. The self-appointed leader of this nucleus of talents who frequented the Café de la Rotonde was the young poet Jean Cocteau. His idea to stage a wartime theatrical event in collaboration with Sergey Diaghilev's Ballets Russes resulted in the production of *Parade,* a work about a circus sideshow that incorporated imagery of the new century, such as skyscrapers and airplanes. Cocteau went to Satie for the music and then to Picasso for the sets and costumes. Work began in 1917, and although Picasso intensely disliked travel, he agreed to go with Cocteau to Rome where they joined Diaghilev and the choreographer of *Parade,* Léonide Massine. It was on this occasion that Picasso also met his future wife, Olga Kokhlova, among the dancers.

Jean Cocteau

Parade was first performed in May 1917 at the Théâtre du Châtelet in Paris, where it was considered no less than an attempt to undermine the solidarity of French culture. Satie seems to have been the principal target of abuse (partly because of his inclusion of airplane propellers and typewriters in the score), while Picasso disarmed the public with the contrast between his basically realistic stage curtain and the startling Synthetic Cubist constructions worn by the characters, the sideshow managers, in the ballet.

New Mediterraneanism. Picasso's paintings and drawings of the late teens often seem unexpectedly naturalistic in contrast to the Cubist works that preceded or sometimes coincided with them ("Passeig de Colom," 1917). After his travels to Italy and a return to Barcelona in 1917 (*Parade* was performed there in November), a new spirit of Mediterraneanism made itself felt in his work, especially in the use of classical forms and drawing techniques. This was reinforced by a conscious looking back to Ingres (for example, in Picasso's portrait drawings of Jacob and Vollard, 1915) and to late Renoir. Even the direction of Picasso's Cubist work was affected. By clarifying planes, forms, and colour, the artist imparted to his Cubist paintings a classical expression (Saint-Raphaël still lifes, 1919; two versions of the "Three Musicians," 1921).

Picasso's only legitimate child, Paulo, was born in 1921. As part of his new status as darling of the socialites (encouraged particularly by his wife and Jean Cocteau) Picasso continued his collaborations with the Ballets Russes and produced designs for Manuel de Falla's *Three-Cornered Hat* (1919); Igor Stravinsky's *Pulcinella* (1920); De Falla's *Cuadro Flamenco* (1921); and Satie's ballet *Mercure* (1924). André Breton called Picasso's designs for this ballet "tragic toys for adults" created in the spirit of Surrealism.

Surrealism. Although Picasso never became an official member of the group, he had intimate connections with the most important art movement between the two world wars, Surrealism. The Surrealist establishment, including its main propagandist, André Breton, claimed him as one of their own, and Picasso's art gained a new dimension from contact with his Surrealist friends, particularly the writers. Inherent in Picasso's work since the "Demoiselles" were many elements that the official circle advocated. The creation of monsters, for instance, could certainly be perceived in the disturbing juxtapositions and broken contours of the human figure in Cubist works; Breton specifically pointed to the strange "Woman in a Chemise" (1913). Moreover, the idea of reading one thing for another, an idea implicit in Synthetic Cubism, seemed to coincide with the dreamlike imagery the Surrealists championed.

What the Surrealist movement gave to Picasso were new subjects—especially erotic ones—as well as a reinforcement of disturbing elements already in his work. The many variations on the subject of bathers with their overtly sexual and contorted forms (Dinard series, 1929) show clearly the impact of Surrealism, while in other works the effect of distortion on the emotions of the spectator can also be interpreted as fulfilling one of the psychological aims of Surrealism (drawings and paintings of the "Crucifixion," 1930–35). In the 1930s Picasso, like many of the Surrealist writers, often played with the idea of metamorphosis. For example, the image of the minotaur, the monster of Greek mythology—half bull and half human—that traditionally has been seen as the embodiment of the struggle between the human and the bestial, becomes in Picasso's work not only an evocation of that idea but also a kind of self-portrait.

Finally, Picasso's own brand of Surrealism found its strongest expression in poetry. He began writing poetry in 1934, and during one year, from February 1935 to the spring of 1936, Picasso virtually gave up painting. Collections of poems were published in *Cahiers d'Art* (1935) and in *La Gaceta de Arte* (1936, Tenerife), and some years later he wrote the Surrealist play *Le Désir attrapé par la queue* (1941, *Desire Caught by the Tail*).

Sculpture. Picasso's reputation as a major 20th-century sculptor came only after his death, because he had kept much of his sculpture in his own collection. Beginning in 1928, Picasso began to work in iron and sheet metal in Julio González's studio in Paris. Then, in 1931, with his new mistress, Marie-Thérèse Walter, he left his wife and moved to a country home at Boisgeloup, where he had room for sculpture studios. There, with Marie-Thérèse as his muse, Picasso began working on large-scale plaster heads. He also began to make constructions incorporating found objects, and until the end of his life Picasso continued working in sculpture in a variety of materials.

Marie-Thérèse Walter

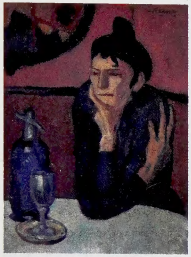
The 1930s. The privacy of his life with the undemanding Marie-Thérèse formed a contrast to the hectic pace of life kept by Olga and her bourgeois circle of society friends. Once in Boisgeloup, Picasso lived openly with Marie-Thérèse (with whom he had a child, Maya, in 1935), and she became the subject of his often lyrical, sometimes erotic paintings, in which he combined intense colour with flowing forms ("Girl Before a Mirror," 1932).

Picasso never completely dissociated himself from the women who had shared his life once a new lover occupied his attention. This is evident in his work, in which one mistress often turns into another; for instance, in a private sketchbook (number 99, 1929) Picasso's portrait drawings betray his double life, for the pictures of his then secret mistress evolve into horrific images of screaming Olgas. And in 1936, while money and a certain amount of attention were given to both Olga and Marie-Thérèse, Picasso moved back to Paris and began to live with the Yugoslav photographer Dora Maar. This change in his own life coincided with a period of personal preoccupation with the Spanish Civil War, which had begun in that year.

Although Picasso never returned to his native country after a visit in 1934, his sympathies always lay with Spain (the short-lived Republican government named him honorary director of the Prado), and in early 1937 he produced a series of etchings and aquatints ("Dream and Lie of Franco") to be sold in support of the Republican cause. His major contribution, of course, was the mural painting "Guernica" (named for the Basque town bombed in 1937 by the Fascists) commissioned by the Republican government for the Spanish pavilion at the 1937 World's Fair in Paris. As compensation Picasso was provided with a studio in Paris on rue des Grands Augustins large enough to accommodate the enormous canvas (11.5 × 25.5 feet; 3.49 × 7.77 metres). Dora Maar worked with him to complete the final work, which was realized in just over three weeks. The imagery in "Guernica"—the gored horse, the fallen soldier, and screaming mothers with dead babies (representing the bullfight, war, and female victims, respectively)—was employed to condemn the useless destruction of life, while at the same time the bull represented the hope of overcoming the unseen aggressor, Fascism.

"Guernica"

World War II and after. The expressive quality of both the forms and gestures in the basically monochromatic composition of "Guernica" found its way into Picasso's other work, especially in the intensely coloured versions of "Weeping Woman" (1937) as well as in related prints and drawings, in portraits of Dora Maar and Nusch Éluard (wife of Picasso's friend, the French poet Paul Éluard), and in still lifes ("Still Life with Red Bull's Head," 1938). These works led to the claustrophobic interiors and skull-like drawings (sketchbook number 110, 1940) of the war years, which Picasso spent in France with Dora Maar as well as with Jaime Sabartés, a friend of his student days in Barcelona. Thereafter Sabartés shared Picasso's life as secretary, biographer, and companion, and more often than not as the butt of endless jokes ("Portrait of Jaime



"The Absinthe Drinker," oil painting by Pablo Picasso, 1901. In the Hermitage Museum, St. Petersburg. 73 x 54 cm.



"Les Demoiselles d'Avignon," oil painting by Pablo Picasso, 1907. In the Museum of Modern Art, New York City. 243.9 x 233.7cm.

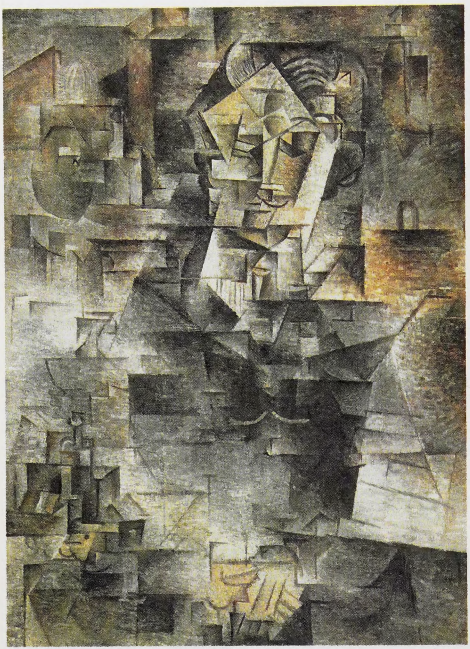


"Woman's Head," sculpture modeled in clay and reproduced by casting in bronze, c. 1909. In the Museum of Modern Art, New York City. Height 41.3 cm.



"The Old Guitarist," oil painting by Pablo Picasso, c. 1903–04. In the Art Institute of Chicago, 122.9 x 82.6 cm.

Plate 1: By courtesy of (top right) the Museum of Modern Art, New York City, acquired through the Lillie B. Bliss Bequest; (bottom right) the Art Institute of Chicago, Helen Birch Bartlett Memorial Collection; (bottom left) the Museum of Modern Art, New York City; photographs, (top left) © Alexander Burkatowski—Corbis; (top right and bottom left) © 2001 The Museum of Modern Art, New York City



"Daniel-Henry Kahnweiler," oil painting by Pablo Picasso, 1910. In the Art Institute of Chicago. 101.1 x 73.3 cm.



"Mother and Child," oil painting by Pablo Picasso, 1921. In the Art Institute of Chicago. 142.9 x 172.7 cm.

Plate 2: By courtesy of (top left) the Art Institute of Chicago, gift of Mrs. Gilbert W. Chapman in memory of Charles B. Goodspeed, (top right) the Art Institute of Chicago, restricted gift of Maymar Corporation, Mrs. Maurice L. Rothschild, Mr. and Mrs. Chauncey McCormick, Mary and Leigh Block Charitable Fund, Ada Turnbull Hertle Endowment; through prior gift of Mr. and Mrs. Edwin E. Hokin; photograph, (bottom) Giraudon/Art Resource, New York City



"Guernica," oil painting by Pablo Picasso, 1937. In the Museo Nacional Centro de Arte Reina Sofía, Madrid. 349.3 x 776.6 cm.

Sabartés," 1939; "Retour de Bruxelles," sketchbook number 137, 1956).

After the war Picasso resumed exhibiting his work, which included painting and sculpture as well as work in lithography and ceramics. At the Autumn Salon of 1944 ("Salon de la Liberation") Picasso's canvases and sculpture of the preceding five years were received as a shock. This plus the announcement that Picasso had just joined the Communist Party led to demonstrations against his political views in the gallery itself. At the same time Picasso opened up his studio to both new and old writer and artist friends, including Jean-Paul Sartre, Pierre Reverdy, Éluard, the photographer Brassaï, the English artist Roland Penrose, and the American photographer Lee Miller, as well as many American GI's.

Membership in the Communist Party

Already in 1943 a young painter, Françoise Gilot, had presented herself at the studio, and within months she became the successor to Dora Maar. In 1946 Picasso moved to the Mediterranean with Gilot (with whom he was to have two children, Claude in 1947 and Paloma in 1949). First they moved to Antibes, where Picasso spent four months painting at the Château Grimaldi ("Joie de Vivre," 1946). The paintings of this time and the ceramics he decorated at the studio in nearby Vallauris, beginning in 1947, vividly express Picasso's sense of identification with the classical tradition and with his Mediterranean origins. They also celebrate his new-found happiness with Gilot, who in works of this period is often nymph to Picasso's fauns and centaurs.

Ceramics. Picasso's ceramics are usually set apart from his main body of work and are treated as less important, because at first glance they seem a somewhat frivolous exercise in the decoration of ordinary objects. Plates, jugs, and vases, mostly made by craftsmen at the Madoura pottery in Vallauris, were reshaped or painted, gouged out, scratched, or marked by fingerprints, and, for the most part, were rendered useless. In turning to craft, Picasso worked with a sense of liberation, experimenting with the play between decoration and form (between two and three dimensions) and between personal and universal meaning.

During this period Picasso's fame increasingly attracted numerous visitors, including artists and writers, some of whom (Hélène Parmelin, Édouard Pignon, Éluard, and especially Louis Aragon) encouraged Picasso's further political involvement. Although he contributed designs willingly (his dove was used for the World Peace Congress poster in Wrocław, Pol., in 1949), it was not so much from a commitment to the Communists as from a sincere and lifelong sympathy with any group of repressed people. "War" and "Peace," two panels painted in 1952 to adorn the Temple of Peace attached to an old chapel in Vallauris, reflect Picasso's personal optimism of those years.

The Picasso myth. After World War II an aura of myth grew up around the name of Picasso, and in the last decades of his life his work had, in a sense, moved beyond criticism. Although there were few critics able to keep pace with his latest work, there were few who attacked him. One exception was the British critic John Berger (*The Success and Failure of Picasso,* 1965), who raised questions about Picasso's economic motives and speculated about his inflated public reputation. Picasso's enormous output (especially in printing and drawing) kept his name before the public, even though his work seemed at the time to be far from mainstream, nonfigurative imagery. For example, in the series that characterized the working methods of his late years he used figurative imagery to weave a kind of narrative within each series' numerous variations.

In 1953 Françoise Gilot with their two children left Picasso, and he spent several years as a bachelor, dividing his time between Paris and his home at La Californie, near Cannes. In 1954 he had met Jacqueline Roque, who worked in the pottery shop in Vallauris, and they married in 1961; she not only became his steadfast companion, but also, as his muse, she became the principal image and source of inspiration for practically all of the late work. They are both buried in the castle at Vauvenargues, which Picasso purchased in 1958. But the years from their marriage to Picasso's death they spent at Mougins.

Jacqueline Roque

History of art. In his late work Picasso repeatedly turned toward the history of art for his themes. He seemed at times obsessed with the need to create variations on the works of earlier artists; thus in his many prints, drawings, and paintings of that period, reference is made to artists such as Altdorfer, Manet, Rembrandt, Delacroix, and Courbet. Repeatedly Picasso did a complete series of variations on one particular work, the most famous being perhaps the series on "Las Meninas" of Velázquez consisting of 58 discrete pictures. At times Picasso reworked a specific work because he identified personally with it. For example, he was attracted to Delacroix's "Femmes d'Alger" because the figure on the right bore resemblance to his wife. More often he seemed moved by the challenge to rework in his own way the complex pictorial and narrative problems the older artists had originally posed for themselves. In a sense Picasso was writing himself into the history of art by virtue of such an association with a number of his predecessors.

There is a renewed sense of play in the work of Picasso's later years. He transformed paper cutouts into monumental sculptures, and in Henri-Georges Clouzot's film "Le Mystère Picasso" (1955), the artist, the sole star, behaves like a conjurer, performing tricks with light as well as with his brush. And finally, just as he turned to the paintings of earlier masters, redoing their works in many variations, so he turned to his own earlier oeuvre, prompted by the same impulse. The circus and the artist's studio became once again the stage for his characters, among whom he often placed himself portrayed as an old acrobat or king.

"Le Mystère Picasso"

Picasso died in Mougins, Fr., at the age of 91, on April 8, 1973.

ASSESSMENT

Because Picasso's art from the time of the "Demoiselles" was radical in nature, virtually no 20th-century artist could escape his influence. Moreover, while other masters such as Matisse or Braque tended to stay within the bounds of a style they had developed in their youth, Picasso continued to be an innovator into the last decade of his life. This led to misunderstanding and criticism both in his lifetime and since, and it was only in the 1980s that his last paintings began to be appreciated both in themselves and for their profound influence on the rising generation of young painters. Since Picasso was able from the 1920s to sell works at very high prices, he could keep most of his oeuvre in his own collection. At the time of his death he owned some 50,000 works in various media from every period of his career, which passed into possession of the French state and his heirs. Their exhibition and publication has served to reinforce the highest estimates of Picasso's astonishing powers of invention and execution over a span of more than 80 years.

MAJOR WORKS

PAINTINGS: "Girl with Bare Feet" (1895; Picasso Museum, Paris); "Science and Charity" (1897; Picasso Museum, Barcelona); "Portrait of Carles Casagemas" (1899; Picasso Museum, Barcelona); "Lovers in the Street" (1900; Picasso Museum, Barcelona); "Moulin de la Galette" (1900; Solomon R. Guggenheim Museum, New York City); "Mourners" (1901; private collection, United States); "Evocation" (1901; Museum of Modern Art of the City of Paris); "Self-Portrait" (1901; Picasso Museum, Paris); "El Bock (Portrait of Jaime Sabartés)" (1901; State Pushkin Museum of Fine Arts, Moscow); "The Soup" (1902; Art Gallery of Ontario, Toronto); "Crouching Woman" (1902; Art Gallery of Ontario); "The Two Sisters" (1902; Hermitage, Leningrad); "Blind Man's Meal" (1903; Metropolitan Museum of Art, New York City); "Old Jew and a Boy" (1903; State Pushkin Museum of Fine Arts); "La Vie" (1903; Cleveland Museum of Art); "The Old Guitarist" (1903; Art Institute of Chicago); "The Tragedy" (1903; National Gallery of Art, Washington, D.C.); "Family of Saltimbanques" (1905; National Gallery of Art); "Girl Balancing on a Ball" (1905; State Pushkin Museum of Fine Arts); "The Actor" (1905; Metropolitan Museum of Art); "The Harem" (1906; Cleveland Museum of Art); "La Toilette" (1906; Albright-Knox Art Gallery, Buffalo); "Portrait of Gertrude Stein" (1906; Metropolitan Museum of Art); "Two Nudes" (1906; Museum of Modern Art, New York City); "Seated Female Nude with Crossed Legs" (1906; National Gallery, Prague); "Self-Portrait with Palette" (1906; Philadelphia Museum of Art); "Woman with Loaves" (1906; Philadelphia Museum of Art); "Les Demoiselles d'Avignon" (1907; Museum of Modern Art); "Three Women" (1908–09;

Hermitage); "Bread and Fruit Dish on a Table" (1909; Public Art Collection, Basel, Switz.); "Factory at Horta de Ebro" (1909; Hermitage); "Woman with Pears" (1909; State Pushkin Museum of Fine Arts); "Portrait of Ambroise Vollard" (1910; State Pushkin Museum of Fine Arts); "Portrait of Daniel-Henry Kahnweiler" (1910; Art Institute of Chicago); "The Accordionist" (1911; Solomon R. Guggenheim Museum); "Ma Jolie" (1911–12; Museum of Modern Art); "Still Life with Chair Caning" (1911–12; Picasso Museum, Paris); "Student with a Pipe" (1913; Museum of Modern Art); "Woman in a Chemise" (1913; private collection, United States); "Harlequin" (1915; Museum of Modern Art); "Portrait of Olga" (1917; Picasso Museum, Paris); "Harlequin" (1917; Picasso Museum, Barcelona); "Passeig de Colom" (1917; Picasso Museum, Barcelona); "Three Musicians" (1921; Philadelphia Museum of Art); "Three Musicians" (1921; Museum of Modern Art); "Pipes of Pan" (1923; Picasso Museum, Paris); "Three Dancers" (1925; Tate Gallery, London); "Woman in an Armchair" (1929; Picasso Museum, Paris); "Crucifixion" (1930; Picasso Museum, Paris); "Seated Bather" (1930; Museum of Modern Art); "Bather with Beach Ball" (1932; private collection, United States); "Girl Before a Mirror" (1932; Museum of Modern Art); "The Female Swimmer" (1934; Marina Picasso Collection, Geneva); "Nude Asleep in a Landscape" (1934; Picasso Museum, Paris); "Guernica" (1937; Casón del Buen Retiro, Prado, Madrid); "Weeping Woman" (1937; private collection, England); "Still Life with Red Bull's Head" (1938; Museum of Modern Art); "Portrait of Jaime Sabartès" (1939; Picasso Museum, Barcelona); "Woman Dressing Her Hair" (1940; private collection, United States); "L'Aubade" (1942; National Museum of Modern Art, Beaubourg Centre, Paris); "The Charnel House" (1944–45; Museum of Modern Art); "Joie de Vivre" (1946; Picasso Museum, Antibes, Fr.); "Femme Fleur" (1946; private collection, France); "Peace" (1952; National Picasso Museum, Vallauris, Fr.); "War" (1952; National Picasso Museum); "Women of Algiers, After Delacroix" (series, 1954–55); "Las Meninas, After Velázquez" (series 1957; Picasso Museum, Barcelona); "Rape of the Sabines" (1963; Museum of Fine Arts, Boston); "Man with Pipe and Cupid" (1969; Am Rhyn-Haus, Luzern, Switz.); "The Kiss" (1969; Picasso Museum, Paris); "Young Bather with Sand Shovel" (1971; private collection, France).

SCULPTURE: (Works without ownership identification exist in multiple castings.) "The Jester" (bronze, 1905); "Head of a Woman" (bronze, 1909); "Guitar" (sheet metal and wire, 1912; Museum of Modern Art); "Absinthe Glass" (painted bronze and silver sugar strainer, 1914; six versions); "Guitar" (floor-cloth, string, nails, and newspaper, 1926; Picasso Museum, Paris); "Construction" (iron wire, 1928; Picasso Museum, Paris); "Bust of a Woman" (bronze, 1932; Picasso Museum, Paris); "Bull's Head" (saddle and handlebars, 1943; Picasso Museum, Paris); "Death's Head" (bronze, 1942; Picasso Museum, Paris); "Man with a Sheep" (bronze, 1944); "She-Goat" (plaster, later bronze, 1950); "Bathers" (wood, 1956; State Gallery, Stuttgart, W.Ger.); "Head" (steel, 1967; Daley Center Plaza, Chicago).

WORKS ON PAPER: "The Frugal Repast" (etching, 1904); "Portrait of Max Jacob" (pencil, 1915; private collection, France); "Sleeping Peasants" (tempera with coloured pencil, 1919; Museum of Modern Art); "Portrait of Erik Satie" (pencil, 1920; Picasso Museum, Paris); "Portrait of Igor Stravinsky" (pencil, 1920; Picasso Museum, Paris); illustrations for Balzac's *Chef-d'œuvre inconnu* (13 etchings and 67 wood engravings, 1927 [published 1931]); illustrations for Ovid's *Metamorphoses* (30 etchings, 1930 [published 1931]); "Crucifixion, After Grünewald" (india ink series, 1932; Picasso Museum, Paris); cover of *Minotaure* (collage, 1933; Museum of Modern Art); "Model and Surrealist Sculpture" (etching, 1933); "Sculptor's Studio" (etchings, 1933–34); "Minotauromachia" (etching, 1935); "Minotaur" (ink and gouache series, 1936; Picasso Museum, Paris); "Dream and Lie of Franco" (etchings and aquatints, 1937); "David and Bathsheba, After Cranach the Elder" (lithograph series, 1947); "Luncheon on the Grass, After Manet" (sketchbook number 163, 1962; private collection, Paris); "Suite 347" (etchings and engravings, 1968); "Man and Seated Nude, Mougins" (ink, 1970; Ludwig Museum, Cologne); "Self-Portrait" (crayons, 1972; Fuji TV Company Gallery, Japan).

CERAMICS: "Bullfight" series (plates, 1949); "Centaur" series (plates, 1949); "Visage de Femme" (jug with spout, 1951; private collection, England); "Hands Holding Fish" (plate, 1953; State Pushkin Museum of Fine Arts).

BIBLIOGRAPHY

Biography and Criticism: JAIME SABARTÉS, *Picasso: An Intimate Portrait,* trans. from Spanish by ANGEL FLORES (1948), is a fundamental biographical source, although many Picasso legends begin here. Picasso himself approved the biography by ANTONINA VALLENTIN, *Picasso* (1963). An important recent biography is JOHN RICHARDSON, *A Life of Picasso* (1991–), two volumes of which have already been published (in 1991 and 1996, respectively). Other biographies and memoirs include FERNANDE OLIVIER, *Picasso and His Friends* (1964; originally published in French, 1933); GERTRUDE STEIN, *Picasso* (1938, reprinted 1984); FRANÇOISE GILOT and CARLTON LAKE, *Life with Picasso* (1964, reissued 1990), which are informative and sometimes biased memoirs; BRASSAÏ, *Picasso and Company* (1966, also published as Picasso & Co., 1967; originally published in French, 1964), which contains conversations with the artist; DORE ASHTON (compiler), *Picasso on Art: A Selection of Views* (1972, reprinted 1988); PATRICK O'BRIAN, *Picasso: Pablo Ruiz Picasso: A Biography* (also published as Pablo Ruiz Picasso, 1976); PIERRE DAIX, *La Vie de peintre de Pablo Picasso* (1977), which is a well-documented chronological study of his work; ROLAND PENROSE, *Picasso: His Life and Work,* 3rd ed. (1981); and JOSEP PALAU I. FABRE, *Picasso, the Early Years: 1881–1907* (1981, reissued 1996; originally published in Catalan, 1980), *Picasso Cubism (1907–1917),* trans. from Catalan (1990, reissued 1996), and *Picasso: From the Ballets to Drama (1917–1926)* (1999; originally published in Catalan, 1999).

More specific works of criticism are ALFRED H. BARR, JR., *Picasso: Fifty Years of His Art* (1945, reprinted 1980); GERT SCHIFF (ed.), *Picasso in Perspective* (1976), a collection of essays about his work; and MARILYN MCCULLY (ed.), *A Picasso Anthology: Documents, Criticism, Reminiscences* (1981). RAY ANNE KIBBEY, *Picasso: A Comprehensive Bibliography* (1977), offers limited general coverage through 1976.

Catalogs: Catalogs of the artist's work in individual media include BERNHARD GEISER (ed.), *L'Oeuvre gravé de Picasso* (1955), which examines his prints; PIERRE DAIX and GEORGES BOUDAILLE, *Picasso: The Blue and Rose Periods,* rev. ed. (1967; originally published in French, 1966), which examines his early paintings; DOUGLAS COOPER, *Picasso Theatre* (1968, reissued 1987; originally published in French, 1967); CHRISTIAN ZERVOS, *Pablo Picasso,* 33 vol. (1932–78), a basic work that reproduces the works in black and white; GEORGES BLOCH, *Pablo Picasso,* 4th ed., 4 vol. (1984), which includes printed graphic work from 1904 to 1972 and painted ceramics from 1949 to 1971; PIERRE DAIX and JOAN ROSSELET, *Picasso: The Cubist Years, 1907–1916* (1979; originally published in French, 1979); GARY TINTEROW, *Master Drawings by Picasso* (1981); WERNER SPIES and CHRISTINE PIOT, *Picasso: The Sculptures* (2000; originally published in French, 2000); SEBASTIAN GOEPPERT, HERMA GOEPPERT-FRANK, and PATRICK CRAMER, *Pablo Picasso, the Illustrated Books: Catalogue Raissonné* (1983; originally published in French, 1983); ARNOLD GLIMCHER and MARK GLIMCHER (eds.), *Je Suis le Cahier: The Sketchbooks of Picasso* (1986, reissued 1996); and WERNER SPIES, *Picasso: Pastelle, Zeichnungen, Aquarelle* (1986).

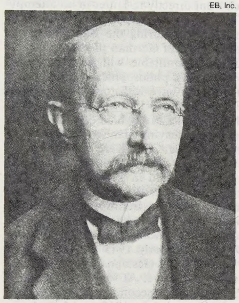
Catalogs addressing specific themes, often in conjunction with museum exhibitions, include WILLIAM RUBIN (ed.), *Pablo Picasso: A Retrospective* (1980); CHRISTIAN GEELHAAR et al., *Pablo Picasso: Das Spätwerk: Themen 1964–1972* (1981); DOUGLAS COOPER and GARY TINTEROW, *The Essential Cubism, 1907–1920: Braque, Picasso & Their Friends* (1983); GERT SCHIFF, *Picasso: The Last Years, 1963–1973* (1983); JÜRGEN GLAESEMER (ed.), *Der Junge Picasso: Frühwerk und blaue Periode* (1984); ELLEN C. OPPLER (ed.), *Picasso's Guernica: Illustrations, Introductory Essay, Documents, Poetry, Criticism, Analysis* (1988); HERSCHEL B. CHIPP, *Picasso's Guernica: History, Transformations, Meanings* (1988); RÉUNION DES MUSÉES NATIONAUX, *Les Demoiselles d'Avignon,* 2 vol. (1988); MUSÉE NATIONAL D'ART MODERNE and TATE GALLERY, *Late Picasso: Paintings, Sculpture, Drawings, Prints, 1953–1972* (1988); WILLIAM RUBIN, *Picasso and Braque: Pioneering Cubism* (1989); WILLIAM RUBIN (ed.), *Picasso and Portraiture: Representation and Transformation* (1996); ELIZABETH COWLING and JOHN GOLDING, *Picasso: Sculptor/Painter* (1994); MARILYN MCCULLY (ed.), *Picasso: The Early Years, 1892–1906* (1997); STEVEN A. NASH and ROBERT ROSENBLUM (eds.), *Picasso and the War Years, 1937–1945* (1998); and MARILYN MCCULLY (ed.), *Picasso: Painter and Sculptor in Clay* (1998).

(M.McCu.)

Planck

Max Planck made many contributions to theoretical physics, but his fame rests primarily on his role as originator of the quantum theory. This theory revolutionized our understanding of atomic and subatomic processes, just as Albert Einstein's theory of relativity revolutionized our understanding of space and time. Together they constitute the fundamental theories of 20th-century physics. Both have forced man to revise some of his most cherished philosophical beliefs, and both have led to industrial and military applications that affect every aspect of modern life.

EB, Inc.



Planck.

Early life. Max Karl Ernst Ludwig Planck was born on April 23, 1858, in Kiel, Germany, the sixth child of a distinguished jurist and professor of law at the University of Kiel. The long family tradition of devotion to church and state, excellence in scholarship, incorruptibility, conservatism, idealism, reliability, and generosity became deeply ingrained in Planck's own life and work. When Planck was nine years old, his father received an appointment at the University of Munich, and Planck entered the city's renowned Maximilian Gymnasium, where a teacher, Hermann Müller, stimulated his interest in physics and mathematics. But Planck excelled in all subjects, and after graduation at age 17 he faced a difficult career decision. He ultimately chose physics over classical philology or music because he had dispassionately reached the conclusion that it was in physics that his greatest originality lay. Music, nonetheless, remained an integral part of his life. He possessed the gift of absolute pitch and was an excellent pianist who daily found serenity and delight at the keyboard, enjoying especially the works of Schubert and Brahms. He also loved the outdoors, taking long walks each day and hiking and climbing in the mountains on vacations, even in advanced old age.

Planck entered the University of Munich in the fall of 1874 but found little encouragement there from physics professor Philipp von Jolly. During his *Wanderjahr* (1877–78) at the University of Berlin, he was unimpressed by the lectures of Hermann von Helmholtz and Gustav Robert Kirchhoff, despite their eminence as research scientists. His intellectual capacities were, however, brought to a focus as the result of his independent study, especially of Rudolf Clausius' writings on thermodynamics. Returning to Munich, he received his doctoral degree in July 1879 (the year of Einstein's birth) at the unusually young age of 21. The following year he completed his *Habilitationsschrift* (qualifying dissertation) at Munich and became a *Privatdozent* (lecturer). In 1885, with the help of his father's professional connections, he was appointed *ausserordentlicher Professor* (associate professor) at the University of Kiel. In 1889, after the death of Kirchhoff, Planck received an appointment to the University of Berlin, where he came to venerate Helmholtz as mentor and colleague. In 1892 he was promoted to *ordentlicher Professor* (full professor). He had only nine doctoral students altogether, but his Berlin lectures on all branches of theoretical physics went through many editions and exerted great influence. He remained in Berlin for the rest of his active life.

Planck recalled that his "original decision to devote myself to science was a direct result of the discovery . . . that the laws of human reasoning coincide with the laws governing the sequences of the impressions we receive from the world about us; that, therefore, pure reasoning can enable man to gain an insight into the mechanism of the [world]. . . ." He deliberately decided, in other words, to become a theoretical physicist at a time when theoretical physics was not yet recognized as a discipline in its own right. But he went further: he concluded that the existence of physical laws presupposes that the "outside world is something independent from man, something absolute, and the quest for the laws which apply to this absolute appeared . . . as the most sublime scientific pursuit in life."

The first instance of an absolute in nature that impressed Planck deeply, even as a *Gymnasium* student, was the law of the conservation of energy, the first law of thermodynamics. Later, during his university years, he became equally convinced that the entropy law, the second law of thermodynamics, was also an absolute law of nature. The second law became the subject of his doctoral dissertation at Munich, and it lay at the core of the researches that led him to discover the quantum of action, now known as Planck's constant h, in 1900.

Development of quantum theory

In 1859–60 Kirchhoff had defined a blackbody as an object that re-emits all of the radiant energy incident upon it; *i.e.,* it is a perfect emitter and absorber of radiation. There was, therefore, something absolute about blackbody radiation, and by the 1890s various experimental and theoretical attempts had been made to determine its spectral energy distribution—the curve displaying how much radiant energy is emitted at different frequencies for a given temperature of the blackbody. Planck was particularly attracted to the formula found in 1896 by his colleague Wilhelm Wien at the Physikalisch-Technische Reichsanstalt (PTR) in Berlin-Charlottenburg, and he subsequently made a series of attempts to derive "Wien's law" on the basis of the second law of thermodynamics. By October 1900, however, other colleagues at the PTR, the experimentalists Otto Richard Lummer, Ernst Pringsheim, Heinrich Rubens, and Ferdinand Kurlbaum, had found definite indications that Wien's law, while valid at high frequencies, broke down completely at low frequencies.

Planck learned of these results just before a meeting of the German Physical Society on October 19. He knew how the entropy of the radiation had to depend mathematically upon its energy in the high-frequency region if Wien's law held there. He also saw what this dependence had to be in the low-frequency region in order to reproduce the experimental results there. Planck guessed, therefore, that he should try to combine these two expressions in the simplest way possible, and to transform the result into a formula relating the energy of the radiation to its frequency.

Planck's formulation was hailed as indisputably correct. To Planck, however, it was simply a guess, a "lucky intuition." If it was to be taken seriously, it had to be derived somehow from first principles. That was the task to which Planck immediately directed his energies, and by December 14, 1900, he had succeeded—but at great cost. To achieve his goal, Planck found that he had to

relinquish one of his own most cherished beliefs, that the second law of thermodynamics was an absolute law of nature. Instead he had to embrace Ludwig Boltzmann's interpretation, that the second law was a statistical law. In addition, Planck had to assume that the oscillators comprising the blackbody and re-emitting the radiant energy incident upon them could not absorb this energy continuously but only in discrete amounts, in quanta of energy; only by statistically distributing these quanta, each containing an amount of energy $h\nu$ proportional to its frequency, over all of the oscillators present in the blackbody could Planck derive the formula he had hit upon two months earlier. He adduced additional evidence for the importance of his formula by using it to evaluate the constant h (his value was 6.55×10^{-27} erg-second, close to the modern value), as well as the so-called Boltzmann constant (the fundamental constant in kinetic theory and statistical mechanics), Avogadro's number, and the charge of the electron. As time went on physicists recognized ever more clearly that—because Planck's constant was not zero but had a small but finite value—the microphysical world, the world of atomic dimensions, could not in principle be described by ordinary classical mechanics. A profound revolution in physical theory was in the making.

Planck's concept of energy quanta, in other words, conflicted fundamentally with all past physical theory. He was driven to introduce it strictly by the force of his logic; he was, as one historian put it, a reluctant revolutionary. Indeed, it was years before the far-reaching consequences of Planck's achievement were generally recognized, and

Einstein's role

in this Einstein played a central role. In 1905, independently of Planck's work, Einstein argued that under certain circumstances radiant energy itself seemed to consist of quanta (light quanta, later called photons), and in 1907 he showed the generality of the quantum hypothesis by using it to interpret the temperature dependence of the specific heats of solids. In 1909 he introduced the wave–particle duality into physics. In October 1911 he was among the group of prominent physicists who attended the first Solvay conference in Brussels. The discussions there stimulated Henri Poincaré to provide a mathematical proof that Planck's radiation law necessarily required the introduction of quanta—a proof that converted James (later Sir James) Jeans and others into supporters of the quantum theory. In 1913 Niels Bohr also contributed greatly to its establishment through his quantum theory of the hydrogen atom. Ironically, Planck himself was one of the last to struggle for a return to classical theory, a stance he later regarded not with regret but as a means by which he had thoroughly convinced himself of the necessity of the quantum theory. Opposition to Einstein's radical light quantum hypothesis of 1905 persisted until after the discovery of the Compton effect in 1922.

Later life. Planck was 42 years old in 1900 when he made the famous discovery that in 1918 won him the Nobel Prize for Physics and that brought him many other honours. It is not surprising that he subsequently made no discoveries of comparable importance. Nevertheless, he continued to contribute at a high level to various branches of optics, thermodynamics and statistical mechanics, physical chemistry, and other fields. He was also the first prominent physicist to champion Einstein's special theory of relativity (1905). "The velocity of light is to the Theory of Relativity," Planck remarked, "as the elementary quantum of action is to the Quantum Theory; it is its absolute core." In 1914 Planck and the physical chemist Walther Hermann Nernst succeeded in bringing Einstein to Berlin, and after the war, in 1919, arrangements were made for Max von Laue, Planck's favourite student, to come to Berlin as well. When Planck retired in 1928, another prominent theoretical physicist, Erwin Schrödinger, the originator of wave mechanics, was chosen as his successor. For a time, therefore, Berlin shone brilliantly as a centre of theoretical physics—until darkness enveloped it in January 1933 with the ascent of Adolf Hitler to power.

In his later years, Planck devoted more and more of his writings to philosophical, aesthetic, and religious questions. Together with Einstein and Schrödinger, he

Worldview

remained adamantly opposed to the indeterministic, statistical worldview introduced by Bohr, Max Born, Werner Heisenberg, and others into physics after the advent of quantum mechanics in 1925–26. Such a view was not in harmony with Planck's deepest intuitions and beliefs. The physical universe, Planck argued, is an objective entity existing independently of man; the observer and the observed are not intimately coupled, as Bohr and his school would have it.

Planck became permanent secretary of the mathematics and physics sections of the Prussian Academy of Sciences in 1912 and held that position until 1938; he was also president of the Kaiser Wilhelm Society (now the Max Planck Society) from 1930 to 1937. These offices and others placed Planck in a position of great authority, especially among German physicists; seldom were his decisions or advice questioned. His authority, however, stemmed fundamentally not from the official appointments he held but from his personal moral force. His fairness, integrity, and wisdom were beyond question. It was completely in character that Planck went directly to Hitler in an attempt to reverse Hitler's devastating racial policies and that he chose to remain in Germany during the Nazi period to try to preserve what he could of German physics.

Planck was a man of indomitable will. Had he been less stoic, and had he had less philosophical and religious conviction, he could scarcely have withstood the tragedies that entered his life after age 50. In 1909, his first wife, Marie Merck, the daughter of a Munich banker, died after 22 years of happy marriage, leaving Planck with two sons and twin daughters. The elder son, Karl, was killed in action in 1916. The following year, Margarete, one of his daughters, died in childbirth, and in 1919 the same fate befell Emma, his other daughter. World War II brought further tragedy. Planck's house in Berlin was completely destroyed by bombs in 1944. Far worse, the younger son, Erwin, was implicated in the attempt made on Hitler's life on July 20, 1944, and in early 1945 he died a horrible death at the hands of the Gestapo. That merciless act destroyed Planck's will to live. At war's end, American officers took Planck and his second wife, Marga von Hoesslin, whom he had married in 1910 and by whom he had had one son, to Göttingen. There, on October 4, 1947, in his 89th year, he died. Death, in the words of James Franck, came to him "as a redemption." (R.H.St.)

BIBLIOGRAPHY. Editions of Planck's works include *The Theory of Heat Radiation* (1914, reprinted 1991; originally published in German, 2nd rev. ed., 1913); *Where Is Science Going?*, trans. from German (1932, reprinted 1981), discussing free will and determinism; and *The Philosophy of Physics*, trans. from German (1936, reissued 1963). Planck described his life and work in his *Scientific Autobiography, and Other Papers*, trans. from German (1949, reissued 1968).

HANS KANGRO, "Max Karl Ernst Ludwig Planck," in CHARLES COULSTON GILLISPIE (ed.), *Dictionary of Scientific Biography*, vol. 11 (1975), pp. 7–17, contains an excellent short biography. J.L. HEILBRON, *The Dilemmas of an Upright Man: Max Planck as Spokesman for German Science* (1986), concentrates on the moral dilemmas Planck faced.

Technical books that treat Planck's work and the history of quantum physics include EDMUND WHITTAKER, *A History of the Theories of Aether and Electricity*, rev. and enlarged ed., vol. 2, *The Modern Theories, 1900–1926* (1953, reissued 1987); MAX JAMMER, *The Conceptual Development of Quantum Mechanics* (1966, reissued 1989); ARMIN HERMANN, *The Genesis of Quantum Theory (1899–1913)* (1971; originally published in German, 1969); ROGER H. STUEWER, *The Compton Effect: Turning Point in Physics* (1975); HANS KANGRO, *Early History of Planck's Radiation Law* (1976; originally published in German, 1970); and THOMAS S. KUHN, *Black-Body Theory and the Quantum Discontinuity, 1894–1912* (1978, reprinted 1987).

Nontechnical books include EMILIO SEGRÈ, *From X-Rays to Quarks: Modern Physicists and Their Discoveries* (1980); ILSE ROSENTHAL-SCHNEIDER, *Reality and Scientific Truth: Discussions with Einstein, von Laue, and Planck* (1980); and ALEX KELLER, *The Infancy of Atomic Physics: Hercules in His Cradle* (1983). Especially noteworthy are three articles by MARTIN J. KLEIN: "Max Planck and the Beginning of the Quantum Theory," *Archive for History of Exact Sciences*, 1(5):459–479 (1962), "Planck, Entropy, and Quanta, 1901–1906," *The Natural Philosopher*, 1:83–108 (1963), and "Thermodynamics and Quanta in Planck's Work," *Physics Today*, 19:23–32 (1966).

(R.H.St./Ed.)

Plants

Plants, as members of the kingdom Plantae, share the following features common to all living things. Their cells undergo complex metabolic reactions that result in the production of chemical energy, nutrients, and new structural components. They respond to internal and external stimuli in a self-preserving manner. They reproduce by passing their genetic information to descendants that resemble them. They have evolved over geologic time by the process of natural selection into a wide array of forms and life-history strategies.

Plants range in size from diminutive duckweeds only a few millimetres in length to the giant sequoias of California that reach 90 metres (300 feet) or more in height. There are approximately 275,000 to 300,000 different species of plants, and new species are continuously being described, particularly from unexplored tropical areas of the world. Having evolved from aquatic ancestors, plants have subsequently migrated over the entire surface of the Earth, inhabiting tropical, Arctic, desert, and Alpine regions. Some plants have returned to an aquatic habitat in either fresh or salt water.

Plants play a vital role in the maintenance of life on Earth. All energy used by living organisms depends on the complex process of photosynthesis, which is carried out by green plants. Radiant energy from the Sun is transformed into organic chemical energy in the form of sugar through a fundamental series of chemical reactions called photosynthesis. In nature all food chains begin with green plants (primary producers). Primary producers, represented by trees, shrubs, and herbs, are a prolific source of energy in the form of carbohydrates (sugars) stored in the leaves. These carbohydrates, produced in photosynthesis, are broken down in a process called respiration; the smaller units of the sugar molecule and its products fuel numerous metabolic processes. Various parts of the plant (*e.g.*, leaves) are the energy sources that support animal life in different community habitats. A by-product of photosynthesis, oxygen, is essential to animals.

The daily existence of human beings is also directly influenced by plants. Plants furnish food and flavourings; raw materials for industry, such as wood, resins, oils, and rubber; fibres for the manufacture of fabrics and cordage; medicines; insecticides; and fuels. More than half of the Earth's population relies on the grasses rice, corn (maize), and wheat as their primary source of food. Apart from their commercial and aesthetic value, plants conserve other natural resources by protecting soils from erosion, by controlling water levels and quality, and by producing a favourable atmosphere.

The following article summarizes the morphological, physiological, and ecological features of plants. The principal focus is on structure and function, physiology, life histories, and ecology, and on how the various plant groups have evolved, dispersed, and become adapted to life on land. The features that define each major plant group and the role they play in the wider ecosystem also are discussed.

For further information, see PHOTOSYNTHESIS and BIOCHEMICAL COMPONENTS OF ORGANISMS.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 313, and the *Index*.

The article is divided into the following sections:

Definition of the kingdom 867
Nonvascular plants 868
  Definition of the category
  Representative members
Vascular plants 869
  Definition of the category
  Nonseed plants
  Seed plants
Reproduction and life histories 872
  Life histories
  Variations involving seed plants
  Asexual reproduction
  Deviations from the usual life history
Plant physiology 874
  General features of plant nutrition
  Pathways and cycles
  Unique features of plant metabolism
  Photosynthesis
Ecology 876
  Plant geography
  Succession and zonation
  Ecosystems and the biosphere
  Dispersal and colonization
  Human effects on plants and natural communities
  Conservation
Evolution and paleobotany 879
Classification 881
  Annotated classification
  Critical appraisal
Bibliography 882

DEFINITION OF THE KINGDOM

The kingdom Plantae includes organisms that range in size from a tiny moss to a giant tree. Despite this enormous variation, all plants are multicellular and eukaryotic (*i.e.*, each cell possesses a membrane-bound nucleus that contains the chromosomes). They generally possess pigments (chlorophylls *a* and *b* and carotenoids), which play a central role in converting the radiant energy of sunlight into chemical energy (in the form of sugar) by means of photosynthesis. Plants, therefore, are independent in their nutritional needs (autotrophic) and store their excess food in the form of macromolecules of starch. The relatively few plants that are not autotrophic have lost pigments and are dependent on other organisms for nutrients (parasitism). Plants are nonmotile organisms, although some produce motile cells (gametes) propelled by whiplike flagella. Plant cells are surrounded by a more or less rigid cell wall composed of the carbohydrate cellulose, and adjacent cells are interconnected by microscopic strands of cytoplasm called plasmodesmata, which traverse the cell walls. Many plants have the capacity for unlimited growth at localized regions of cell division, called meristems. Plants, unlike animals, can use inorganic forms of the element nitrogen (N), such as nitrates and ammonia, and thus do not require an external source of protein (in which nitrogen is a major constituent) to survive.

The life histories of plants include two phases, or generations, one of which is diploid (the nuclei of the cells contain two sets of chromosomes), while the other is haploid (with one set of chromosomes). The diploid generation is known as the sporophyte, which literally means spore-producing plant. The haploid generation, called the gametophyte, produces the sex cells, or gametes. The complete life cycle of a plant thus involves an alternation of haploid and diploid generations. The sporophyte and gametophyte generations of plants are structurally quite dissimilar.

Changes in classifications

The concept of what constitutes a plant has undergone significant change over time. At one time the photosynthetic aquatic organisms commonly referred to as algae were considered members of the plant kingdom. The various major algal groups, such as the green algae, brown algae, and red algae, are now placed in the kingdom Pro-

tista because they lack one or more of the features that are characteristic of plants. The organisms known as fungi also were once considered to be plants since they reproduce by spores and possess a cell wall. The fungi, however, uniformly lack chlorophyll (thus they are heterotrophic) and are chemically distinct from the plants; they are now placed in a separate kingdom, Fungi.

The earliest plants undoubtedly evolved from an aquatic, green algal ancestor (as evidenced by similarities in pigmentation, cell wall chemistry, biochemistry, and method of cell division), and different plant groups have become adapted to terrestrial life to varying degrees. Land plants face severe environmental threats or difficulties, such as desiccation, drastic changes in temperature, support, nutrient availability to each of the cells of the plant, regulation of gas exchange between the plant and the atmosphere, and successful reproduction. Thus, many adaptations to land existence have evolved in the plant kingdom and are reflected among the different major plant groups. An example is the development of a waxy covering (the cuticle) that covers the plant body, preventing excess water loss. Specialized tissues and cells (vascular tissue) enabled early land plants to absorb and transport water and nutrients to distant parts of the body more effectively and, eventually, to develop a more complex body composed of organs called stems, leaves, and roots. The evolution and incorporation of the substance lignin into the cell walls of plants provided strength and support. Details of the life history are often a reflection of a plant's adaptation to a terrestrial mode of life and may characterize a particular group; for example, the most highly evolved plants reproduce by means of seeds, and, in the most advanced of all plants (angiosperms), a reproductive organ called a flower is formed.

NONVASCULAR PLANTS

Definition of the category. The plant division Bryophyta includes tiny plants that lack specialized vascular tissue (xylem and phloem) for internal water and food conduction and support. Bryophytes are, therefore, nonvascular plants and, correlatively, possess no true roots, stems, or leaves. Some larger mosses, however, contain a central core of elongated, thick-walled cells called hydroids that are involved in water conduction and that have been compared to the xylem elements of other plants. The Bryophyta are second in diversity among the land plants only to the flowering plants (angiosperms) and are generally regarded as composed of three classes; some evidence indicates, however, that these groups may not be so closely related: Musci (the mosses), Hepaticae (the liverworts), and Anthocerotae (the hornworts).

Adaptations against desiccation

Since bryophytes generally lack conducting cells and a well-developed cuticle that would limit dehydration, they depend on their immediate surroundings for an adequate supply of moisture. As a result, most bryophytes live in moist or wet, shady locations, growing on rocks, trees, and soil. Some, however, have become adapted to totally aquatic habitats; others have become adapted to alternately wet and dry environments by growing during wet periods and becoming dormant during dry intervals. Although bryophytes are widely distributed, occurring in practically all parts of the world, none are found in salt water. Ecologically, some mosses are considered pioneer plants because they can invade bare areas.

Bryophytes are typically land plants but seldom attain a height of more than a few centimetres. They possess the photosynthetic pigment chlorophyll (both *a* and *b* forms) and carotenoids in cell organelles called chloroplasts. The life histories of these plants show a well-defined alternation of generations, with the independent and free-living gametophyte as the dominant photosynthetic phase in the life cycle. (This is in contrast to the vascular plants, in which the dominant photosynthetic phase is the sporophyte.) The sporophyte generation develops from, and is almost entirely parasitic on, the gametophyte. The gametophyte produces multicellular sex organs (gametangia). Female gametangia are called archegonia; and male gametangia, antheridia. At maturity, archegonia each contain one egg, and antheridia produce many sperm. Because the egg is retained and fertilized within the archegonium, the early stages of the developing sporophyte are protected and nourished by the gametophytic tissue. The young undifferentiated sporophyte is called an embryo. Although bryophytes have become adapted to life on land, an apparent vestige of their aquatic ancestry is that the motile (flagellated) sperm depend on water to effect gamete transport and fertilization.

Bryophytes are widely believed to have evolved from complex green algae that invaded land more than 400 million years ago. Bryophytes share some traits with green algae, such as motile sperm, similar photosynthetic pigments, and the general absence of vascular tissue. However, bryophytes have multicellular reproductive structures, whereas those of green algae are unicellular, and bryophytes are mostly terrestrial and have complex plant bodies, while the green algae are primarily aquatic and have less complex forms. An alternative view envisions the bryophytes as directly derived from early vascular plants or as an early offshoot of the lineage that gave rise to vascular plants.

Representative members. *Class Musci.* Moss is a term erroneously applied to many different plants (Spanish moss, a flowering plant; Irish moss, a red alga; pond moss, filamentous algae; and reindeer moss, a lichen). Mosses are classified as the class Musci in the division Bryophyta. Two major groups of mosses are recognized, the true mosses and the commercially important peat mosses (*Sphagnum*). (The granite mosses form a small group comprising the genera *Andreaea* and *Andreaeobryum*.) *Sphagnum* often grows in dense mats in acidic boglands. Because the peat mosses are very absorbent, they are widely used to improve soil texture and to surround plant roots during shipment and replanting in order to prevent desiccation.

Moss structure

The moss gametophyte possesses leaflike structures (phyllids) that usually are a single cell layer thick, have a costa (midrib), and are spirally arranged on a stemlike axis (caulid). The moss gametophyte is an independent plant and is the familiar, erect "leafy" shoot. Multicellular rhizoids anchor the gametophyte to the substrate. The sporophyte plant develops from the tip of the fertile leafy shoot. After repeated cell divisions, the young sporophyte (embryo) transforms into a mature sporophyte consisting of foot, elongate seta, and capsule. The capsule is often covered by a calyptra, which is the enlarged remains of the archegonium. The capsule is capped by an operculum (lid), which falls off, exposing a ring of teeth (the peristome) that regulates the dispersal of spores.

Class Hepaticae. Liverworts, the second major class of bryophytes, are found in the same types of habitat as mosses, and species of the two classes are often intermingled on the same site. The curious name "liverwort" is a relic of the medieval belief in the "doctrine of signatures," which held that the external form of a plant provided a clue to which diseased body organ could be cured by a preparation made from that particular plant. There are two types of liverworts (also called hepatics) based on reproductive features and thallus structure. The more numerous "leafy" liverworts superficially resemble mosses, but most notably differ in having lobed or divided leaves that are without a midrib and which are positioned in three rows. Thalloid (thallose) liverworts have a ribbonlike, or strap-shaped, body that grows flat on the ground. They have a high degree of internal structural differentiation into photosynthetic and storage zones. Liverwort gametophytes have unicellular rhizoids. Liverworts have an alternation of generations similar to that of mosses and, as with mosses, the gametophyte generation is dominant. The sporophytes, however, are not microscopic and are often borne on specialized structures. They sometimes resemble small umbrellas and are called antheridiophores and archegoniophores.

Class Anthocerotae. The third class of bryophytes comprises the hornworts, a minor group numbering fewer than 100 species. The gametophyte is a small, ribbonlike thallus that resembles a thallose liverwort. The name hornwort is derived from the unique, slender, upright sporophytes, which are about three to four centimetres long at maturity and dehisce longitudinally into two valves that twist in

response to changing humidity, thereby releasing spores in small numbers over a fairly long period of time.

VASCULAR PLANTS

Definition of the category. Vascular plants (tracheophytes) differ from the nonvascular bryophytes in that they possess specialized supporting and water-conducting tissue, called xylem, and food-conducting tissue, called phloem. The xylem is composed of nonliving cells (tracheids and vessel elements) that are stiffened by the presence of lignin, a hardening substance that reinforces the cellulose cell wall. The living sieve elements that comprise the phloem are not lignified. Xylem and phloem are collectively called vascular tissue and form a central column (stele) through the plant axis. The ferns, gymnosperms, and flowering plants are all vascular plants. Because they possess vascular tissues, these plants have true stems, leaves, and roots. Before the development of vascular tissues, the only plants of considerable size existed in aquatic environments where support and water conduction were not necessary. A second major difference between the vascular plants and bryophytes is that the larger, more conspicuous generation among vascular plants is the sporophytic phase of the life cycle.

Xylem and phloem

The vegetative body of vascular plants is adapted to terrestrial life in various ways. In addition to vascular tissue, the aerial body is covered with a well-developed waxy layer (cuticle) that decreases water loss. Gases are exchanged through numerous pores (stomata) in the outer cell layer. The root system is involved in the uptake from the soil of water and minerals that are used by the root system as well as the stem and leaves. Roots also anchor the plant and store food. The stem conducts water and minerals absorbed by the root system upward to various parts of the stem and leaves; stems also conduct carbohydrates manufactured through the process of photosynthesis from the leaves to various parts of the stem and root system. Leaves are supported by the stem and are oriented in a manner conducive to maximizing the amount of leaf area involved in trapping sunlight for use in photosynthesis.

Modifications of roots, stems, and leaves have enabled species of vascular plants to survive in a variety of habitats encompassing diverse and even extreme environmental conditions. The ability of vascular plants to flourish in so many different habitats is a key factor in their having become the dominant group of terrestrial plants.

Major divisions

The vascular plants are divisible into the nonseed plants (lower vascular plants, or cryptogams) and those that reproduce by seeds (higher vascular plants, or phanerogams). The ferns (Filicophyta) are a group of the lower vascular plants; other groups include the whisk ferns (Psilotophyta), club and spike mosses (Lycophyta), and horsetails (Sphenophyta, or Arthrophyta). Collectively, the latter four groups are sometimes referred to as pteridophytes, since each reproduces by spores liberated from dehiscent sporangia (free sporing). Although the lower vascular plants have adapted to terrestrial life, they are similar to bryophytes in that, as an apparent vestige of their aquatic ancestry, all produce motile (flagellated) male gametes (antherozoids, or sperm) and must rely on water for fertilization to be effected.

Nonseed plants. *Division Filicophyta.* Ferns are a diverse group of plants technically classified in the division Filicophyta. Although they have a worldwide distribution, ferns are more common in tropical and subtropical regions. They range in size and complexity from small floating aquatic plants less than 2 centimetres long to tall tree ferns 20 metres (65 feet) high. Tropical tree ferns possess erect, columnar trunks and large, compound (divided) leaves more than 5 metres long. As a group, ferns are either terrestrial or epiphytic (growing upon another plant). Fern stems never become woody (composed of secondary tissue containing lignin) since all tissues of the plant body originate at the stem apex.

Ferns typically possess a rhizome (horizontal stem) that grows partially underground; the deeply divided fronds (leaves) and the roots grow out of the rhizome. Fronds are characteristically coiled in the bud (fiddleheads) and uncurl in a type of leaf development called circinate vernation. Fern leaves are either whole or variously divided. The leaf types are differentiated into rachis (axis of a compound leaf), pinnae (primary divisions), and pinnules (ultimate segments of a pinna). Fern leaves often have prominent epidermal hairs and large chaffy scales. Venation of fern leaves is usually open dichotomous (forking into two equal parts).

Each frond is a potential sporophyll (spore-bearing leaf) and as such can bear structures that are associated with reproduction. When growth conditions are favourable, a series of brown patches appear on the undersurface of the sporophylls. Each one of the patches (called a sorus) is composed of many sporangia, or spore cases, which are joined by a stalk to the sporophyll. The spore case is flattened, with a layer of sterile, or nonfertile, cells surrounding the spore mother cells. Each spore mother cell divides by reduction division (meiosis) to produce haploid spores, which are shed in a way characteristic to the ferns.

Spores

Each fern spore has the potential to grow into a green, heart-shaped, independent gametophyte plant (prothallus) capable of photosynthesis. In contrast to bryophytes, in which the sporophyte is nutritionally dependent on the gametophyte during its entire existence, the fern sporophyte is dependent on the gametophyte for nutrition only during the early phase of its development; thereafter, the fern sporophyte is free-living. In some ferns the sexes are separate, meaning a gametophyte will bear only male or female sex organs. Other species have gametophytes bearing both sex organs. Features important in the identification of ferns include such aspects of the mature sporophyte plant as differences in the stem, frond, sporophyll, sporangium, and position of the sporangium and the absence or presence, as well as shape, of the indusium (a membranous outgrowth of the leaf) covering the sporangia.

Division Psilotophyta. Psilotophyta (whisk ferns) is a division represented by two living genera (*Psilotum and Tmesipteris*) and several species that are restricted to the subtropics. This unusual group of small herbaceous plants is characterized by a leafless and rootless body possessing a stem that exhibits a primitive dichotomous type of branching: it forks into equal halves. The photosynthetic function is assumed by the stem, and the underground rhizome anchors the plant. The vascular tissue is organized into a poorly developed central cylinder in the stem.

Division Lycophyta. This division is represented by four or more living genera, with the principal genera being *Lycopodium* (club mosses), *Selaginella* (spike mosses), and *Isoetes* (quillworts). Extant members of Lycophyta occur in both temperate and tropical regions and represent the survivors of a group of vascular plants that was extremely diverse and numerous. As a group, the lycopods were prominent in the great coal-forming swamp forests of the Carboniferous Period (360 to 280 million years ago). Although all living lycopods are small herbaceous plants, some extinct types were large trees. Lycopods are differentiated into stem, root, and leaf (microphylls). Sporangia are positioned on the upper (adaxial) surface of the leaf (sporophyll). Some species form distinct cones or strobili, while others do not.

Division Sphenophyta. Sphenophyta (also called horsetails and scouring rushes) is a division represented by a single living genus (*Equisetum*). It has a worldwide distribution but occurs in greater variety in the Northern Hemisphere. Like the lycopods, the sphenophytes were a diverse and prominent group of vascular plants during the Carboniferous Period, when some genera attained great size in the coal-forming swamp forests. Sphenophytes are differentiated into stem, leaf (microphylls), and root. Green aerial stems have longitudinal ridges and furrows extending the length of the internodes, and stems are jointed (articulated). Surface cells are characteristically filled with silica. Branches, when they occur, are borne in whorls at the node, as are the scale leaves. Sporangia are borne in terminal strobili. The Sphenophyta are an independent line of vascular plant evolution that had its origin in the Devonian Period (408 to 360 million years ago).

Seed plants. Gymnosperms and angiosperms (flowering plants) share with ferns a dominant, independent sporophyte generation; the presence of vascular tissue; differentiation of the plant body into root, stem, and leaf

derived from a bipolar embryo (having stem and root-growing apexes); and similar photosynthetic pigments. Unlike ferns, however, the seed plants have stems that branch laterally and vascular tissue that is arranged in strands (bundles) around the pith (eustele). Among seed plants, as in ferns, the stem tissues that arise directly from the shoot apex are called primary tissues. Primary tissues contribute to the longitudinal growth of the stem, or primary growth. Secondary growth, resulting in an increase in the width of the axis, is produced by meristematic tissue between the primary xylem and phloem called vascular cambium. This meristem consists of a narrow zone of cells that form new secondary xylem (wood) and secondary phloem (secondary vascular tissues).

Major evolutionary advancements of these plants are demonstrated by the generally more complex plant body and by reproduction via seeds. Seeds represent an important evolutionary innovation within the plant kingdom. Each seed has an embryonic plant (sporophyte), food-storage tissue, and hardened protective covering (seed coat). The seed thus contains and protects the embryonic plant and, as the primary dispersal unit of the seed plants, represents a significant improvement over the spore, with its limited capacity for survival.

Seeds

In comparing ferns and seed plants and their life histories, certain significant differences are seen. The gametophyte in seed plants has been reduced in size, usually consisting of a few to a dozen cells. Thus, it is no longer itself a plant body, as in the bryophytes and ferns. The gametophyte is not free-living but is embedded in the sporophyte and is thus less vulnerable to environmental stress than the gametophytes of bryophytes and ferns. Finally, the spores of seed plants are male and female, as are the sporangia that contain them. The spores are not dispersed as in the bryophytes and ferns but develop into gametophytes within the sporangia. In the most advanced seed plants, the male gametes (sperm) are carried to the egg by a later extension of the pollen grain called the pollen tube. The advantage of this system is that the nonflagellated sperm are no longer dependent on water to reach the egg.

Another terrestrial adaptation of the seed plants not found in ferns is pollen dispersed by wind or animals. Pollen is a unit of genetic material as well as part of the seed formation process. The dispersal of pollen by wind or insects, in addition to dispersal of seeds, promotes genetic recombination and distribution of the species over a wide geographic area.

Gymnosperms. The term gymnosperm ("naked seeds") represents four extant divisions of vascular plants whose ovules (seeds) are exposed on the surface of cone scales. The cone-bearing gymnosperms are among the largest and oldest living organisms in the world. They dominated the landscape about 200 million years ago. Today gymnosperms are of great economic value as major sources of lumber products, pulpwood, turpentine, and resins.

Naked seeds

Conifer stems are composed of a woody axis containing primitive water- and mineral-conducting cells called tracheids. Tracheids are interconnected by passages called bordered pits. Leaves are often needlelike or scalelike and typically contain canals filled with resin. The leaves of pine are borne in bundles (fascicles), and the number of leaves per fascicle is an important distinguishing feature. Most gymnosperms are evergreen, but some, such as larch and bald cypress, are deciduous (the leaves fall after one growing season). The leaves of many gymnosperms are adapted to water conservation in having a thick cuticle and in having a stomata below the leaf surface.

The tree or shrub is the sporophyte generation. In conifers, the male and female sporangia are produced on separate structures called cones or strobili. Individual trees are typically monoecious (male and female cones are borne on the same tree). A cone is a modified shoot with a single axis, on which is borne a spirally arranged series of pollen- or ovule-bearing scales or bracts. The male cone, or microstrobilus, is usually smaller than the female cone (megastrobilus) and is essentially an aggregation of many small structures (microsporophylls) that encase the pollen in microsporangia.

The extant cycads (division Cycadophyta) are a group of ancient seed plants that are survivors of a complex which has existed since the Mesozoic Era (beginning 245 million years ago). They are presently distributed in the tropics and subtropics of both hemispheres. Cycads are palmlike in general appearance, with an unbranched, columnar trunk and a crown of large, pinnately compound (divided) leaves. The sexes are always separate, resulting in male and female plants (*i.e.,* cycads are dioecious). Most species produce conspicuous cones (strobili) on both male and female plants, and the seeds are very large.

The ginkgophytes (division Ginkgophyta), although abundant, diverse, and widely distributed in the past, are represented now by a sole surviving species, *Ginkgo biloba* (maidenhair tree). The species was formerly restricted to southeastern China, but it is now probably extinct in the wild. The plant is commonly cultivated worldwide, however, and is particularly resistant to disease and air pollution. The ginkgo is multibranched, with stems that are differentiated into long shoots and dwarf (spur) shoots. A cluster of fan-shaped, deciduous leaves with open dichotomous venation occurs at the end of each lateral spur shoot. Sexes are separate, and distinct cones are not produced. Female trees produce plumlike seeds with a fleshy outer layer and are noted for their foul smell when mature.

Ginkgo

The gnetophytes (division Gnetophyta) comprise a group of three unusual genera. *Ephedra* occurs as a shrub in dry regions in tropical and temperate North and South America and in Asia, from the Mediterranean Sea to China. Species of *Gnetum* occur as woody shrubs, vines, or broad-leaved trees and grow in moist tropical forests of South America, Africa, and Asia. *Welwitschia,* restricted to extreme deserts (less than 25 millimetres [1 inch] of rain per year) in a narrow belt about 1,000 kilometres (600 miles) long in southwestern Africa, is an unusual plant composed of an enormous underground stem and a pair of long, strap-shaped leaves that lie along the ground. The three genera differ from all other gymnosperms in possessing vessel elements (as compared to tracheids) in the xylem and in specializations in reproductive morphology. The gnetophytes have figured prominently in the theories about gymnospermous origins of the angiosperms.

Angiosperms. Approximately 130 million years ago flowering plants (angiosperms) evolved from gymnosperms, although the identity of the specific gymnospermous ancestral group remains unresolved. The primary distinction between gymnosperms and angiosperms is that angiosperms reproduce by means of flowers. Flowers are modified shoots bearing a series of leaflike, modified appendages and containing ovules (immature seeds) surrounded and protected by the female reproductive structure, the carpel or pistil. Along with other features, angiospermy, the enclosed condition of the seed, gave the flowering plants a competitive advantage and enabled them to come to dominate the extant flora. Flowering plants have also fully exploited the use of insects and other animals as agents of pollination (the transfer of pollen from male to female floral structures). In addition, the water-conducting cells and food-conducting tissue are more complex and efficient in flowering plants than in other land plants. Finally, flowering plants possess a specialized type of nutritive tissue in the seed, endosperm. Endosperm is the chief storage tissue in the seeds of grasses; hence, it is the primary source of nutrition in corn, rice, wheat, and other cereals that have been utilized as major food sources by humans and other animals.

The flowering plants are represented by two divergent evolutionary lines, the monocotyledons and the dicotyledons, treated as separate classes within the division. These two major groups are distinguished by the number of embryonic seed leaves (cotyledons), arrangement of vascular tissue in the stem, leaf venation, and manner of leaf attachment to the stem. A further distinction between the two is in the number of flower parts. Generally, monocots have flower parts in multiples of three, and dicots have flower parts in multiples of four or five. The pollen of monocots is uniaperturate (with a single germinal aperture or germ pore), whereas the pollen of dicots is most commonly triaperturate or a derived form.

The plant body of angiosperms consists of a central axis

of two parts, the shoot and root. Shoots have two kinds of organs, the stem and the leaves, while roots have one type of organ, the root itself. Systems of classification are often based upon the longevity of the portions of plant above ground. Woody plants are trees and shrubs whose shoots are durable and survive over a period of years. They are further classified into deciduous and evergreen plants. Deciduous plants drop their leaves at the end of every growing season, while evergreens keep their leaves until the leaves of the following year are produced. Herbaceous plants have soft, flexible aerial portions that die each year.

Annuals, biennials, and perennials

Another system of classification, based on the duration of the life history, is particularly applicable to angiosperms of the temperate region. Annuals are plants that complete the entire life history (germinate from seeds, mature, flower, and produce seed) in one growing season. Examples of annuals are corn, wheat, peas, and tobacco. Biennials complete their life history in two seasons, blooming during the second season. Beets, celery, cabbage, carrots, and turnips are biennials, but their flowers are rarely seen since they are harvested during the first season. Annuals and biennials are both generally herbaceous plants. Perennials are plants that live from year to year. Trees and shrubs are perennial, but some herbaceous plants are also perennials.

A number of modifications of the stem occur in angiosperms, and many of these modifications provide a means for herbs to become dormant and survive for a period of years. Rhizomes are horizontally growing underground stems that serve as organs of asexual reproduction and food storage. Tubers are rhizomes with thickened portions (for example, the Irish potato). Corms are short, upright underground stems surrounded by a few thin scale leaves (as in *Crocus* and *Gladiolus*). Bulbs have a greatly reduced stem with thick, fleshy scale leaves surrounding it (as in the onion). Runners are thin, surface stems characteristic of such plants as strawberries; new plants may form on the runner as it spreads along the ground. Stolons are like runners and extend along the ground. Many of the most prolific weeds have stolons by which they propagate asexually.

In herbaceous dicotyledonous stems the vascular conducting tissue (xylem and phloem) is organized into discrete strands or vascular bundles, each containing both xylem and phloem. The cells between the vascular bundles are thin-walled and often store starch. The peripheral region of cells in the stem is called the cortex; cells of the central portion make up the pith. The outermost cells of the stem compose the epidermis. No bark is formed on the herbaceous stem. In contrast, woody dicot stems develop an outer layer of dead, thick-walled cells called cork cells, which together with the underlying phloem compose the bark of the tree. The major portion of the woody stem's diameter is a cylinder of xylem (wood) that originates from a region of cell division called the vascular cambium. The water-conducting cells that make up the xylem are nonliving. The accumulated xylem forms annual rings composed of two zones: a relatively wide zone of spring wood (made up of large cells, characteristic of rapid growth) and a narrower zone of summer wood (smaller cells). Xylem rays, radiating like spokes of a wagon wheel, are formed in the xylem and connect with the peripheral phloem. Stems of monocotyledons are composed of numerous vascular bundles that are arranged in a seemingly scattered manner within the ground tissue. Monocot vascular bundles lack a vascular cambium, and thus monocot stems do not become woody in a manner comparable to dicots.

Leaves and roots

Leaves are the other plant organ that, along with stems, constitutes the shoot of the vascular plant body. Their principal function is to act as the primary site of photosynthesis in the plant. Leaves of dicots possess a network of interconnecting veins and veinlets between the larger veins of the leaf (a pattern called net venation). Leaves of monocots possess major veins that extend parallel to the long axis of the leaf (parallel venation). Leaves are classified based on leaf arrangement and on whether they are simple or compound. A leaf may be deeply lobed but still simple; a compound leaf is composed of two or more distinctly separate leaflets.

Structurally, leaves are composed of an outermost layer of cells called the epidermis. Epidermal cells secrete a waxy substance (cutin) that forms a cuticle impermeable to water. The pores (stomates) in the epidermis that allow for gas exchange are formed between specialized epidermal cells called guard cells. Vascular bundles (veins) are embedded in the mesophyll, the tissue that includes all of the cells between the upper and lower epidermis. The cells of the mesophyll contain the photosynthetic pigments.

The root system begins its development from the embryonic root (radicle), which grows out of the seed after the seed has absorbed water. This is the primary root of a new plant. The tip of the root is covered by a mass of loose cells called the root cap. Just beneath the root cap is the region of cell division of the root. Epidermal outgrowths just above the root tip are root hairs that are active in water and mineral absorption. Two types of root system are commonly distinguished, fibrous roots and taproots. Fibrous root systems are composed of large numbers of roots nearly equal in size; root systems of this type are found, for example, in the grasses. A taproot system is one in which the primary root remains the largest, and a number of smaller secondary roots are formed from it; taproots are found in such plants as carrots and dandelions. Roots that arise other than by branching from the primary roots are called adventitious roots. The prop roots of corn, for example, are adventitious.

Flowers

As noted above, a primary distinction between the gymnosperms and angiosperms is that the latter have flowers. Flowers represent modified shoots that have become differentiated for reproduction. The flower bears whorls of floral organs attached to a receptacle, the expanded end of a flower stalk on which the flower parts are borne. Sepals (collectively called the calyx) are modified leaves that encase the developing flower. They are sterile floral parts and may be either green or leaflike or composed of petallike tissue. Petals (collectively called the corolla) are also sterile floral parts that usually function as visually conspicuous elements serving to attract specific pollinators to the flower. The calyx and corolla together are referred to as the perianth. Flowers that lack one or both of the above perianth parts are called incomplete. Stamens (collectively called the androecium) are the male parts of the flower. Stamens are composed of saclike anthers (microsporangia) and filaments, which are stalks that support the anthers. Anthers are usually compartmentalized and contain the pollen grains (microgametophytes). The pistil, or female part of the flower, is composed of one or a number of carpels (collectively called the gynoecium) that fuse to form an essentially enclosed chamber. The three regions of the pistil (from the base up) are the ovary, which contains the ovules; the style, a stalked structure atop the ovary that elevates the stigma; and the stigma, a sticky knob whose surface receives the pollen during pollination.

Flowers may contain both male and female parts (a condition called perfect) or parts related to just one sex (imperfect), or they may have no sexual parts (sterile). Female and male flowers may be located on separate plants (dioecious) or on the same plant (monoecious). Flowers can also be borne singly or in aggregations called inflorescences.

Primitive flowers are radially symmetrical (actinomorphic) and characterized by numerous spirally arranged floral parts. Floral parts are free (unfused) and borne on an elongated floral axis. Sepals, petals, and stamens are attached below the ovary. Advanced flowers are bilaterally symmetrical and are characterized by a reduction in the number of floral parts. Floral parts are fused (often forming a long floral tube). Sepals, petals, and stamens are attached to the floral tube above the ovary.

Pollination is the transfer of pollen to the stigma of the same or another flower. Agents of pollination encompass a vast and diverse array of animals, including insects, birds, bats, and slugs. Flowers exhibit various adaptations to pollinators, such as showy corollas, the production of nectar (a sugary liquid), and even visual cues visible only to insects that can perceive ultraviolet wavelengths of light. Flowers pollinated by wind generally are small and lack petals. The stigma is the pollen receptor site and must be

chemically compatible with any pollen that lands on it for the pollen grain to germinate. This ensures that only genetically compatible sperm are transferred to the egg.

In flowering plants, ovules are enclosed and protected in an ovary. As the ovule develops into a seed, the ovary matures into a fruit. The formation of fruits is a characteristic feature of the flowering plants. Fruits are extremely variable. In some fruits the ovary wall (pericarp) is thick and fleshy; in others it is thin and dry.

Fruits and seeds

Angiosperms have evolved many different adaptations for seed dispersal involving such agents as wind, water, and animals. Adaptations to wind dispersal include wings or plumules attached to the seed or as part of the fruit, or simply very minute seeds that are easily windborne. Adaptations to water dispersal are seeds that float or fruits that float and carry the seeds with them. Some seeds are a source of food to animals, which bury the seeds in the ground, where they later germinate. Other plants produce a fleshy fruit that is eaten along with the seeds inside it by animals, which pass the seeds through their digestive tracts unharmed. Another adaptation for animal dispersal is the development of barbed fruits or seeds that stick to the coats or skins of wandering animals. Some plants, such as witch hazels or jewelweed, can project their seeds through the air some distance from the parent plant.

Seeds have many adaptations that enable them to survive long periods of harsh conditions. Seeds can remain viable in a dormant condition for a few days or, in some species, for hundreds of years. (W.C.D.)

REPRODUCTION AND LIFE HISTORIES

Each organism from inception to death goes through a sequence of genetically programmed developmental events constituting a life history. In eukaryotic organisms, development involves cellular events such as mitosis, meiosis, and syngamy (fertilization), which variously proceed by nuclear division (karyokinesis), cytoplasmic division (cytokinesis), cytoplasmic fusion without the union of nuclei (plasmogamy), or nuclear fusion (karyogamy).

This discussion focuses on the life histories of land plants—that is, nonvascular (bryophytes) and vascular plants, the latter comprising nonseed vascular plants (pteridophytes) and seed plants (gymnosperms and angiosperms). Although algae and fungi were traditionally regarded as plants, fungi are now universally considered as constituting the kingdom Fungi, whereas algae are often included in the kingdom Protista.

Life histories. The chromosome number in cells may be haploid, with one set of chromosomes per cell (written $1n$); diploid, with two sets ($2n$); polyploid, with three or more sets; or dikaryotic, with a pair of nuclei in a cell ($n + n$), a condition that occurs mainly in fungi. Three types of sexual life histories have been recognized for the eukaryotic organisms: $1n$, or haplontic; $2n$, or diplontic; and $1n$-$2n$ ($2n$-$1n$). The former two types have collectively been called haplobiontic or monobiontic, because the life histories include only one phase; the third type has been called haplodiplontic, diplohaplontic, diplobiontic, dibiontic, or sporic, because the life history involves two alternating multicellular phases, or generations. Algae and fungi have many variants of all three types, especially the first, while land plants have the third type exclusively. In addition, all land plants are strictly oogamous, having motile sperm and nonmotile eggs. (In contrast, the algae and fungi may be oogamous or, frequently, isogamous or anisogamous, the latter conditions characterized by morphologically similar gametes that are either of the same size or with the female gametes of a larger size, respectively.)

The $1n$-$2n$ life history of bryophytes and vascular plants comprises the entire sequence of developmental events from zygote formation via syngamy (fertilization) to spore formation via meiosis. Syngamy and meiosis are successive events in a sexual life history. Syngamy involves the union of two $1n$ gametes to form a $2n$ zygote, which eventually develops into a $2n$ sporophyte. Meiosis involves the division of a $2n$ sporocyte (meiocyte, spore mother cell, pollen mother cell) to produce four $1n$ spores. These four spores constitute a tetrad. Gametes are $1n$ cells that fuse to form a zygote, whereas spores are $1n$ cells that develop into gameophytes without uniting with another cell.

Syngamy and meiosis

Syngamy and meiosis are generally inseparable and alternate; consequently, the sporophyte develops from a zygote formed via syngamy, whereas the haploid gametophyte results after the sporophyte undergoes meiosis to form spores. These two phases, or generations, are multicellular in land plants. This type of life history involves an alternation of generations, a phenomenon not occurring in haplobiontic ($1n$ or $2n$) life histories. In the latter, such alternation is evident in both the morphological and the nuclear (chromosomal) changes that occur.

Meiosis and syngamy (fertilization) are the critical events that separate the sporophytic and gametophytic generations. (All nonmeiotic cell divisions involved in development are mitotic, in which chromosomes replicate, giving each daughter cell a full complement.) The zygote is the first stage and the sporocyte the last stage of the sporophytic generation, whereas the spore is the first stage and the gametes (eggs, sperm) the last stage of the gametophytic generation. The sporocytes of the multicellular $2n$ sporophyte divide meiotically to form $1n$ spores (sporogenesis). Each meiotic division results in a tetrad of four spores. (The land plants produce only sexual spores resulting from meiosis.) Each spore divides mitotically to form a multicellular $1n$ gametophyte, which eventually produces gametes mitotically (gametogenesis). The gametes (an egg and a sperm) fuse in the process of syngamy to form a $2n$ zygote. The zygote divides mitotically to form a multicellular embryo (embryogenesis), which is protected by either gametophytic tissues (such as remnants of archegonia in the nonseed land plants) or sporophytic tissues (the seed in the seed plants). An embryo, which is actually an immature sporophyte, is universally found among the land plants and often becomes dormant, but eventually grows into the mature sporophyte. During organogenesis in vascular plants the embryo develops into the mature sporophyte, with its vegetative organs (root and shoot, the latter consisting of stem and leaves) and reproductive structures (cones or strobili, flowers, etc.).

In all land plants, the two alternating generations are morphologically dissimilar, and the gametophyte is initially dominant over the sporophyte. In all bryophytes the gametophyte remains dominant, and the sporophyte is physiologically dependent on it. A homosporous life history, in which only one morphological type of spore is produced, is found in most bryophytes, although a few (*e.g., Macromitrium*) exhibit an anisosporous life history, in which the same sporangium produces morphologically similar spores of two different sizes.

Morphological differences among generations

In vascular plants, the sporophyte ultimately becomes physiologically independent of the gametophyte. A homosporous life history occurs in *Psilotum* (Psilotophyta), *Lycopodium* (Lycophyta), *Equisetum* (Sphenophyta), and most ferns (Filicophyta). A functionally heterosporous life history, in which the same sporangium produces morphologically similar but physiologically different spores, has been reported in a few pteridophytes—*e.g., Equisetum*. Finally, a heterosporous life history, in which different sporangia produce morphologically different types of spores, occurs in *Selaginella, Isoetes* (Lycophyta), a few aquatic ferns (Filicophyta), and all seed plants.

Homosporous life histories. A homosporous life history occurs in nearly all bryophytes and in most pteridophytes. It is characterized by morphologically identical spores that germinate to produce bisexual (both male and female) gametophytes in pteridophytes, but either bisexual or more usually unisexual (either male or female) gametophytes in bryophytes. Each mature gametophyte bears gametangia (sex organs) that produce gametes. Each antheridium (male gametangium) forms many motile, flagellate sperm, and each archegonium (female gametangium) forms one nonmotile egg. Fusion of an egg and sperm (syngamy) creates a zygote and restores the $2n$ ploidy level. Various mechanisms prevent the fusion of eggs and sperm from a bisexual gametophyte (inbreeding). For example, the sex organs may mature at different times (usually antheridia mature first), or inbreeding may be chemically or genetically inhibited. The zygote divides mitotically to form the embryo, which then develops into the sporophyte.

Gametangia and gametes

This eventually produces sporangia, which bear meiocytes (sporocytes) that divide meiotically to form spores. The number of spores produced per sporangium ranges from 16 or 32 in some pteridophytes to more than 65,000,000 in some mosses. The sporangia may be borne in specialized structures, such as sori in ferns or as cones (strobili) in many other pteridophytes. The leaflike structures that bear sporangia are called sporophylls.

In most homosporous life histories of pteridophytes the spores are both morphologically and physiologically identical and produce bisexual gametophytes. In some species of horsetail (*Equisetum*) the spores may be physiologically different and produce male or female gametophytes. This uncommon situation is called functional heterospory and may represent the means by which the heterosporous condition (see below) in vascular plants evolved from the homosporous condition.

Anisosporous life histories. In anisosporous life histories, an unusual phenomenon in bryophytes, there is a size difference among spores produced in the same sporangium. Each meiotic division results in a tetrad of two small spores that produce male gametophytes and two larger spores that produce female gametophytes.

Heterosporous life histories. A heterosporous life history occurs in some pteridophytes and in all seed plants. It is characterized by morphologically dissimilar spores produced from two types of sporangia: microspores, or male spores, and megaspores (macrospores), or female spores. In pteridophytes, megaspores are typically larger than microspores, but the opposite is true in most seed plants.

The spores produce two types of gametophytes: each microspore develops into a microgametophyte (male gametophyte), which ultimately produces male gametes (sperm), and each megaspore produces a megagametophyte (female gametophyte), which ultimately produces female gametes (eggs). Fusion of an egg and a sperm creates a zygote and restores the $2n$ ploidy level. The zygote divides mitotically to form the embryo, which then develops into the sporophyte. Eventually the sporophyte produces sporangia, which bear sporocytes (meiocytes) that undergo meiosis to form spores. Microsporangia (male sporangia) produce microsporocytes (micromeiocytes) that yield microspores. Megasporangia (female sporangia) produce megasporocytes (megameiocytes) that yield megaspores. The sporangia may be borne in specialized structures such as sori in ferns, cones (strobili) in some pteridophytes and most gymnosperms, or flowers in angiosperms. The leaflike structures bearing microsporangia and megasporangia are called, respectively, microsporophylls and megasporophylls. In angiosperms these sporophylls represent, respectively, the stamens and carpels of the flower; in gymnosperms these sporophylls may constitute parts of, respectively, microstrobili (male cones, or pollen cones) and megastrobili (female cones, ovule cones, or seed cones).

The essential difference between the homosporous and heterosporous life history is the presence in the latter of two spore types (microspores and megaspores) and their concomitant precursory structures (microsporocytes and megasporocytes; microsporangia and megasporangia; etc.) and subsequent structures (microgametophytes and megagametophytes).

Variations involving seed plants. The gymnosperms and angiosperms not only lack some reproductive structures found in the homosporous and heterosporous pteridophytes but also have certain reproductive structures peculiar to the seed plants. Heterosporous pteridophytes, like their homosporous counterparts, have archegonia, antheridia, and motile, flagellate sperm. The seed plants completely lack antheridia, and of the extant groups only the ginkgo and the cycads have flagellate sperm. Archegonia occur in most gymnosperms except *Gnetum* and *Welwitschia,* but they are lacking in all angiosperms.

Structures unique to seed plants

Pollen grains and pollen tubes (male reproductive structures), ovules and seeds (female reproductive structures), and seedlings are structures unique to all seed plants. The ovule is a single megasporangium (in seed plants this is called the nucellus) surrounded by one or two integuments (in rare cases, none or three) and containing inside the nucellus a single megasporocyte (spore mother cell). The megasporocyte undergoes meiosis to form four megaspores, three of which typically degenerate, the remaining one developing into the megagametophyte (female gametophyte). Ovules never dehisce to release their megaspores, unlike the megasporangia of most pteridophytes. The pollen grain is the partly or completely developed microgametophyte (male gametophyte). It is usually multicellular, consisting of two or three cells in angiosperms and usually two to five cells in gymnosperms, although in conifers it is occasionally one cell (for example, the families Taxaceae and some Cupressaceae) or 6 to 43 cells (the families Araucariaceae and some Podocarpaceae).

During pollination, pollen is transferred from its source to a receptive surface: in gymnosperms from the microsporangium to the integument or, especially, the pollination droplet of the ovule (rarely to the cone scale); in angiosperms from the microsporangium (pollen sac) of the anther to the stigma of the carpel. Once pollen has reached the appropriate receptive source, it germinates to form the pollen tube, a structure that grows toward the megagametophyte and in so doing conveys the sperm directly to the egg. All angiosperms and most gymnosperms, except the ginkgo, cycads, and some fossil seed plants, lack swimming sperm. The presence of swimming sperm apparently represents a more primitive, transitional evolutionary condition. After fertilization, the ovule transforms into a seed. The integument or integuments become modified into the seed coat. The seed typically becomes dormant for a period of time before it germinates to produce a seedling.

Double fertilization

Double fertilization is a phenomenon unique to angiosperms. Each pollen grain produces two sperm; one fuses with an egg to form the zygote, and the other fuses with one or more polar nuclei in the female gametophyte (megagametophyte, or also "embryo sac") to form an endosperm, which has a ploidy level that varies from $2n$ to $15n$. In approximately 70 percent of the known cases, the second sperm fuses with two endosperm nuclei to produce a $3n$ (triploid) endosperm. The endosperm is a special nutritive tissue for the embryo and, after seed germination, for the seedling. In contrast, the megagametophyte is the comparable nutritive tissue in the gymnosperms.

Asexual reproduction. Both homosporous and heterosporous life histories may exhibit various types of asexual reproduction (vegetative reproduction, somatic reproduction). Asexual reproduction is any reproductive process that does not involve meiosis or the union of nuclei, sex cells, or sex organs. Depending on the type of life history, asexual reproduction can involve the $1n$ or $2n$ generation.

Significance of reproductive method

The significance of sexual reproduction is that it is responsible for the genetic variation arising in a population as a result of the segregation and recombination of genetic material via meiosis and syngamy, respectively (the cells that result from sexual reproduction are genetically different from their parent cells). The significance of asexual reproduction is that it is a means for a rapid and significant increase in the numbers of individuals. (Weeds, for instance, are successful partly due to their great capacity for vegetative reproduction.) The cells that result from asexual reproduction are genetically identical to their parent cells. In addition, vegetative reproduction in the bryophytes and pteridophytes is a means of bypassing the somewhat lengthy and moisture-dependent sexual process—that is, the motile, swimming sperm characteristic of these groups require the presence of water, which may be a limiting factor in drier times.

Deviations from the usual life history. In most life histories, a $2n$ sporophyte typically alternates with a $1n$ gametophyte, but there are significant deviations. Apospory is the development of $2n$ gametophytes, without meiosis and spores, from vegetative, or nonreproductive, cells of the sporophyte. In contrast, apogamy is the development of $1n$ sporophytes without gametes and syngamy from vegetative cells of the gametophyte. The $2n$ aposporous gametophytes and $1n$ apogamous sporophytes are usually infertile under natural conditions because of disruption of cytological events. Various compensating genetic mechanisms, however, may occur to complete the life history. Parthenogenesis is the formation of a $1n$ embryo directly

from an unfertilized egg. Apospory and apogamy occur in bryophytes, pteridophytes, and angiosperms, whereas parthenogenesis occurs in ferns and angiosperms. Apogamy is more common in pteridophytes, but apospory is more common in bryophytes.

Some ferns (certain species of *Trichomanes* and *Vittaria*) have lost the ability to produce sporophytes. The species exist as gametophytes that spread by gemmae (units of asexual reproduction); although gametangia are produced, no sporophytes result. (Ru.Sc.)

PLANT PHYSIOLOGY

General features of plant nutrition. Plant nutrition includes the nutrients necessary for the growth, maintenance, and reproduction of individual plants; the mechanisms by which plants acquire such nutrients; and the structural, physiological, and biochemical roles these nutrients play in metabolism.

Mode of nutrition. All organisms obtain their nutrients from the environment, but not all organisms require the same nutrients, nor do they assimilate these nutrients in the same way. There are two basic nutritional types, autotrophs and heterotrophs. Heterotrophs require both inorganic and organic (carbon-containing) compounds as nutrient sources. Autotrophs obtain their nutrients from inorganic compounds, and their sole source of carbon is carbon dioxide (CO_2). An autotroph is photoautotrophic if light energy is required to assimilate CO_2 into the organic constituents of the cell. Furthermore, a photoautotroph that also uses water as a nutrient source and liberates oxygen in the energy-trapping process of photosynthesis is an oxygenic photoautotroph. The Earth's first such organisms are believed to have been the major sources of the present-day oxygen content of the atmosphere (approximately 21 percent). All plants, as well as many monerans and protists, are characteristically oxygenic photoautotrophs.

Autotrophs and heterotrophs

Plants, as autotrophic organisms, use light energy to photosynthesize sugars from CO_2 and water. They also synthesize amino acids and vitamins from carbon fixed in photosynthesis and from inorganic elements garnered from the environment. (Animals, as heterotrophic organisms, cannot synthesize many nutrients, including certain amino acids and vitamins, and so must take them from the environment.)

Essential elements and minerals. Certain key elements are required, or essential, for the complex processes of metabolism to take place in plants. Plant physiologists generally consider an element to be essential if the plant is unable to complete its life cycle (*i.e.,* grow and reproduce) in its absence; the particular structural, physiological, or biochemical roles of the element cannot be satisfied by any other element; and the element is directly involved in the plant's metabolism (*e.g.,* as part of an enzyme or other essential organic cellular constituent). Beneficial elements are those that stimulate plant growth by ameliorating the toxic effects of other elements or by substituting for an element in a less essential role (*e.g.,* as a nonspecific osmotic solute). Some elements are beneficial in that they are necessary for the growth of some, but not all, plant species.

The required concentrations of each essential and beneficial element vary over a wide range. The essential elements required in relatively large quantities for adequate growth are called macroelements. Nine minerals make up this group: carbon (C), hydrogen (H), oxygen (O), nitrogen (N), potassium (K), calcium (Ca), magnesium (Mg), phosphorus (P), and sulfur (S). Seven other essential mineral elements are required in smaller amounts (0.01 percent or less) and are called microelements. These are iron (Fe), chlorine (Cl), manganese (Mn), boron (B), copper (Cu), molybdenum (Mo), and zinc (Zn). The specific required percentages may vary considerably with species, genotype (or variety), age of the plant, and environmental conditions of growth.

Macro- and microelements

A macronutrient is the actual chemical form or compound in which the macroelement enters the root system of a plant. The macronutrient source of the macroelement nitrogen, for example, is the nitrate ion (NO_3^-). Carbon dioxide from the atmosphere provides the carbon atoms and two-thirds of the oxygen required by plants. Water taken from the soil provides about one-third of the oxygen and much of the hydrogen. Soil provides macroelements and microelements from mineral complexes, parent rock, and decaying organisms. Factors that determine plant root uptake include the solubility of the chemical in question, the adsorptive properties of the charged soil surfaces, and the metabolic uptake capabilities of the individual plant.

The macroelements carbon, hydrogen, oxygen, and nitrogen constitute more than 96 percent of the dry weight of plants. Thus, they are the major constituents of the structural and metabolic compounds of the plant. Their presence within cells also helps regulate osmotic pressure. Microelements are generally either activators or components of enzymes, although the macroelement potassium also serves these roles.

General overview of metabolic cycles. Metabolism denotes the sum of the chemical reactions in the cell that provide the energy and synthesized materials required for growth, reproduction, and maintenance of structure and function. In plants the ultimate source of all organic chemicals and the energy stored in their chemical bonds is the conversion of CO_2 into organic compounds (CO_2 fixation) by either photosynthesis or chemosynthesis. The general and specific features of plant metabolism ultimately derive from oxygenic photosynthesis, which underlies the autotrophic nutrition of plants.

Pathways and cycles. Chemical reactions in the cell occur in a sequence of stages called a metabolic pathway. Each stage is catalyzed by an enzyme, a protein that changes (usually increases) the rate at which the reaction proceeds but does not alter the reactants or end products. Certain thermodynamic conditions must be met for a reaction to proceed, even in the presence of enzymes. If the end product of the reaction is also the reactant (or substrate) that starts the pathway, then the sequence of reactions is called a metabolic cycle. The intermediate chemicals that are formed and used in the various stages of the sequence are called intermediary metabolites.

Metabolic pathways and cycles are either catabolic (energy-releasing) or anabolic (energy-consuming). Catabolic reactions break down complex metabolites into simpler ones, while anabolic reactions build up (biosynthesize) new molecules. When chemical bonds are broken, energy is released, which drives anabolic reactions to form new bonds. The energy released generally has been stored in high-energy bonds of an intermediate energy carrier molecule, such as the terminal phosphate bond of adenosine triphosphate (ATP). (When the terminal phosphate is split from the ATP molecule, adenosine diphosphate, or ADP, is formed and inorganic phosphate is released, along with energy.) The simpler metabolites formed via catabolic reactions are often the building-block metabolites used in anabolic reactions to synthesize more complex molecules (*e.g.,* starch, proteins, or lipids).

Catabolic and anabolic reactions

Control mechanisms. The cells of all plants are eukaryotic, since they possess a nucleus and membrane-bound organelles, such as chloroplasts, mitochondria, glyoxyosomes, peroxisomes, and vacuoles. The thousands of metabolic reactions that take place in the cell are regulated within these organelles and their subcompartments. When compared with cells of other eukaryotic organisms, plant cells may have the highest degree of metabolic compartmentalization.

The primary mechanism of metabolic control, however, remains the enzymes themselves. Although all enzymes of the pathway help determine the net and directional flow of carbon, certain key stages are controlled by regulatory enzymes. Regulatory enzymes may catalyze either the first stage in the metabolic pathway, or they may catalyze reactions in which key branch points occur. The activity of such enzymes, in turn, may be controlled by the amount synthesized (coarse control)—in that further action by the enzyme is inhibited when some critical concentration of the reaction product is reached—or by special metabolites, called effectors, that interact directly with the enzyme (fine control). The latter metabolites may be either part of, or totally unrelated to, the metabolites of the pathway.

Another mechanism by which metabolic reactions are regulated is through transport systems in the membranes

of organelles. These systems control the nature, direction, and amount of metabolites entering the metabolic pathways and are often uniquely related to the autotrophic nutrition of plants.

Principal pathways and cycles. The 6-carbon sugar glucose, a product of photosynthesis, may be translocated in the form of sucrose (a 12-carbon sugar) to nourish nonphotosynthesizing parts of the plant, or it may be polymerized into starch for storage. (Trehalose, another 12-carbon sugar, replaces sucrose in some nonvascular plants.) When required, sucrose and starch are hydrolyzed to glucose and then enter glycolysis or the pentose phosphate pathway. The reactions of both pathways take place in the cytoplasm of the cell.

Result of glycolysis

The net result of glycolysis is the metabolism of glucose into two molecules of the three-carbon organic acid pyruvate. This 10-reaction metabolic pathway involves phosphate-containing intermediates and is regulated by two enzymes, which catalyze those reactions that contain the substrates fructose phosphate and phosphoenolpyruvate (PEP). Glycolysis yields ATP molecules and hydrogen; the latter is accepted by the coenzyme (coenzymes are smaller, nonprotein participants associated with certain enzymes) nicotinamide adenine dinucleotide (NAD) to form NADH. The hydrogen on NADH then reacts either with molecular oxygen (O_2) to catalyze the release of energy (trapped in the high-energy bonds of ATP) or with another metabolite to reduce the molecule by the addition of hydrogen. Some intermediates are used in the biosyntheses of fat or certain amino acids.

The pentose phosphate pathway, an alternative pathway for the catabolism of glucose, operates in the presence of oxygen. It furnishes reducing power (*i.e.,* it accepts hydrogen atoms and carries them on the coenzyme nicotinamide adenine dinucleotide phosphate [NADP]) for use in the synthesis of fat and five-carbon (pentose) sugars that are required as components of nucleic acids, some vitamins, and key metabolites. It is regulated by the rate at which the product of the pentose pathway, NADPH, is oxidized. The end products are cycled back into the glycolytic pathway.

Pyruvate is transported into the mitochondria, where it enters a sequence of 10 reactions called the tricarboxylic acid (TCA) cycle, or Krebs cycle. Pyruvate is metabolized into the two-carbon intermediate, acetyl coenzyme A (CoA), which combines with a four-carbon acid, oxaloacetate. The product, citrate, has three acid (or carboxylic) groups; hence the name tricarboxylic acid cycle. Citrate is systematically catabolized (broken down) with progressive losses of successive carbon atoms as CO_2 into five-carbon and, finally, four-carbon, acids. The latter acid, oxaloacetate, begins the cycle again. With each oxidation reaction, a hydrogen atom is transferred to the coenzyme NAD or, in one reaction, the coenzyme flavin adenine dinucleotide (FAD) to form NADH and FADH, respectively. The reduced coenzymes NADH and FADH enter into a sequence of reactions called the respiratory chain on the inner membrane of the mitochondrion. This chain is a series of carriers (ubiquinone and several iron-containing chemicals called cytochromes) that ultimately transfer the hydrogen and electrons of these coenzymes to molecular oxygen, forming water. The energy generated from the oxidation by the respiratory chain is trapped in the three ATP molecules formed per NADH molecule oxidized. The mechanism is chemiosmotic in that it involves building a hydrogen ion (proton) gradient on one side of the mitochondrial membrane.

A net of 36 ATP molecules are gained from all hydrogen-carrying coenzymes formed in glycolysis and the TCA cycle, and they represent the principal energy source for most anabolic (biosynthetic) reactions in plants. In addition, the TCA cycle furnishes metabolites for the biosynthesis of important organic molecules of the cell.

Isoprenoid pathway

Another metabolic cycle, the isoprenoid pathway, produces essential oils, carotenoid pigments, certain plant hormones, and rubber. These metabolites (often called secondary metabolites) are unique to plants and serve such functions as attracting pollinating insects, photosynthesis, and growth and development. Plant seedlings use the glyoxylic acid cycle to convert fats (principally from seeds) into glucose. This occurs initially in the glyoxysome and subsequently in the mitochondria and cytoplasm.

Unique features of plant metabolism. The pathways outlined above exist in essentially the same form in all organisms, but metabolism in plants does have certain unique features. Plant mitochondria, for example, have specific transport systems for the NADH produced in glycolysis and for the oxaloacetate produced from a direct fixation of CO_2 into PEP. Unlike animal mitochondria, plant mitochondria metabolize malate and the amino acid glycine. A special enzyme converts malate to pyruvate, thereby allowing an alternative to the glycolytic pathway. Glycine is a product of the unique plant pathway of photorespiration II (see below *Photosynthesis*).

Plant mitochondria possess a cyanide-resistant cytochrome-respiratory chain in addition to the cyanide-sensitive respiratory chain also found in other organisms. Oxidation of NADH through this alternative pathway produces energy in the form of heat, but no ATP. Some physiologists suggest that this pathway is a mechanism for burning off excess carbohydrate. Others believe that this pathway allows a TCA cycle, in times of increased need, to produce more than the usual amount of metabolites for biosyntheses, which, in the presence of ATP acting as a feedback inhibitor, could not normally be produced. This system functions at a high rate in the flowers of the arum (Araceae) lily. Temperatures of this organ may reach 40° C (104° F) and may contribute to the attraction of pollinators. The overall function of the cyanide-insensitive respiratory pathway, however, is not clearly understood.

Photosynthesis. As discussed above, the autotrophic mode of nutrition of plants is derived from oxygenic photosynthesis. Energy-rich organic compounds are synthesized from low-energy atmospheric CO_2 using the energy of absorbed sunlight. (Some bacteria are nonoxygenic photosynthesizers, utilizing hydrogen sulfide, H_2S, rather than water.) The resultant organic compounds initiate the flow of energy and carbon through the food chains of agricultural and natural ecosystems, intrinsically linking plants with the heterotrophic life-forms of the remaining four kingdoms of organisms. The oxygen liberated by plants (and certain photosynthetic protists and monerans) over geologic time, has oxygenated the Earth's atmosphere and has produced fossil fuels such as coal, gas, and oil.

This section describes the basic mechanisms of photosynthesis—the acquisition of energy and the fixation of carbon dioxide—used by plants of diverse evolutionary lines.

Absorption ranges for plants

Basic mechanisms. Electromagnetic radiation having wavelengths between approximately 430 and 700 nanometres can be seen as light by the eye and constitutes the range absorbed by plants for photosynthesis. Blue light has a wavelength around 450 nanometres, and red light, a wavelength of 650–700 nanometres.

Double-membraned cell organelles called chloroplasts contain the photosynthetic apparatus: light-absorbing pigments, other electron-carrying chemicals (cytochromes and quinones), and enzymes. (Pigments absorb light of a particular wavelength; those wavelengths that are not absorbed are reflected and may be perceived as colour; hence, for example, the green colour of many plants.) The inner membrane of the chloroplast is folded into flat tubes, the edges of which are joined to hollow, sacklike disks called thylakoids. Stacks of thylakoids embedded with pigment molecules are called grana. The inner matrix of the chloroplast is called the stroma.

Photosynthesis consists of two interdependent series of reactions, the photochemical light reactions and the metabolic dark reactions; the former are dependent on light, the latter on temperature. Light reactions occur in the grana and dark reactions in the stroma. The overall formula for photosynthesis is:

$$6CO_2 + 12H_2O \rightarrow \underset{\text{(glucose)}}{C_6H_{12}O_6} + 6O_2 + 6H_2O.$$

The light reactions, the first stage of photosynthesis, convert light energy into chemical energy (ATP and NADPH). Light reactions comprise two interdependent systems, called photosystems I and II. The dark reactions,

the second stage of photosynthesis, use the chemical energy products of the light reactions to convert carbon from carbon dioxide to simple sugars.

Light reactions consist of several hundred light-absorbing pigment molecules so arranged as to maximize the gathering of light energy. These "antennae" are coupled to a minicircuit of electron-carrying chemicals. The pigments are chlorophyll *a* and chlorophyll *b* and various carotenoids. Absorbed light energy is transferred to specialized chlorophyll molecules called P_{700} and P_{680} in photosystems I and II, respectively. Once these specialized chlorophyll molecules acquire sufficient energy, electrons are given up to the electron carriers within their photosystems, initiating an electron flow. (The carrier molecules include quinones and cytochromes.) The effect of this, when photosystems I and II function synchronously, is the formation of a chemiosmotic gradient of protons that phosphorylates (adds a phosphate group to) ADP, resulting in ATP. These electrons also effect the formation of NADPH from NADP. The P_{680} chlorophyll, upon loss of its electron, becomes a strong oxidizing agent that subsequently causes the water molecule to dissociate into protons and oxygen gas.

The dark reactions are responsible for the conversion of carbon dioxide to glucose. The essential reaction involves the combining of CO_2 with the five-carbon sugar ribulose 1,5-bisphosphate (RuBP) in a series of reactions called the Calvin-Benson cycle. This reaction yields an unstable intermediate, which breaks down into two molecules of phosphoglycerate (PGA), a three-carbon acid. Each reaction is catalyzed by a specific enzyme. Six revolutions of the cycle means that six CO_2 molecules react with six RuBP molecules to produce 12 molecules of PGA; two three-carbon PGA molecules combine to form the six-carbon glucose, and 10 PGAs are recycled to regenerate six molecules of RuBP. The ATP and NADPH from light reactions provide the energy and reducing power to form glucose and refurbish the CO_2 acceptor, RUBP.

Calvin-Benson cycle

Specific variations in photosynthesis. Chlorophylls *a* and *b* (bound to a protein) and carotenes constitute the principal light-absorbing complex of most plants. Differences in chloroplast structure, though not major, occur among phylogenetically diverse plant groups. All such variations, however, represent evolutionary adaptations to utilize more efficiently the light energy that drives the reactions common to all oxygenic photosynthesizers, *i.e.*, photosystems I and II.

The enzyme ribulose 1,5-bisphosphate carboxylase/oxygenase (rubisco) catalyzes the formation of organic molecules from CO_2. As the major enzyme of all photosynthetic cells, rubisco is probably the most abundant protein on the Earth. There is, however, a major catalytic flaw in the ability of this enzyme to convert CO_2 to sugars. In the presence of molecular oxygen, rubisco catalyzes reactions in which oxygen is introduced (*i.e.*, it acts as an oxygenase), and CO_2 is formed rather than converted.

Rubisco evolved in photosynthetic organisms that lived in the atmosphere of primitive Earth, an atmosphere which contained only traces of molecular oxygen. As photosynthesis in the Cyanobacteria of Precambrian times (3.8 billion to 570 million years ago) oxygenated the atmosphere, the ratio of carbon dioxide to oxygen fell drastically, and rubisco began to function more and more as an oxygenase. This greatly reduces the net fixation of CO_2 into sugars and, therefore, photosynthetic efficiency. Rubisco as an oxygenase only splits RuBP into PGA and the two-carbon acid phosphoglycolate, which initiates the photorespiratory carbon-oxidation cycle, or photorespiration. (This cycle probably evolved to recycle PGA back into the photosynthetic pathway, thereby preventing an even greater loss of carbon.) Photorespiration involves three organelles (chloroplasts, peroxisomes, and mitochondria), each with unique transport mechanisms for the cycle's intermediates.

C_3, C_4, and C_3-C_4 plants

All plants are classified as C_3 (plants that use only the Calvin-Benson cycle), C_4 (plants that use an additional CO_2-fixation mechanism and the Calvin-Benson cycle), C_3-C_4 (plants intermediate between C_3 and C_4), and CAM (plants that have a variant of the C_4 pathway).

The majority of plants fix CO_2 directly into RuBP, and their first stable product is the three-carbon acid PGA, hence the designation C_3. These plants have an active photorespiratory cycle, especially in high light intensity and warm temperatures.

Sometime during the later periods of the Cenozoic Era (66.4 million years ago to the present), certain of the angiosperms (grasses and the dicotyledonous plants) of mainly tropical climates evolved a CO_2-fixation system that precedes the Calvin-Benson cycle. The first fixation is into the three-carbon acid phosphoenolpyruvate (PEP) by PEP carboxylase (an enzyme that has no oxygenase function) in the outer mesophyll cells of the leaf. The first stable fixation product is the four-carbon acid oxaloacetate, hence the designation C_4 plants. Oxaloacetate is converted to malate, which is transferred to a thick-walled bundle sheath cell that shields the subsequent reactions from the high concentration of molecular oxygen in the atmosphere. Malate is decarboxylated, and rubisco of the primitive and susceptible Calvin-Benson cycle functions more efficiently because here photorespiration is suppressed. There is thus a spatial separation of initial CO_2 fixation and the Calvin-Benson cycle. This efficiency is not without cost, however, as an additional ATP is required to recycle PEP. For this reason, C_3 plants may be more efficient in cold climates with less light (and, therefore, less photorespiration).

There are also plants with enzymatic and leaf anatomical characteristics intermediate between C_3 and C_4 plants, called C_3-C_4 intermediate species. These plants are thought to be in the pathway of evolution to full C_4 photosynthetic status.

Succulent plants of the desert regions (*e.g.*, cacti) also initially fix CO_2 into oxaloacetate. This occurs only at night when conditions are cooler, however. Normally, the stomates in leaves or stems, through which plants lose water and acquire carbon dioxide, are open in the day and closed at night; those of the succulent plants do the opposite through a special mechanism that prevents great loss of water during the hot days. The resultant oxaloacetate is converted into malate, stored in the vacuole, and released during the day when the stomates are closed. Malate is decarboxylated, and the CO_2 that released is fixed by rubisco in the usual Calvin-Benson cycle. Both the C_4 and C_3 processes take place in the same cell. This process is called crassulacean acid metabolism (hence CAM plants) after a family of succulent plants (Crassulaceae). (J.H.Y.)

ECOLOGY

Plant geography. Plants occur over the surface of the Earth in well-defined patterns that are closely correlated with both climate and the history of the planet. Forests are the most important of these natural communities from the standpoint of area, carbon content, annual carbon fixation, the cycling of nutrient elements, and influence on energy and water budgets, as well as being the principal reservoir of biotic diversity on land. The most extensive forests are the boreal coniferous forests of North America, Scandinavia, northern Europe, and northern Asia. The moist forests of the tropics are the most diverse, often containing as many as 100 species of trees per hectare (1 hectare = 2.47 acres) and occasionally many more.

Importance of forests

Forests are commonly distinguished from woodlands and savannas in having a closed canopy of trees that may be 10 metres to more than 50 metres in height. At their wetter and cooler limits, forests are replaced by tundra, a community of low-growing trees and shrubs with a ground cover of *Sphagnum* and other mosses, dwarfed shrubs, and perennial herbs. The transition from forest to tundra occurs at high latitudes and elevations and may extend over a considerable distance, throughout which trees are scattered, restricted to sheltered areas, and stunted in growth. Such taiga stands are common at the northern limit of trees throughout the Northern Hemisphere. The alpine transition tends to be more abrupt. Both are often determined locally by the distribution of snow.

Forests yield at their warmer and drier margins to grassland (called prairie in North America and steppe in Asia). In North America the rich grasslands of the eastern midcontinent, which is affected by moisture moving north-

ward from the Gulf of Mexico, once supported a diversity of herbaceous plants, including grasses that might reach several feet in height, and were known as tall-grass prairie. Westward, where water was less abundant, the grasses diminished in diversity and size to form the short-grass plains. Southward and westward the plains yielded under increasing aridity to desert, as old as any other zone in an evolutionary sense and very diverse in species. Similar patterns of vegetation occur across the broad waist of Asia.

The distinction between grassland and forest is blurred in many places, especially in the semiarid subtropics and in other regions where fires may be common. This transitional zone is savanna, an open, grassed woodland. The southeastern coastal plain of North America was originally such a region, a pine savanna that produced plants adapted to, in fact dependent on, burning for survival. The longleaf pine (*Pinus palustris*), for instance, has a "grass" stage, which lasts for several years of early growth, with the bud protected at the very surface of the ground by a thick tuft of long, grasslike leaves that shield it from the heat of a fire. Once sufficient energy has been stored in the taproot and short stem, the tree virtually explodes into growth and passes rapidly through a period of vulnerability to fire. As a tree of 3 metres (10 feet) or more in height, it is safe from all but the hottest of fires. Oaks and other species are replacing pines as savannas disappear from the southeastern United States because of the continuing division of land ownership and protection from fire.

In contrast, the extensive savannas of eastern Africa have been maintained through climate, fires, and grazing and support an extraordinary diversity of migratory animals that have evolved to use different parts of the ecosystem. The patterns of use are complex and phased in time, and they work to maintain both the diversity and productivity of the savanna. These African populations are the last of the large and diverse mammalian and avian populations that once also grazed similar areas of North America, Europe, and Asia.

The forests of the moist tropics remain the most complex and fascinating of terrestrial ecosystems. They extend from mountain slopes to river swamps and occurred throughout the tropics wherever precipitation allowed trees to survive. Human activities have virtually eliminated forests from vast areas of the tropics and threaten to destroy the residual forests globally. The largest areas of tropical forest occur now in the Amazon basin of South America and in the Congo basin of west-central Africa. The extensive tropical forests of Southeast Asia have been reduced to fragments of the area they occupied as recently as 1950.

The forests of the Amazon basin have evolved as a part of a river system whose water level fluctuates annually by as much as 15 metres (50 feet) or more along the middle and lower Amazon. There are substantial further differences in the quality of water. The Negro River, for example, drains an area of sands low in nutrient elements, where organic matter has accumulated sufficiently in soils to produce the humic acids that give the river its dark colour and sufficient acidity to affect the plant and animal life it can support. The areas flooded in the annual cycle are forested and are known by the Portuguese word *várzea.* Trees in this zone survive flooding for several weeks annually and provide the basis of a food web that includes fish adapted for grazing on tree fruits and seeds. The grazing fish possess large flat molars adapted for masticating seeds and other coarse organic matter, and they compete for seeds dropped from pods in *várzea* trees on the river.

The distribution of the various zones of vegetation over the Earth has changed with climates throughout time. The greatest changes in the recent past have been due to the periods of glacial advance and retreat over the last several hundred thousand years. As recently as 15,000 years ago glacial ice covered much of eastern North America as far south as the present Hudson estuary and covered as well all of Scandinavia and northern Europe. Sea level was tens to hundreds of feet lower at various times, and extensive coastal areas now flooded were exposed. Such drastic changes in climate have sorted and resorted plant species, allowed the establishment of a diversity of forest types over time, and worked against the establishment or preservation of the mutual dependencies among species so common in the tropics. These higher-latitude forests, while similar in form and genera and sometimes in species around the world, are apparently highly mutable, the antithesis of the stable and self-perpetuating "organism" they were once thought to be. Nonetheless, the concept of the plant community as an organism remains viable and useful in ecology and is the basis of much of the most progressive analysis of how to manage the Earth for successful support of large numbers of people.

Glacial advances

Succession and zonation. It is known from studies of plant residues and pollen preserved in the highly acid sediments of bogs and from observations of contemporary glaciers that the vegetation southward from the glacial front in the Northern Hemisphere was banded in much the same way the vegetation is zoned today: tundra occurred in a zone closest to the ice; coniferous forests occurred in a warmer and drier zone southward; and deciduous forests occurred still farther southward. As the habitat changed—that is, as the glacial ice melted and the glacial front retreated—the vegetation migrated onto the new landscape: first the pioneer species of the tundra—a few hardy low-growing or crustose lichens and mosses of small stature—followed by dwarfed willows and birch and, ultimately, the full panoply of tundra plants. As the climate ameliorated further, the forest followed, always according to pattern, with a few pioneers followed by the full array of species characteristic of the forest.

Primary succession

The process of invasion of a new landscape not previously occupied by plants has long been called "primary succession." In this case the succession was in response to both the availability of a new habitat and a climatic warming that allowed the replacement of tundra by forest and of the coniferous forest ultimately by deciduous forest as the warming continued. Further warming might result in the replacement of deciduous forest by grassland, as has occurred worldwide at the steppe- or prairie-forest border in response to climatic changes over recent centuries. This boundary is strongly influenced by the fires that have swept grasslands throughout time. Disturbances such as clearings for agriculture, fires, diseases, and storms severe enough to open gaps in a forest may start secondary successional changes that also reestablish the normal vegetation for that climatic zone.

Throughout much of the tropics, where forests have been destroyed over large areas to make pasture, conditions appropriate for what might be a normal succession within disturbance-caused gaps in an otherwise intact forest do not exist, and the sites are permanently impoverished. The process generates grassland, shrublands, or sometimes bare earth. One-third of the land area of India has been impoverished in this way and is lost to agriculture, forests, or forestry. This class of land is the most rapidly growing class globally, according to interpretations of statistics of the United Nations Food and Agriculture Organization in 1985.

Ecosystems and the biosphere. The thin mantle of the Earth that supports life extends not more than a few feet into the sediments of the abyssal depths of the oceans and to a few tens of thousands of feet into the atmosphere, where pollen grains or other spores may survive. It is this thin layer at the surface of the Earth that is the biosphere, the place where life occurs. Human activity has changed the Earth irreversibly to form the present biosphere, and humans are now a force for further irreversible change.

Dispersal and colonization. The methods by which plants are distributed over the Earth are as diverse and complex as the plants themselves. The most widely occurring plants are the small-bodied, rapidly reproducing forms whose spores can be carried by wind and water and by birds and other animals.

Among the seed plants, whose spores are less mobile, explanations for the current distribution of plants become more complicated and include such profound changes over evolutionary time as the breakup of Pangaea some 300 million years ago, the opening of the Atlantic Ocean, and the isolation of North and South America, Australia, and Madagascar from larger continental landmasses. Progressive isolation produced endemism, evolutionary di-

vergence sufficient to generate whole floras peculiar to a particular region, with many species, even genera, not known elsewhere. Volcanic islands are much younger than the continents and support floras derived from chance invaders carried by wind, sea, or animals, including humans. Island floras also come to exhibit endemism. Charles Darwin's observations of the fauna and flora of the Galápagos Islands off the western coast of South America led him to recognize the general process of biotic evolution. The islands are diverse in form and climate, isolated in varying degrees from each other and from the continent, and support a highly diversified flora and fauna clearly derived from South America, but modified over the time of isolation to contain forms peculiar to the islands.

Methods of dispersal

Diverse mechanisms exist for dispersing spores and seeds at appropriate times and in appropriate habitats to ensure the survival of the new plant. In forests adapted to fire, pinecones may remain closed until a fire occurs. The heat of the fire opens the cones and releases the seeds into a fire-seared habitat, where the seedlings find soil with an abundance of nutrients left by the burning of organic matter and reduced competition from other plants. Seeds of trees frequently have "wings" that permit wider distribution by wind. Other plants, such as thistles and burdock, have awns or other appendages that catch on animals and obtain a wide distribution in areas frequented by them. Still others have elaborate mechanisms for spreading via rhizomes, stems, or other propagules.

Especially during the past two centuries, human activities have both deliberately and inadvertently spread the higher plants and microbial plagues of higher plants around the Earth. Examples among the higher plants are numerous; while many of the transfers of such plants have been benign and some clearly advantageous, others have been disastrous. Tree species have been moved freely around the world, sometimes with remarkable effectiveness. The Monterey pine (*Pinus radiata*), for instance, is a diminutive and unproductive tree on the coast of California, but it has become a major timber tree as a result of rapid growth in plantations in New Zealand. The rubber tree (*Hevea*) was carried in the 19th century from the moist tropics of Brazil to Java and elsewhere in the South Pacific, where its excellent growth, free of the diseases and competitors that affect it in its native American habitat, nearly destroyed the market for Brazilian rubber. Other inadvertent introductions have been far from benign; some have severely impoverished the landscape. *Bromus tectorum,* for example, is an annual grass introduced from Europe and the Transcaspian steppes to the arid intermontane west in North America, probably as a contaminant in fodder in the latter part of the 19th century. It spread rapidly and became a continuous ground cover over extensive areas. The plant completes its life cycle early in the season, sets seeds, and stands through the rest of the year as a continuous cover of dry grass not useful for fodder. It does, however, carry fire, which changes and can even destroy the native vegetation.

Microbial diseases of plants have been inadvertently spread around the world by human activities. Such diseases continue to exact a high cost in that they progressively impoverish the vegetation. One of the most profound changes in any forested region was produced by the introduction of the chestnut blight (*Endothia parasitica*) to North America. The fungus, introduced from Asia, found a home in the Fagaceae of eastern North America but was lethal to the American chestnut (*Castanea dentata*), the dominant species of extensive stands in the southern Appalachians and elsewhere. Once a common and popular shade tree and a source of smooth-grained, easily worked wood and of the abundant mast that was one of the principal foods of indigenous wild turkeys and other animals, the American chestnut never developed resistance to this now widely distributed fungus, which has several hosts in the Fagaceae and which produces wind-borne spores.

Dutch elm disease

The American elm (*Ulmus americana*), a magnificent tree of the moist forests of lowlands throughout eastern North America, has suffered a similar but less comprehensive loss through the ravages of another exotic fungus, Dutch elm disease (*Ceratocystis ulmi*), which was imported from Europe on infected wood probably early in the 1900s. The fungus is spread by two elm bark beetles. Individual trees sometimes escape the fungus for many years but ultimately succumb before reaching the large sizes that were common throughout its range early in the 20th century.

Human effects on plants and natural communities. Humans have influenced the structure and development of natural communities for many thousands of years. The major influence has been through fire, which has been used deliberately in herding game, in rejuvenating plants used for fodder, in opening forested plots for agriculture, in maintaining savannas and grasslands, and in keeping forests open for easy passage for hunting. With the rapid expansion of human populations globally in recent decades, however, the changes in vegetation have been profound. Forests have advanced into grasslands and savannas. Succession has replaced pine forests with deciduous forests as in the southeastern coastal plain and piedmont regions of North America. In many instances, temporary protection from fire has resulted in the accumulation of a large mass of flammable vegetation that, when burning, makes a hot fire which kills the overstory trees and many species that normally survived the more frequent and less severe fires of the past. In such circumstances the vegetation may be impoverished, a condition that may prevail indefinitely, once established.

Domestication. The origins of domesticated plants and agriculture are buried in a dim and unrecorded past of 10,000 years and more. Recent experience with Stone Age cultures of the Amazon basin and elsewhere has shown that such cultures have a sophisticated knowledge of plants for such purposes as food, medicine, and the tools and poisons used for hunting and fishing. Amerindian cultures improved a wide range of species by deliberate selection of the more productive and useful forms. Manioc (*Manihot utilissima*) remains a staple of large sections of Latin America, especially Brazil and the Amazon basin. A woody, tuberous plant whose origin in the savannas of South America has long been lost, it is propagated vegetatively by planting a piece of tuber or a segment of stem. The tuber is ground to make a flour that must be washed to remove toxic quantities of hydrocyanic acid before being eaten directly or baked into a flat bread called *cazabi* (cassava).

Corn

Corn, or maize (*Zea mays*), was domesticated 10,000 to 8,000 years ago either from teosinte (a perennial *Zea* that exists today) or from a lost ancestor that existed in the highlands of what is now central Mexico. Its culture had spread as far north as southern Maine by the time of European settlement of North America. Corn is now the third largest plant-based food source in the world. Other plants domesticated from the region include the common bean, squash, chili, tomato, avocado, papaya, guava, sapodilla, cotton, sisal, and vanilla.

The process of domestication appears to have involved selection and cultivation of the most promising and productive of plants long harvested in the wild. Continued selection brought drastic shifts in the population, shifts in the frequency of genes, and the development of new races totally dependent on humans to maintain them. Races developed in this way have been called cultigens to emphasize their dependence on cultivation. Such is the case with corn: the original stock has probably been lost, although teosinte may offer clues as to the source.

Major crop plants have been domesticated over the last several thousand years from sources identified by the Russian botanist N.I. Vavilov. Common wheat (*Triticum vulgare*) and rye (*Secale cereale*) probably were first domesticated from the grasses of Central Asia. Various millets (*Panicum*) and barley (*Hordeum hexastichum*) originated in the mountainous regions of central and western China, rice (*Oryza sativa*) probably in the Indian region.

Changes in biosystems: pollution. There is no plant community anywhere on Earth that has not felt the direct influence of the expansion of the human enterprise. The influence has had the form of direct intrusion through hunting and gathering, cultivation, or extensive harvest of trees or other plants or has had indirect effects through the

Chemical pollutants

harvest of fish or other animals or through changes in the chemistry of the environment. The latter changes are now pervasive and range from the introduction of pesticides, such as DDT, and other industrial toxins, such as the PCBs, to the acidification of rain, with substantial modification of the sulfur, nitrogen, and hydrogen ion budgets of ecosystems over large areas. Widespread destruction of forest tree species occurred during the 1980s in eastern North America and throughout Europe because of the combined effects of acid rain and other air pollutants.

Symptoms of incipient damage from air pollution and other chronic disturbances are now common in forests and other types of vegetation around the world, although symptoms may be subtle and difficult to assign as being caused by a particular pollutant. They range from such changes as the elimination of populations of leafy lichens from the bark of trees to the mortality of the trees themselves. Under extreme circumstances, such as in regions around smelters in Sudbury, Ont., Can., virtually all higher plants have been eliminated from an area of hundreds of square miles.

One of the most spectacular and informative examples of the patterns of impoverishment of forests exposed to chronic disturbance was induced experimentally in an oak-pine forest at Brookhaven National Laboratory in central Long Island, N.Y., U.S., using ionizing radiation. A single radiation source was used in the centre of the forest. The exposure, begun in the fall of 1961, was sufficient within months to eliminate all plants from a central area within a few metres of the source and to produce a systematic zonation of the forest at lower exposures. The zones ranged through a lichen and moss community, followed at lower exposures by a continuous ground cover of a single species of sedge (*Carex pensylvanica*). At slightly lower exposures a community of low shrubs occurred, then high shrubs at still lower exposures. An impoverished forest of hardy oaks well within the normal range of variability of the oak-pine forest of the region occurred down the gradient of exposure and beyond that zone; at exposures of one roentgen or less daily, the pines survived and the forest appeared intact and normal. Effects could, however, be detected by careful measurements of growth down to exposures on the order of 0.1 roentgen per day. The experiment offered a specific example of the progressive impoverishment of vegetation where both cause and effect could be measured, something not possible in most instances of pollution. Such experience is useful not only in defining the effects of exposure to ionizing radiation but also in identifying stages in the impoverishment of natural communities as effects accumulate.

Similar changes occur in aquatic systems in response to chronic pollution. The early stages of pollution in bodies of water usually involve enrichment with nutrient elements, especially nitrogen and phosphorus, a process known as eutrophication.

General effects of pollution

The effects of pollution on land and in water are to favour small-bodied, rapidly reproducing organisms that do not depend on complex food webs. The process of simplification and impoverishment is now global and affects terrestrial and aquatic communities alike. It is the continuously expanding result of chronic intrusions on natural systems by human influences. The impoverishment threatens all life because it reduces systematically the capacity of the Earth to support plants.

Conservation. One of the most serious threats to life is the possibility of a rapid warming of the Earth from the accumulation of heat-trapping gases in the atmosphere. The most important of the gases is carbon dioxide, but methane, nitrous oxide, and the chlorofluorocarbon compounds (CFCs) also contribute significantly. The largest sources of carbon dioxide are the combustion of fossil fuels, the destruction of forests, and, probably, the warming itself. Methane is a product of the use of fossil fuels and of the decay of organic matter in soils and swamps and bogs. Its rate of evolution is accelerated by warming.

The effects of the warming are expected to be profound: If the warming occurs as rapidly as some suggest, there could be a change in the average temperature of the Earth of as much as 0.5°–1° C (0.9°–1.8° F) per decade during the early part of the 21st century. A change of 1° in temperature is equivalent to 60–100 miles of latitudinal change in the middle latitudes. Such high rates of change in climatic zones would not only disrupt agriculture but also destroy forests and other natural communities at their warmer and drier margins more rapidly than they can be regenerated elsewhere. One of the effects would be to speed the decay of organic matter in soils and accelerate the accumulation of carbon dioxide and methane in the atmosphere, an effect that would further accelerate the warming. Avoiding such change will require abandoning fossil fuels as the major source of energy for industrial societies. An alternative is solar energy.

Classical approaches to conservation involve specific efforts at preserving species that are endangered. By far the most desirable approach to the preservation of the Earth's 5–30 million or more species is through management of natural ecosystems. Studies of tropical forests have emphasized that the area required to assure the habitat of bird populations as well as populations of other animals and plants is very large indeed, on the order of 10,000 hectares. There is evidence to support earlier experience that the best way to assure the preservation of the indigenous biota is to maintain a matrix of the indigenous vegetation with more than 50 percent of the land area undisturbed and an additional 25 to 50 percent in successional stages. This type of land-use planning exists in very few places around the world but offers the advantage of a landscape that maintains itself with stable, near-maximum flows of energy through living systems and offers, regionally, maximum sustainable support for people at low cost. The difficulty is that the apparently high stocks of land under light use—the forests, the water, and energy in various forms—are an open invitation to immediate, short-term exploitation, and their existence may suggest that the landscape can be more intensively occupied. (G.M.W.)

EVOLUTION AND PALEOBOTANY

The fossil record

The evolutionary history of plants is recorded in fossils preserved in lowland or marine sediments. Some fossils preserve the external form of plant parts, others show cellular features, and still others consist of microfossils such as pollen and spores. In rare instances, fossils may even display the ultrastructural or chemical features of the plants they represent. The fossil record reveals a pattern of accelerating rates of evolution coupled with increasing diversity and complexity of biological communities that began with the invasion of land and continued with the progressive colonization of the continents. At present, fossil evidence of land plants dates to the Ordovician Period (505 to 438 million years ago) of the Paleozoic Era.

By far the most diverse and conspicuous living members of the plant kingdom are vascular plants (tracheophytes), in which the sporophyte phase of the life history (see above *Reproduction and life histories*) is dominant. Fossil remains of vascular plants provide evidence for evolutionary changes in the structure of the plant body (sporophyte and gametophyte), in the variety of plant forms, in the complexity of the life history, in the tolerance for ecological conditions, and in systematic diversity. Nonvascular plants, or bryophytes (mosses, liverworts, and hornworts), are much smaller and less diverse than vascular plants. The first evidence for liverworts is from rocks of the Devonian Period (408 to 360 million years ago), while the earliest moss fossils are from the Permian Period (286 to 245 million years ago). In contrast to tracheophytes, most fossil bryophytes are relatively similar to living forms. Understanding of the evolution of nonvascular plants is, therefore, less complete than for tracheophytes.

Evolution from algae

Botanists now believe that plants evolved from the algae; the development of the plant kingdom may have resulted from evolutionary changes that occurred when photosynthetic multicellular organisms invaded the continents. The earliest evidence for land plants consists of isolated spores, tracheid-like tubes, and sheets of cells found in Ordovician rocks. The abundance and diversity of these fossils increases into the Silurian Period (438 to 408 million years ago), where the first macroscopic (megafossil) evidence for land plants has been found. These megafossils consist of

slender forking axes that are only a few centimetres long. Some of the axes terminate in sporangia that bear trilete spores (*i.e.,* spores that divide meiotically to form a tetrad). Since a trilete mark indicates that the spores are the product of meiosis, the fertile axes may be interpreted as the sporophyte phase of the life cycle.

Fossils of this type could represent either vascular plants or bryophytes. Another possibility is that they are neither but include ancestors of vascular plants, bryophytes, or both. The earliest fossils also include at least one or more additional plant groups that became extinct early in the colonization of the land and therefore have no living descendants. By the Early Devonian Period (408 to 360 million years ago), some of the fossils that consist of forking axes with terminal sporangia also produced a central strand of tracheids, the specialized water-conducting cells of the xylem. Tracheids are a diagnostic feature of vascular plants and are the basis for the division name, Tracheophyta.

The simplest and presumably most primitive vascular plants from the Late Silurian and Early Devonian periods (421 to 387 million years ago) were the Rhyniopsida. They included plants such as *Cooksonia* and *Rhynia,* which were herbaceous colonizers of moist habitats. Most were less than 30 centimetres (12 inches) tall. The plant body was not differentiated into stems, leaves, and roots; rather, the forking, above-ground axes bore terminal sporangia and produced stomata which demonstrate that the plants were green and photosynthetic. Surface or underground axes served to root the plant and were anchored by rhizoids. Because such plants produced only one type of spore, they were nonseed plants with a homosporous life cycle and free-living gametophytes. A small number of such gametophytes have been described from Devonian deposits.

Some plants of the Early Devonian had multicellular emergences of tissue along their above-ground axes, which are thought to have increased the light-capturing surface of the photosynthetic tissue. Such emergences (enations) gave rise to the leaves (microphylls) of the Lycopsida, thus producing an above-ground shoot system that consisted of branching stems with leaves. Underground axes that lacked leaves would have become the roots. Lycophytes were the first plants with well-differentiated shoot systems, and they are the most ancient groups of vascular plants with living representatives.

The leaves of several other plant groups were derived from modifications of the forking axes. There was a variety in structure among Devonian plants. The axes of some plants forked equally, while other plants—*e.g., Drepanophycus*—were more specialized, comprising a large, centrally located axis as well as smaller axes borne laterally. In plants where the lateral systems branched in only one plane, the side branches were flat like leaves. Filling in (webbing) of the spaces between forks of the laterals with photosynthetic tissue produced leaves called megaphylls. There is evidence for the evolution of megaphylls in several plant groups of the Late Devonian Period (374 to 360 million years ago) and Early Carboniferous Period (360 to 320 million years ago). Although most of these groups have no living representatives, by the Carboniferous Period they had given rise to homosporous ferns, sphenopsids (horsetails), and seed plants (gymnosperms).

Development of megaphylls

As Devonian plants with microphylls and those with specialized branching systems diversified, many grew to the size of shrubs. By the Middle Devonian Period (387 to 374 million years ago) there were shrub-size representatives of several lineages, but a further increase in the size of herbaceous plants was restricted by the limited diameter that above-ground stems and rooting systems could achieve. The development of lateral (secondary) growth overcame this size restriction. The ability to produce secondary growth evolved independently in several groups. In the lycophytes, much of this secondary growth occurred in cortical tissues; in the ancestors of seed plants and several other lineages, however, the production of wood accounted for most of the growth in stem diameter. About the same time, downward-growing, central rooting systems evolved independently in lycophytes and other plant groups. As a result there were forests with a canopy of giant lycophyte and gymnosperm trees near the beginning of the Early Carboniferous Period.

As plants developed more complex growth forms, they also underwent systematic diversification and evolved more specialized modes of sexual reproduction. The most primitive vascular plants had a homosporous life cycle, in which reproduction and dispersal involved a single type of spore. Extant homosporous plants include most ferns and many lycophytes. The homosporous life cycle is an effective means for long-distance dispersal of species. Although it permits the fertilization of an egg by a sperm from the same gametophyte plant, genetic recombination, considered important for more rapid evolution, is not possible. Moreover, because the gametophytes of homosporous plants are exposed to the environment for an extended period of time, mortality is relatively high.

By the Middle Devonian Period (387 to 374 million years ago), the heterosporous life cycle had evolved independently in several groups, including lycophytes and the ancestors of seed plants. In heterosporous plants there are two sizes of spores; the smaller (a microspore) produces a male gametophyte, and the larger (a megaspore) produces a female gametophyte. The incidence of genetic recombination is increased by this obligate cross-fertilization. Both types of gametophytes develop quickly within the protective spore wall. Compared with homosporous plants, reproduction is more rapid and mortality is reduced in heterosporous plants. By the end of the Devonian, heterosporous plants became the dominant species in most wetland environments; however, the need for an abundant source of water from the environment to effect fertilization prevented the heterosporous plants from establishing communities in drier habitats.

A series of changes in the reproductive biology among some heterosporous plants during the Late Devonian overcame this environmental restriction and allowed them to colonize a much wider range of habitats. These changes also led to the evolution of seed plants. In seed plants, the megaspore is retained in the megasporangium and the microspore is brought to a pollen chamber at the tip of this organ. The megasporangium is then sealed, and gametophyte development and fertilization occur within a protected environment. When such a megasporangium is enclosed in a seed coat, the structure with its enclosed embryo is called a seed. Seeds of the earliest such plants were exposed to the external environment; these "naked seed" plants are referred to as gymnosperms.

Evolution of seed plants

From the Late Devonian through the base of the Late Cretaceous Period (97.5 to 66.4 million years ago), gymnosperms underwent dramatic evolutionary radiations and became the dominant group of vascular plants in most habitats. Extant gymnosperms include conifers, cycads, and *Ginkgo biloba,* but these represent only a small fraction of the gymnosperms that inhabited the Earth during the Mesozoic Era (245 to 66.4 million years ago). Among the Mesozoic forms were species with a wide variety of mechanisms for effecting pollination, protecting the seeds, dispersing the seeds, and increasing the natural selection of the most successful varieties.

Primitive forms of the flowering plants (angiosperms) arose from among this diverse array of complex gymnosperms. From their earliest diversification in the Cretaceous Period (144 to 66.4 million years ago), angiosperms rapidly came to dominate the land flora. Today there are approximately 250,000 to 300,000 species of flowering plants, which account for more than 90 percent of the diversity of vascular plants. Among the many structures that contribute to this success are flowers, fruits, complex vein patterns in the leaves, and highly specialized cells of the conducting system.

Just as an ecological succession of plant forms can transform bare ground into a complex plant community, so an evolutionary succession can be thought of as having transformed bare continents into rich terrestrial biotas. In ecological succession, herbaceous weeds colonize bare earth and modify the environment for the development of more complete ground cover. This further modifies the environment for the successive establishment of larger herbs, perennial shrubs, fast-growing trees, and, finally, slower-

Development of plant communities

growing trees, vines, and epiphytes (plants that grow on other plants rather than in soil).

As discussed above, primitive land plants of the Late Silurian and Early Devonian periods were primarily small herbs that inhabited the moist lowlands near oceans, lakes, and streams. The first of these grew on bare ground because there would have been little or no organic soil. They may be thought of as the evolutionary equivalent of primary colonizing weeds, and their establishment on land was responsible for the first of three dramatic increases in the diversity of land plants.

After the initial production of organic soils by these evolutionary colonizers, communities of larger herbs and shrubs were able to develop, and they became common in the Middle Devonian. In the Late Devonian these communities were themselves succeeded by communities dominated by heterosporous, tree-sized plants. During the Early Carboniferous Period, nonseed plants continued to dominate many wetland habitats, while communities dominated by gymnosperm trees colonized drier habitats than had been previously available to the heterosporous forms.

Communities dominated by trees were the first to provide suitable habitats for vines and for epiphytes. Such communities also provided habitats for herbivorous animals. As a result, biological communities of considerable complexity evolved by the Late Carboniferous Period (320 to 286 million years ago), giving rise to a second dramatic increase in the diversity of land plants.

With the rapid diversification of the gymnosperms and an increasing complexity of interactions between plants and animals, there was a significant evolutionary turnover among land plants during the Permian, Triassic, and Jurassic periods (286 to 144 million years ago). New groups of gymnosperms replaced some of the primitive gymnosperms as well as the most prominent nonseed plants from earlier periods. Among the most successful gymnosperms of these periods were conifers, cycads, *Ginkgo*, and several major groups with no extant representatives.

Dominance of angiosperms

Flowering plants probably also originated during this time but did not become a significant part of the fossil flora until the middle of the Cretaceous Period. The fossil evidence provides a clear picture of the rapid diversification and spectacular rise to dominance of the angiosperms during the Late Cretaceous Period. This was accompanied by a dramatic increase in the complexity of plant community structure, produced at least in part by an expanding array of interactions between plants and animals. Improved pollination and seed dispersal were among the benefits of such interactions to plants. That animals equally benefited is evidenced by the coevolution of several groups of animals (particularly insects) and angiosperms during the Late Cretaceous and Tertiary periods. Such interactions contributed significantly to a third rapid increase in global plant diversity and helped angiosperms achieve the overwhelming dominance of the land flora characteristic of modern vegetation. (G.W.R.)

CLASSIFICATION

Annotated classification.

KINGDOM PLANTAE

Multicellular, photosynthetic, eukaryotic organisms containing chlorophylls *a* and *b* and carotenoids and forming true starch; chlorophyll contained in stacks of membranes (grana) within chloroplasts; cellulosic cell walls; aerial tissues covered with a waxy cuticle and provided with openings (stomata) in the epidermis bordered by two guard cells and serving in gas exchange; alternation of generations with multicellular gametangia and sporangia and possessing a multicellular embryo stage; motile sperm, when present, with whiplash flagellum; sperm with a microtubular cytoskeleton; cell division associated with the dispersal of the nuclear membrane and the formation of a cell plate across the mitotic spindle (phragmoplast).

Division Bryophyta

Small, mostly nonvascular, archegoniate plants with a dominant, photosynthetic, free-living gametophyte; sporophyte has little or no chlorophyll and is dependent on gametophyte; biflagellate sperm.

Class Musci (mosses)

Gametophytes "leafy" and radially symmetrical, with leaves arising spirally from stemlike axis; leaves rarely notched or lobed, with thickened "midrib"; many chloroplasts per leaf cell; gametophyte with multicellular rhizoids; sporophytes with complex spore-containing capsule with peristome and operculum, columella, and stomata present, but elaters absent; between 10,000 and 14,000 species; representative genera include *Polytrichum, Mnium, Funaria,* and *Sphagnum.*

Class Hepaticae (liverworts)

Gametophytes either "leafy" or dorsiventrally flattened (strap-shaped) and thalloid; leafy forms with leaves in 3 rows, 2 lateral and 1 below; leaves usually notched or lobed, with thickened midrib lacking; gametophytes with unicellular rhizoids; many chloroplasts per cell; sporophytes ephemeral, surrounded by gametophytic tissue, lacking opercula, peristome, stomata, and columella; elaters mixed with spores in capsule; capsule opening into 4 or more valves; between 6,000 and 9,000 species; representative genera include *Porella, Frullania, Marchantia, Conocephalum,* and *Riccia.*

Class Anthocerotae (hornworts)

Gametophyte thalloid, with a single, large chloroplast per cell, mucilage cavities present; sporophytes persistent, erect (hornlike), photosynthetic, arising from upper surface of gametophyte, possessing stomata, columella, basal meristem, and pseudoelaters opening by 2 valves but lacking an operculum; representative genus, *Anthoceros.*

Division Lycophyta (club mosses, spike mosses, quillworts)

Vascular plants; sporophyte differentiated into stem, leaf, and root; leaves spirally arranged on stem, of the microphyll type; sporangia borne on upper (adaxial) surface of leaves (sporophylls), one sporangium per sporophyll; vascular tissue basically forming a central core in stem (protostelic); homosporous or heterosporous; ligulate (having basal, leaflike protuberances) or eligulate; strobili (cones) present or lacking; all living genera with primary growth only; gametophytes subterranean or surface-dwelling; motile sperm; about 5 living genera and between 600 and 1,000 species; representative genera include *Lycopodium, Selaginella,* and *Isoetes.*

Division Psilotophyta (whisk ferns)

Vascular plants; sporophyte lacking roots and often leaves; stems with small enations, dichotomously branched; vascular tissue forming a central core in stem (protostelic); sporangia fused into synangiate structure, apparently terminal on short stem; homosporous; gametophytes subterranean, with motile sperm; representative genus, *Psilotum.*

Division Sphenophyta (horsetails, scouring rushes)

Vascular plants; sporophyte differentiated into stem, leaf, and root; stems ribbed and jointed, monopodial; minute leaves whorled at the nodes; vascular tissue organized into bundles; sole living genus with primary growth only; sporangia borne on specialized stalks (sporangiophores) in strobili; homosporous; gametophytes photosynthetic, surface-dwelling; motile sperm; 1 genus, *Equisetum.*

Division Filicophyta (ferns)

Vascular plants; sporophyte differentiated into stem (rhizome), roots, and leaves (fronds); leaves entire or, more often, divided; arrangement of vascular tissue in stem variable, primary growth only; sporangia usually clustered into sori, often located on the under (abaxial) surface of sporophylls; mostly homosporous; gametophytes (prothallia) either subterranean and nongreen or, more commonly, surface-dwelling and photosynthetic; sperm motile; between 9,000 and 12,000 species; representative genera include *Pteridium, Polypodium, Polystichum, Adiantum,* and *Cyathea.*

Division Cycadophyta (cycads)

Palmlike, gymnospermous plants with typically short, thick, unbranched aerial trunks, sometimes subterranean, and large, divided leaves; leaves usually thick and leathery; cones present, often large and terminal on the stem; pollen (male) and seed (female) cones borne on separate plants; gametophytes reduced, not free-living; sperm motile (flagellated), but brought to vicinity of egg by a pollen tube; 10 genera and about 120 species; representative genera include *Zamia* and *Cycas.*

Division Ginkgophyta (ginkgoes)

Gymnospermous plants; tall, much-branched tree with well-developed cylinder of wood; resin ducts present; xylem with tracheids only; stem differentiated into long shoots and short spur shoots; simple, fan-shaped leaves with open dichotomous venation terminate short shoots; leaves deciduous; sexes on separate trees; distinct cones lacking; gametophytes reduced, not free-living; sperm motile (flagellated); mature seeds with fleshy, foul-smelling outer region; 1 living species, *Ginkgo biloba.*

Division Coniferophyta (conifers)

Gymnospermous plants; mostly trees with abundant xylem composed of tracheids only; resin ducts present; leaves simple, needlelike, scalelike, with a single vein or, less commonly, strap-shaped with multiple veins; reproduction by well-defined cones; seeds exposed on ovuliferous scales; gametophyte gener-

ation reduced, microscopic, not free-living; sperm nonmotile, transported to egg by pollen tube; approximately 50 to 55 genera and between 550 and 575 species; representative genera include *Pinus, Abies, Sequoia, Taxodium, Juniperus, Cupressus,* and *Agathis.*

Division Gnetophyta (gnetophytes)

Diverse and unusual group of 3 gymnospermous genera; growth habits of all 3 are atypical among gymnosperms in possessing vessel elements in the xylem and reproductive structures that are somewhat flowerlike; gametophytes reduced as in other gymnosperms; sperm nonmotile; extant genera *Gnetum, Ephedra,* and *Welwitschia.*

Division Magnoliophyta (flowering plants, angiosperms)

Vascular plants, xylem typically with vessel elements; reproduction by flowers; ovules or young seeds enclosed by female reproductive structure (carpel); gametophyte generation extremely reduced, consisting of only a few cells; archegonia and antheridia lacking; nonmotile sperm transported to egg by pollen tube; pollen transported to specialized receptive surface (stigma) on carpel; double fertilization, one sperm uniting with the egg to form a zygote, another fusing with the polar nuclei to form the primary endosperm nucleus; nutritive tissue of seed triploid endosperm; seeds enclosed by mature ovary that ripens into a fruit; approximately 230,000 species worldwide, more than 300 families.

Class Magnoliopsida (*Dicotyledonae;* dicots)

Angiosperms with 2 seedling leaves (cotyledons); leaves with reticulate (net) venation; stem with vascular bundles arranged in a ring, each bundle with a functional vascular cambium; secondary growth common; flower parts in multiples of 4 or 5, pollen typically tricolpate (3 germinal apertures) or some derived type; about 165,000 species.

Class Liliopsida (*Monocotyledonae;* monocots)

Angiosperms with 1 seedling leaf (cotyledon); leaves with parallel venation; stem with scattered vascular bundles; bundles without a functional vascular cambium; secondary growth typically absent; flower parts in multiples of 3; pollen typically have a single germinal aperture; approximately 50,000 to 55,000 species; representative types include grasses, orchids, palms, and lilies.

Critical appraisal. In the classification above, only the major divisions and classes of living plants are listed, and a number of entirely extinct divisions are omitted. The classification outlined is somewhat conservative but is one that best conforms to available data and has gained wide acceptance.

Biological classifications were initially mechanical or "artificial"; that is to say, they had no basis in evolution. This was followed by a period of "natural system" construction, whereby plants were grouped together on the basis of their overall similarities or differences, using as many characteristics as possible. Contemporary systems of biological classification are phylogenetic, which means that various plants are arranged together because they are thought to be related by descent from a common ancestor. As more evidence is accumulated, it is to be expected that classifications also will change to accommodate this new information.

At the turn of the 19th century, the plant kingdom was frequently divided into two major groups, the cryptogamia (algae, fungi, bryophytes, and ferns) and the phanerogamia (gymnosperms and angiosperms). Subsequently, it was common practice among systematic botanists to group all vascular plants together under a single division, Tracheophyta. In contrast, all of the major individual groups of plants are now most often elevated to divisional rank, a view that interprets the individual major groups to be less closely related to one another than previously believed. Difficult and complex questions still exist in the definition and circumscription of certain groups. Although it is convenient to treat the Bryophyta (mosses, liverworts, and hornworts) as a single division, some authorities prefer to recognize three or five separate divisions for these plants, each representing an independent evolutionary lineage. Furthermore, the phylogenetic relationships, if any, of the bryophytic plants with primitive vascular plants remains unclear. The treatment of the Psilotophyta is controversial. Although here recognized as a separate division of lower vascular plants, some botanists prefer to include the group among the ferns (Filicophyta). The flowering plants have classically been recognized as composed of two distinct classes; however, the origin of each class and the phylogenetic relationships of their component taxa have yet to be satisfactorily resolved. (W.C.D.)

BIBLIOGRAPHY. General surveys providing definitions of the kingdom and characterizations of major groups of plants include PETER H. RAVEN, RAY F. EVERT, and SUSAN E. EICHHORN, *Biology of Plants,* 4th ed. (1986), an excellent introductory text; HAROLD C. BOLD, CONSTANTINE J. ALEXOPOULOS, and THEODORE DELEVORYAS, *Morphology of Plants and Fungi,* 5th ed. (1987), a detailed and comprehensive treatment that includes algae and fungi; ERNEST M. GIFFORD and ADRIANCE S. FOSTER, *Morphology and Evolution of Vascular Plants,* 3rd ed. (1989), an advanced text; and V.H. HEYWOOD *et al.* (eds.), *Flowering Plants of the World* (1978, reissued 1985), a beautifully illustrated guide. HENRY S. CONARD, *How to Know the Mosses and Liverworts,* 2nd ed., rev. by PAUL L. REDFEARN, JR. (1979); and W.B. SCHOFIELD, *Introduction to Bryology* (1985), are well-illustrated guides to the bryophytes. JOHN T. MICKEL, *How to Know the Ferns and Fern Allies* (1979), is an introduction to the ferns and lower vascular plants of the United States.

Plant physiology is the subject of many specialized studies. T. WALLACE, *Trace Elements in Plant Physiology* (1950), is the classic exposition of the pioneering experimentation on and founding principles of the subject. EMANUEL EPSTEIN, *Mineral Nutrition of Plants: Principles and Perspectives* (1972), treats the subject from the perspective of the basic and applied plant sciences; JOHN SKUJINS, "Essential Nutrient Elements and Nutrient Cycles in Soil," pp. 41–48 in vol. 3 of DAVID W. NEWMAN and KENNETH G. WILSON (eds.), *Models in Plant Physiology and Biochemistry,* 3 vol. (1987), offers a concise overview of the subject; A.D.M. GLASS, *Plant Nutrition. An Introduction to Current Concepts* (1989), provides a complete exposition suitable for the informed general reader; and A. LÄUCHLI and R.L. BIELESKI (eds.), *Inorganic Plant Nutrition* (1983), is a definitive advanced treatment of mineral nutrition. Metabolic cycles and photosynthesis in plants are studied in CECIE STARR and RALPH TAGGART, *Biology: The Unity and Diversity of Life,* 5th ed. (1989); DAVID D. DAVIES (ed.), *Metabolism and Respiration* (1980), *Biochemistry and Metabolism* (1987), and *Physiology of Metabolism* (1987); G. RAY NOGGLE and GEORGE J. FRITZ, *Introductory Plant Physiology,* 2nd ed. (1983), which provides an overview of plant metabolism; and FRANK B. SALISBURY and CLEON W. ROSS, *Plant Physiology,* 3rd ed. (1985), which offers a more advanced treatment. *Annual Review of Plant Physiology and Plant Molecular Biology* covers current research in all subjects of plant physiology.

Ecological considerations are discussed in MICHAEL G. BARBOUR, JACK H. BURK, and WANNA D. PITTS, *Terrestrial Plant Ecology,* 2nd ed. (1987); VICTOR V. RENDIG and HOWARD M. TAYLOR, *Principles of Soil-Plant Interrelationships* (1989); FRIEDRICH G. BARTH, *Insects and Flowers: The Biology of a Partnership* (1985; originally published in German, 1982); and OSWALD TIPPO and WILLIAM LOUIS STERN, *Humanistic Botany* (1977), on the impact of plants on human beings. On the cultivation of plants, see N.I. VAVILOV, *The Origin, Variation, Immunity, and Breeding of Cultivated Plants: Selected Writings,* trans. from Russian (1951); and R. WALDEN, *Genetic Transformation in Plants* (1989).

Analyses of evolutionary relationships are provided in BARRY A. THOMAS, *The Evolution of Plants and Flowers* (1981), a spectacularly illustrated introduction; WILLIAM G. CHALONER, "Early Land Plants: The Saga of a Great Conquest," pp. 301–316 in WERNER GREUTER and BRIGITTE ZIMMER (eds.), *Proceedings of the XIV International Botanical Congress* (1988), a summary of what is known about the earliest land plants; WILSON N. STEWART, *Paleobotany and the Evolution of Plants* (1983), on the fossil history and the evolution of plant structures; and ELSE MARIE FRIIS, WILLIAM G. CHALONER, and PETER R. CRANE (eds.), *The Origins of Angiosperms and Their Biological Consequences* (1987), a collection of writings on the geography, climate, ecology, plant-animal interactions, structure, reproductive biology, and interrelationships of flowering plants in their evolution from the Cretaceous to the modern biota.

For classification, see LYNN MARGULIS and KARLENE V. SCHWARTZ, *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth,* 2nd ed. (1988), a brief account of the diversity and important traits of the major groups of living organisms according to the five-kingdom system of classification; ALBERT E. RADFORD *et al.*, *Vascular Plant Systematics* (1974), a comprehensive sourcebook on all areas of higher plant systematics and a complete survey of terminology; and ARTHUR CRONQUIST, *An Integrated System of Classification of Flowering Plants* (1981), a scholarly outline of the modern system of angiosperm classification based on available information.

(W.C.D./Ru.Sc./J.H.Y./G.M.W./G.W.R.)

Plate Tectonics

Plate tectonics is a theory dealing with the dynamics of the Earth's outer shell, the lithosphere. Resting on a broad synthesis of geologic and geophysical data, it dominates current thinking in the Earth sciences. According to the theory, the lithosphere consists of about a dozen large plates and several small ones. These plates move relative to each other and interact at their boundaries, where they diverge, converge, or slip relatively harmlessly past one another. Such interactions are thought to be responsible for most of the seismic and volcanic activity of the Earth, although earthquakes and volcanoes are not wholly absent in plate interiors. While moving about, the plates cause mountains to rise where they push together and continents to fracture and oceans to form where they pull apart. The continents, sitting passively on the backs of plates, drift with them and thereby bring about continual changes in the Earth's geography.

The theory of plate tectonics, formulated during the late 1960s, is now almost universally accepted and has had a major impact on the development of the Earth sciences. Its adoption represents a true scientific revolution, analogous in its consequences to the Rutherford and Bohr atomic models in physics or the discovery of the genetic code in biology. Incorporating the much older idea of continental drift, the theory of plate tectonics has made the study of the Earth more difficult by doing away with the notion of fixed continents, but it has at the same time provided the means of reconstructing the past geography of continents and oceans. While its impact has, to a considerable degree, run its course in marine geology and shows signs of reaching the limits of usefulness in the study of mountain-building processes, its influence on the scientific understanding of the Earth's history, of ancient oceans and climates, and of the evolution of life is only beginning to be felt.

For details on the specific effects of plate tectonics, see EARTHQUAKES, VOLCANISM, and GEOMORPHIC PROCESSES. A detailed treatment of the various land and submarine relief features associated with plate motion is provided in CONTINENTAL LANDFORMS: *Tectonic landforms* and OCEANS.

For coverage of other related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 213, 231, 232, 241, and 243, and the *Index*.

This article is divided into the following sections:

Principles of plate tectonics 883
Historical overview 885
  Precursors
  Alfred Wegener and the concept of continental drift
  Renewed interest in continental drift
Plate tectonics as an explanation for Earth processes 889
  Pangaea and its breakup
  Paleoclimates and ancient oceans
  Plate tectonics and mountain building
  Plate tectonics and life
  Early plate activity
Dissenting opinions and unanswered questions 891
  The dissenters
  Unanswered questions
Bibliography 892

PRINCIPLES OF PLATE TECTONICS

The plate tectonics theory has a long and tortuous history. Yet, the theory itself is elegantly simple.

The surface layer of the Earth, from 50 to 100 kilometres (31 to 62 miles) thick, is assumed to be composed of a set of large and small plates, which together constitute the rigid lithosphere. The lithosphere rests on and slides over an underlying, weaker layer of partially molten rock known as the asthenosphere. The constituent lithospheric plates move across the Earth's surface, driven by forces as yet not fully agreed upon, and interact along their boundaries, diverging, converging, or slipping past each other (Figure 1). While the interiors of the plates are presumed to remain essentially undeformed, their boundaries are the sites of many of the principal processes that shape the terrestrial surface, including earthquakes, volcanism, and orogeny (*i.e.*, the deformation that builds mountain ranges).

The most conspicuous feature of the Earth's surface is its division into continents and ocean basins, a division that owes its existence to differences in thickness and composition between the continental and the oceanic crust. The continents have a crust of granitic composition and hence are somewhat lighter than the basaltic ocean floor. Also, they are 30 to 40 kilometres thick as compared to the oceanic crust, which measures only 6 to 7 kilometres in thickness. Their greater buoyancy causes them to float much higher in the mantle than does the oceanic crust, thus accounting for the difference between the two principal levels of the Earth's surface. The boundary between the continental or oceanic crust and the underlying mantle, the Mohorovičić discontinuity, has been clearly defined by seismic studies.

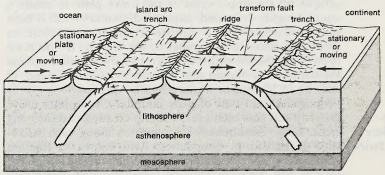


Figure 1: Three-dimensional diagram showing crustal generation and destruction according to the theory of plate tectonics; included are the three kinds of plate boundaries—divergent, convergent (or collision), and strike-slip (or transform).

As conceived by the theory of plate tectonics, the lithospheric plates are much thicker than the oceanic or the continental crust; their boundaries do not usually coincide with those between oceans and continents; and their behaviour is only partly influenced by whether they carry oceans, continents, or both. The Pacific Plate, for example, is purely oceanic, but most of the others contain continents.

Divergent plate boundaries

At a divergent plate boundary, magma wells up from below as the release of pressure produces partial melting of the underlying mantle and generates new crust. Because the partial melt is basaltic in composition, the new crust is oceanic. Consequently, diverging plate boundaries, even if they originate within continents, eventually come to lie in ocean basins of their own making. In fact, most divergent plate boundaries seem to have formed within continents rather than in oceans, probably because a hot, weak layer, sandwiched at a depth of about 15 kilometres between two stronger ones, renders the continental crust more vulnerable to fragmentation than its oceanic counterpart. The creation of the new crust is accompanied by much volcanic activity and by many shallow tension earthquakes as the crust repeatedly rifts, heals, and rifts again.

Convergent plate boundaries

The continuous formation of new crust produces an excess that must be disposed of elsewhere. This is accomplished at convergent plate boundaries where one plate descends—*i.e.*, is subducted—beneath the other. At depths

between 300 and 700 kilometres, the subducted plate melts and is recycled into the mantle. Because the plates form an integrated system that completely covers the surface of the Earth, it is not necessary that new crust formed at any given divergent boundary be completely compensated at the nearest subduction zone, as long as the total amount of crust generated equals that destroyed.

It is in subduction zones that the difference between plates carrying oceanic and continental crust can be most clearly seen. If both plates have oceanic edges, either one may dive beneath the other; but, if one carries a continent, the greater buoyancy prevents this edge from sinking. Thus, it is invariably the oceanic plate that is subducted. Continents are permanently preserved in this manner, while the ocean floor continuously renews itself. If both plates possess a continental edge, neither can be subducted and a complex sequence of events from crumpling to under- and overthrusting raises lofty mountain ranges. Much later, after these ranges have been largely leveled by erosion, their remains continue as a reminder that this is the "suture" where continents were once fused.

The subduction process, which involves the descent into the mantle of a slab of cold rock about 100 kilometres thick, is marked by numerous earthquakes along a plane inclined 30°–60° into the mantle—the Benioff zone. Most earthquakes in this planar dipping zone result from compression, and the seismic activity extends 300–700 kilometres below the surface. At a depth of 100 kilometres or more the subducted oceanic sediments, together with part of the upper basaltic crust, melt to an andesitic magma, which rises to the surface and gives birth to a line of volcanoes a few hundred kilometres behind the subducting boundary. This boundary is usually marked by an oceanic deep, or trench, where the overriding plate scrapes off the upper crust of the lower plate to create a zone of highly deformed, largely sedimentary rock. If both plates are oceanic, the deformed sediments and volcanoes form two island arcs parallel to the trench. If one plate is continental, the sediments are usually accreted against the continental margin and the volcanoes form inland, as they do in Mexico or western South America.

Strike-slip plate boundaries

Along the third type of plate boundary, two plates move laterally and pass each other without creating or destroying crust. Large earthquakes are common along such strike-slip, or transform, boundaries. Also known as fracture zones, these plate boundaries are perhaps best exemplified by the San Andreas fault in California and the North Anatolian fault system in Turkey.

Most of the seismic and volcanic activity on Earth is therefore concentrated along plate boundaries where mid-ocean ridges, trenches with island arcs, and mountain ranges are generated. Some seismic and volcanic activity also occurs within plates. Interesting examples of this interplate activity are linear volcanic chains in ocean basins, such as the Hawaiian Islands and their westward continuation as a string of reefs and submerged seamounts. An active volcano usually exists at one end of an island chain of this type, with progressively older extinct volcanoes occurring along the rest of the chain. Such topographic features have been explained by J. Tuzo Wilson of Canada and W. Jason Morgan of the United States as the product of "hot spots," magma-generating centres of controversial origin located deep in the mantle far below the lithosphere. A volcano builds at the surface of a plate positioned above a hot spot. As the plate moves on, the volcano dies, is eroded, and eventually sinks below the surface of the sea, while a new one forms above the hot spot. Hot spot volcanism is not restricted to the ocean basins; other manifestations occur within continents, as in the case of Yellowstone National Park in western North America.

The movement of a plate across the surface of the Earth can be described as a rotation around a pole, and it may be rigorously described with the theorem of spherical geometry formulated by the Swiss mathematician Leonhard Euler during the 18th century. Similarly, the motions of two plates with respect to each other may be described as rotations around a common pole, provided that the plates retain their shape (Figure 2). The requirement that plates are not internally deformed has become one of the

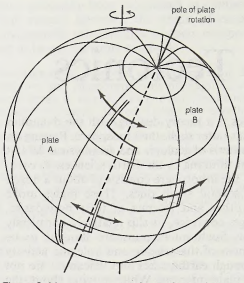


Figure 2: Movement on a sphere of two plates, A and B, can be described as a rotation around a common pole. Circles around that pole correspond to transform faults (single lines connecting divergent plate boundary marked by double line). Note that the pole of rotation of the two plates is not the same as the pole of rotation (spin axis) of the sphere.

postulates of plate tectonics. It is not totally supported by evidence, but it appears to be a reasonable approximation of what actually happens in most cases. It is needed to permit the mathematical reconstruction of past plate configurations.

The joint pole of rotation of two plates can be determined from their transform boundaries, which are by definition parallel to the direction of motion and so form small circles around the pole. The rate of motion can be computed from the increase in age of the crust away from the divergent plate boundary usually by means of magnetic anomalies (see below *Seafloor spreading*). Because all plates form a closed system, all movements can be defined by dealing with them two at a time.

It is, of course, conceivable that the entire lithosphere might slide around over the asthenosphere like a loose skin, altering the positions of all plates with respect to the spin axis of the Earth and the equator. To determine the true geographic positions of the plates in the past, which is so important in paleoclimatology and paleoceanography, investigators have to define their motions, relative not to each other but rather to this independent frame of reference. The hot spot island chains serve this purpose, their trends providing the direction of motion of a plate; the speed of the plate can be inferred from the increase in age of the volcanoes along the chain. It is assumed, of course, that the hot spots themselves remain fixed with respect to the Earth, an assumption that appears to be reasonably accurate for at least some hot spots.

Hot spot island chains as a frame of reference

Quite another method of determining absolute plate movements relies on the fact that the equatorial waters of the ocean are, and always have been, very fertile. The high biological productivity yields an enormous quantity of calcareous microfossils, which, like a gigantic natural chalk line, marks a narrow equatorial zone. The displacement of the equatorial deposits over time, traced by means of deep-sea drill cores, enables investigators to determine the direction and rate of plate movement.

Because the plates all interlock, any change in motion anywhere must reverberate throughout the entire system. If two continents collide, their edges will crumple and shorten, but eventually all motion must stop at this boundary and large parts of the system elsewhere will have to adjust. Earth scientists are thus able to reconstruct the positions and movements of plates in the past so long as they have the ancient oceanic crust to provide them with plate speeds and directions. Since old oceanic crust is continuously consumed to make room for new crust, this kind of evidence is eventually exhausted. Little oceanic crust of Early Cretaceous age (about 144 to 97.5 million years old) remains, and none is older than the Jurassic—the geologic period that began approximately 208 million

years ago. Consequently, this method fails for the history of drifting continents during earlier geologic periods (*i.e.*, the Paleozoic and Precambrian eras), making it necessary for investigators to turn to another, less effective technique (see below *Paleomagnetism, polar wandering, and continental drift*).

HISTORICAL OVERVIEW

Precursors. Any major new idea in science appears to lead instantly to a search of the past for those who might once have proposed similar concepts and with whom the current proponents should therefore share the credit. In the case of plate tectonics, the primary candidate is obvious: Alfred Wegener of Germany, who explicitly presented the concept of continental drift for the first time at the outset of the 20th century. Though plate tectonics is by no means synonymous with continental drift, it encompasses this idea and derives much of its impact from it.

There might have been predecessors even to Wegener. The outlines of the continents bordering the Atlantic Ocean are so similar that many probably noticed the correspondence, and some might have drawn the conclusion that the lands on both sides were once joined together. The earliest reference to this peculiar geographic feature was made by the English philosopher Francis Bacon. In his *Novum Organum* (1620), Bacon pointed out the correspondence but did not go beyond that. Such was also the extent of the contribution of the great French naturalist Georges-Louis Leclerc, Count du Buffon, a century later. Neither can François Paget qualify as a forerunner of continental drift theorists: even though he stated in 1666 that an undivided continent existed before Noah's flood, he explained the creation of the Atlantic Ocean by having part of that continent sink into the sea.

The first credible proponent of continental drift was Antonio Snider-Pellegrini, a belated advocate of catastrophism (the view that geologic history consists of a sequence of numerous violent catastrophic events), who in 1858 ascribed the biblical flood to the former existence of a single continent that was torn apart to restore the balance of a lopsided Earth. More recent and much more sophisticated was the work of the American geologist Frank B. Taylor, who, disdaining the then-prevailing contraction model of mountain building, postulated in 1908 that the arcuate mountain belts of Asia and Europe resulted from the equatorward creep of the continents. His analysis of tectonic features foreshadowed in many ways modern thought regarding plate collisions, and he anticipated Wegener's publications by only a few years. Curiously, however, his work instantly sank into oblivion.

Alfred Wegener and the concept of continental drift. Wegener was by training and profession a meteorologist (he was highly respected for his work in climatology and paleoclimatology), but he is best remembered for the foray into geology that led to his formulation of the concept of continental drift. In 1910, again because of the geography of the Atlantic coastlines, Wegener came to consider the existence of a single supercontinent during the late Paleozoic era (about 350 to 225 million years ago) and named it Pangaea. He searched the geologic and paleontological literature for evidence attesting to the continuity of geologic features across the Indian and Atlantic oceans, which he assumed had formed during the Mesozoic era (about 245 to 66.4 million years ago). His efforts proved rewarding, and he presented the idea of continental drift and some of the supporting evidence in a lecture in 1912. This was followed in 1915 by his major work, *Die Entstehung der Kontinente und Ozeane* (*The Origin of Continents and Oceans*).

The supercontinent Pangaea

The idea of large ancient continents composed of several of the present-day smaller ones had been put forth in the late 19th century by the Austrian geologist Eduard Suess. Suess, however, was not thinking of continental drift. In the spirit of his day, he assumed that portions of a single enormous southern continent—designated Gondwana, or Gondwanaland—foundered to become the Atlantic and Indian oceans. Such sunken lands, along with vanished land bridges, were frequently invoked in the late 1800s to explain sediment sources apparently present in the ocean and to account for floral and faunal connections between continents. They remained popular until the 1950s, stimulated believers in ancient Atlantis, and even made their way into literary works.

Yet, it was already known that the concept of isostasy rendered large sunken continental blocks geophysically impossible, and Wegener characteristically introduced his continental drift proposal by pointing this out. Only then did he proceed to conclude that, if the continents had been once joined together, drift of their fragments rather than their foundering would have been the consequence. The assumption of a former single continent could be tested geologically, and Wegener next displayed a large array of data. Even today his evidence, ranging from the continuity of fold belts across oceans and similarities of sequences of strata on their opposite sides to paleobiogeographic and paleoclimatological arguments, would be judged worthy of serious consideration. He further argued that, if continents could move up and down in the mantle as a result of buoyancy changes produced by, say, erosion or deposition, they should be able to move horizontally as well. The driving forces he considered, however, were unconvincing: both pole fleeing and the westward tidal force appeared to most to be entirely inadequate.

Wegener's proposition was attentively received by many European geologists, and in England Arthur Holmes pointed out that the lack of a driving force was hardly sufficient grounds to scuttle the entire concept. As early as 1929, Holmes proposed an alternative mechanism—namely, convection of the mantle, which remains today a serious candidate for the force driving the plates. Wegener's ideas also were appreciated by geologists in the Southern Hemisphere. One of them, the South African Alexander Du Toit, remained a lifelong believer. After Wegener's death, Du Toit continued to amass further evidence in support of continental drift.

Evidence supporting the hypothesis. Much was thus to be said for the idea that the continents were joined together in the Paleozoic, and supporting evidence has continued to accumulate to this day. The opposing Atlantic shores match well, especially at the 1,000-metre (3,300-foot) depth contour, which is a better approximation of the edge of the continental block than the present shoreline, as Sir Edward Bullard cogently demonstrated in 1964 with the aid of computer analysis. Similarly, the structures and stratigraphic sequences of Paleozoic mountain ranges in eastern North America and northwestern Europe can be matched in detail. This fact was already known to Wegener and has been strengthened substantially in subsequent years.

Often cited as evidence have been the strikingly similar Paleozoic sequences on all southern continents and also in India. This Gondwana sequence—so called after one of Suess's large continents—consists of glacial tillites, followed by sandstones and finally coal measures. Placed on a reconstruction of Gondwana, the tillites mark two ice ages that occurred during the long march of this continent across the South Pole from its initial position north of Libya about 500 million years ago until its final departure from southern Australia 250 million years later (Figure 3). The first of these ice ages left its glacial deposits in the southern Sahara during the Silurian period (which extended from about 438 to 408 million years ago), and the second did the same in southern South America, South Africa, India, and Australia from 380 to 250 million years ago. At each location the tillites were subsequently covered by desert sands of the subtropics, and these in turn by coal measures, indicating that the region had arrived near the equator.

Gondwana sequence

During the 1950s and '60s, patient work in isotopic dating showed that the massifs of Precambrian time (from about 3.8 billion to 570 million years ago) found on opposite sides of the South Atlantic did indeed closely correspond in age and composition, as Wegener had surmised. It is now evident that they originated as a single assemblage of Precambrian continental nuclei later torn apart by the breakdown of Pangaea.

Disbelief and opposition. More common than interest or approval, however, was a disbelief so strong that it often

bordered on indignation. One of the strongest opponents was the British geophysicist Sir Harold Jeffreys, who spent years attempting to demonstrate that continental drift is impossible because the strength of the mantle should be far greater than any conceivable driving force. He refused to abandon this viewpoint in spite of the massive evidence in favour of plate tectonics. It was in North America, however, that opposition to Wegener's ideas was vigorous to the point of excess and very nearly unanimous. This is illustrated by reports from a symposium on continental drift organized in 1928 by the American Association of Petroleum Geologists. The mood that prevailed at the gathering was expressed by an unnamed attendant quoted with sympathy by the great American geologist Thomas C. Chamberlin: "If we are to believe Wegener's hypothesis, we must forget everything which has been learned in the last 70 years and start all over again." The same reluctance to start anew was again displayed some 40 years later by the same organization when its publications provided the principal forum for the opposition to plate tectonics.

Wegener was attacked from virtually every possible vantage point, his paleontological evidence attributed to land bridges, the similarity of strata on both sides of the Atlantic called into question, the fit of Atlantic shores declared inaccurate, and his very competence doubted. It also did not escape attention that he did not possess proper credentials as a geologist.

What might have been the cause of this overwhelmingly negative response in the light of such a substantial amount of supporting evidence? The unsatisfactory quality of Wegener's driving mechanism has commonly been cited as the reason, but that seems too simple, especially since the absence of a mechanism did not delay the acceptance of plate tectonics. The roots of the resistance most certainly reached far deeper. It would be unusual for the practitioners of any science to flock to a new concept—particularly a revolutionary one of such profound consequences—before the need for a thorough overhaul of the existing conceptual edifice had become compelling and obvious to most, its supporting evidence daily crumbling, and its explanatory power reduced below any acceptable level.

Whatever the cause, continental drift, having been rejected by the vast majority of geologists the world over, retreated into obscurity and remained there for roughly three decades. Ironically, though, Du Toit so successfully kept the fires burning in the Southern Hemisphere that it remained quite respectable there to profess oneself an adherent of continental drift during the very years that such a confession north of the equator would have exposed one to ridicule and disbelief.

Renewed interest in continental drift. *Paleomagnetism, polar wandering, and continental drift.* The fact that some rocks are strongly magnetized has been known for centuries, and geologists recognized more than 100 years ago that many rocks preserve the imprint of the Earth's magnetic field as it was at the time of their formation. Volcanic rocks such as basalt are especially good recorders of paleomagnetism, but some sediments also align their magnetic particles with the Earth's field at the time of deposition. Investigators therefore have at their disposal fossil compasses that indicate, like any magnet suspended in the Earth's field, the direction to the magnetic pole and that yield the latitude of their origin.

From T.H. van Andel, *New Views on an Old Planet* (1985); Cambridge University Press

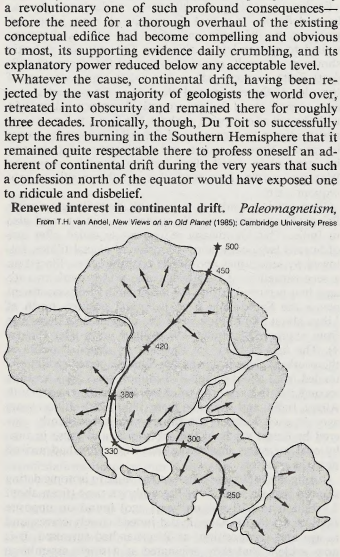


Figure 3: The trail of the South Pole across Gondwanaland during the Paleozoic era. The numbers indicate its progress by giving the dates of pole positions in millions of years. The arrows show the flow directions of ice caps appearing twice during the long march as ice ages overtook the Earth.

The paleomagnetic studies of Runcorn

During the 1950s, paleomagnetic studies, notably those of Stanley K. Runcorn and his coworkers in England, showed that in the late Paleozoic the north magnetic pole—as seen from Europe—seems to have wandered from a Precambrian position near Hawaii to its present location by way of Japan. This, of course, might mean that the magnetic pole itself had migrated or that Europe had moved relative to a fixed pole. Either continental drift or polar wandering was therefore a reasonable explanation. Paleomagnetic data from other continents, however, soon yielded apparent polar wandering paths different from the European one. Separate wanderings of many magnetic poles are not acceptable, but the paths could be brought to coincide by joining the continents in the manner and at the time suggested by Wegener.

Impressed by this result, Runcorn became the first of a new generation of geologists and geophysicists to accept continental drift as a serious proposition worthy of careful testing. Yet, the band of the converted remained small, and most geologists found sufficient reason to doubt the paleomagnetic results, which were often conflicting due to the primitive nature of the early techniques. Since then, more sophisticated methods capable of removing the overprint of later magnetizations have made paleomagnetic data strong supporting evidence for continental drift and a major tool for reconstructing the geography of the past. In the meantime, however, much progress had been made on an entirely different and quite independent front. It was from evidence extracted from the oceanic crust that plate tectonics would be born.

Gestation and birth of plate tectonics. When Wegener developed his ideas, and for many years thereafter, relatively little was known regarding the nature of the ocean floor. After World War II, however, rapid advances were made in the study of the relief, geology, and geophysics of the ocean basins. Due in large part to the efforts of Bruce C. Heezen and Henry W. Menard of the United States, these features, which constitute more than two-thirds of the Earth's surface, became well enough known to permit serious geologic analysis.

Ocean basin studies

Several major topographic and tectonic features distinguish the ocean basins from the continents. The first of these is the mid-ocean ridge system. Mid-ocean ridges are broad, elongated elevations of the ocean floor rising to about 2.5 or 3 kilometres below sea level, with widths ranging from a few hundred to more than 1,000 kilometres. Their crests tend to be rugged and are often endowed with a longitudinal rift valley where fresh lava flows, high heat flow, and shallow earthquakes of the extensional type are found. Mid-ocean ridges nearly girdle the globe.

Trenches constitute another type of seafloor feature. In contrast to mid-ocean ridges, they are long, narrow depressions containing the greatest depths of the ocean basins. Trenches virtually ring the Pacific; a few also occur in the northeastern part of the Indian, and some small ones are found in the Atlantic. Elsewhere they are absent. Trenches have low heat flow, are often filled with thick sediments, and lie at the upper edge of the Benioff zone of compressive earthquakes. Trenches border continents, as in the case of western Central and South America; but they also may occur in mid-ocean, as, for example, in the southwestern Pacific.

Mid-ocean ridges and, more rarely, trenches are offset by fracture zones—transverse features consisting of linear ridges and troughs approximately perpendicular to the ridge crest that they offset by a few to several hundred kilometres. Fracture zones often extend over long distances in

the ocean basins but generally end abruptly against continental margins. They are not volcanic, and their seismic activity is restricted to the area between offset ridge crests where earthquakes indicating horizontal slip are common.

The existence of these three types of striking, large seafloor features, which had gradually become evident during the late 1940s and '50s, clearly demanded a global rather than local tectonic explanation. The first comprehensive attempt at such an explanation was made by Harry H. Hess of the United States in a widely circulated manuscript that he had written in 1960 but not formally published for several years. In this paper Hess, drawing on Holmes's model of convective flow in the mantle, suggested that the mid-ocean ridges were the surface expressions of rising and diverging convective flow while trenches and Benioff zones with their associated island arcs marked descending limbs. At the ridge crests new oceanic crust would be generated and then carried away laterally to cool, subside, and finally be destroyed in the nearest trenches. Consequently, the age of the oceanic crust should increase with distance away from the ridge crests, and, because recycling was its ultimate fate, very old oceanic crust would not be preserved anywhere. This, incidentally, took care of an old and troubling paradox: only rocks younger than Mesozoic had ever been encountered in the oceans, whereas the continents bear ample evidence of the presence of oceans for more than three billion years.

Hess's seafloor-spreading model

Hess's model, later dubbed seafloor spreading by the American oceanographer Robert S. Dietz, appeared to account for most observations and was received with interest by many marine geologists. Confirmation of the production of oceanic crust at ridge crests and its subsequent lateral transfer was not long in coming. Fracture zones had thus far been widely regarded as transcurrent faults that gradually displaced one crestal block to the right or left relative to the other. Given this interpretation, the abrupt termination of many fracture zones against continental margins raised intractable problems. The aforementioned Canadian geologist J. Tuzo Wilson solved these problems in 1965 with a single ingenious stroke. Suppose, he argued, that the offset between two ridge crest segments is present at the outset. Each segment generates new crust, which moves laterally away. Along that part of the fracture zone lying between crests, the crustal slabs move in opposite directions, even though the axes or rift valleys themselves remain stationary. Beyond the crests, adjacent portions of crust move in parallel and are eventually absorbed in a trench. Wilson called this a transform fault and noted that on such a fault the seismicity should be confined to the part between ridge crests, as is indeed the case (Figure 4).

From T.H. van Andel, *New Views on an Old Planet* (1985); Cambridge University Press

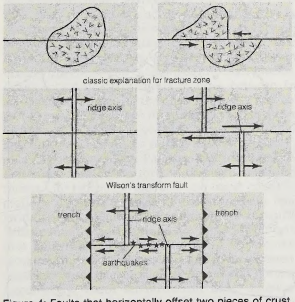


Figure 4: Faults that horizontally offset two pieces of crust are called strike-slip faults (top). Fracture zones were initially regarded as classic strike-slip faults until J. Tuzo Wilson demonstrated that they result from the generation of new crust at already offset mid-ocean ridges. The concantration of earthquakes is in the axial part only, and the directions of motion confirm the hypothesis.

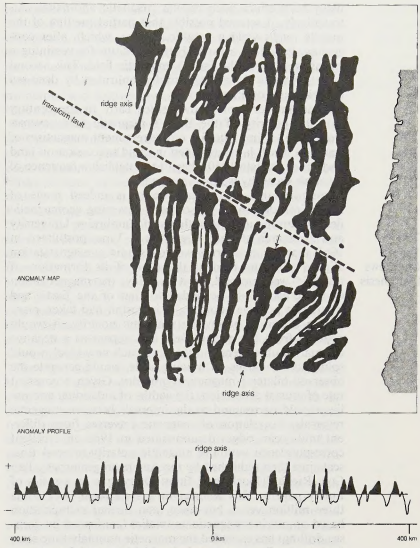


Figure 5: *Map and profile of magnetic anomalies.* (Top) Magnetically, the oceanic crust is distinguished by a pattern of values alternately higher (black) and lower (white) than the average of the Earth's field. Note the offset of the magnetic bands along a transform fault. (Bottom) A corresponding magnetic anomaly record showing the symmetry of anomalies on both sides of the ridge axis produced by seafloor spreading.

From T.H. van Andel, *New Views on an Old Planet* (1985); Cambridge University Press

Shortly afterward Lynn R. Sykes, an American seismologist, showed that the motions deduced from earthquakes on transform faults conform to the directions of motion postulated by Wilson and are opposite those observed on a transform fault.

The seafloor-spreading model also required that the oceanic crust should increase in age with distance from the ridge axis, and Wilson had already pointed out that volcanic islands in the Atlantic indeed show this pattern. Such islands are few, however, and it is in the nature of these piles of lava and ash that the moment of their birth is difficult to ascertain. Additional evidence was needed, and it soon came from magnetic surveys of the oceanic crust.

A magnetic survey of the eastern Pacific floor off the coast of Oregon and California had been published in 1961 by two geophysicists, Arthur D. Raff and Ronald G. Mason. The results were puzzling and gave rise to many farfetched interpretations. Unlike on the continents, where magnetic anomaly patterns tend to be confused and seemingly random except on a fine scale, the seafloor possesses a remarkably regular set of magnetic bands alternately higher and lower than the average Earth field. These positive and negative anomalies are strikingly linear and parallel with the mid-ocean ridge axis, show distinct offsets along fracture zones, and, when displayed in black and white, generally resemble the pattern of a zebra skin (Figure 5). The axial anomaly tends to be higher and wider than the adjacent ones, and in most cases the sequence on one side is the approximate mirror image of that on the other.

Magnetic anomalies

In his convection/seafloor-spreading model, Hess had attributed the formation of the oceanic crust mainly to the hydration of a peridotitic mantle, a process not judged

likely to produce such regular magnetic anomalies. Alternatively, it seemed possible that partial melting of the mantle might yield a basaltic magma, which, after congealing, would be a much better medium for retaining a strong imprint of the Earth's magnetic field. This second hypothesis has since been amply confirmed by deep-sea dredging and drilling.

It had further been known since early in the century that the polarity of the Earth's magnetic field reverses from time to time. Studies of the remanent magnetism of stacks of basalt lavas extruded in rapid succession on land had, since the late 1950s, begun to establish a sequence of reversals dated by isotopic methods.

The Vine-Matthews hypothesis

Assuming that the oceanic crust is indeed made of basalt intruded in an episodically reversing geomagnetic field, Drummond H. Matthews of Cambridge University and a research student, Frederick J. Vine, postulated in 1963 that the new crust would assume a magnetization aligned with the field at the time of its formation. If the field were normal, as it is today, the magnetization of the crust would be added to that of the Earth and produce a positive anomaly. If intrusion had taken place during a period of reverse magnetic polarity, it would subtract from the present field and appear as a negative anomaly. Subsequent to intrusion, each new block would split and the halves, in moving aside, would generate the observed bilateral magnetic symmetry. Given a constant rate of crustal generation, the widths of individual anomalies should correspond to the intervals between magnetic reversals. Correlation of magnetic traverses from different mid-ocean ridges demonstrated in 1966 an excellent correspondence with the magnetic polarity-reversal time scale just then published by the American geologists Allan Cox, Richard Doell, and Brent Dalrymple in a series of timely papers. This reversal time scale went back some three million years, but since then further extrapolation based on marine magnetic anomalies (confirmed by deep-sea drilling) has extended the magnetic anomaly time scale far into the Cretaceous period, which spanned from about 144 to 66.4 million years ago.

As an aside, it is of interest to note that in Canada Laurence W. Morley, simultaneously with Vine and Matthews but entirely independently, had come to the same explanation for the marine magnetic anomalies, but publication of his paper, delayed by unsympathetic referees and technical problems, occurred long after Vine's and Matthews' work had already firmly taken root.

These confirmations persuaded a large number of marine geologists that seafloor spreading was a reality. Continental drift, however, was not much in their minds, as they focused mainly on the explanations that the concept provided for a host of oceanic features. Land geologists were disinterested, viewing the affair as primarily an issue for their marine colleagues.

Two concerns, however, remained. The spreading seafloor was generally seen as a thin skin most likely having its base at the Mohorovičić discontinuity—*i.e.,* boundary between the crust and mantle considered of such major importance in the early 1960s that plans were undertaken to sample it by deep drilling in the oceans. If only oceanic crust were involved, as seemed to be the case in the Pacific Ocean, the thinness of the slab was not disturbing, even though the ever-increasing number of known fracture zones with their close spacing implied oddly narrow, long convection cells. More troubling was the fact that the Atlantic Ocean, though it had a well-developed mid-ocean ridge, lacked trenches adequate to dispose of the excess oceanic crust. There, the adjacent continents needed to travel with the spreading seafloor, a process that, given the thin but clearly undeformed slab, strained credulity.

Working independently but along very similar lines, Dan P. McKenzie and Robert L. Parker of Britain and W. Jason Morgan of the United States resolved these issues. McKenzie and Parker showed with a geometric analysis that, if the moving slabs of crust were thick enough to be regarded as rigid and thus to remain undeformed, their motions on a sphere would lead precisely to those divergent, convergent, and transform boundaries that are indeed observed. Morgan demonstrated that the directions and rates of movement had been faithfully recorded by magnetic anomaly patterns and transform faults. He also proposed that the plates extended approximately 100 kilometres to the base of a rigid lithosphere, which had long been known to be underlain by a weaker asthenosphere marked by strong attenuation of earthquake waves. In 1968 the French geophysicist Xavier Le Pichon refined these propositions with a computer analysis of all plate data and proved that they did indeed form an integrated system where the sum of all crust generated at mid-ocean ridges is balanced by the cumulative amount destroyed in all subduction zones. That same year, the American geophysicists Bryan Isacks, Jack Oliver, and Lynn R. Sykes showed that the theory, which they enthusiastically labeled the "new global tectonics," was capable of accounting for the larger part of the Earth's seismic activity. Almost immediately others began to consider seriously the abil-

The "new global tectonics"

From Kevin C. Burke and J. Tuzo Wilson, "Hot Spots on the Earth's Surface," copyright © (August 1976) by Scientific American, Inc. All rights reserved.

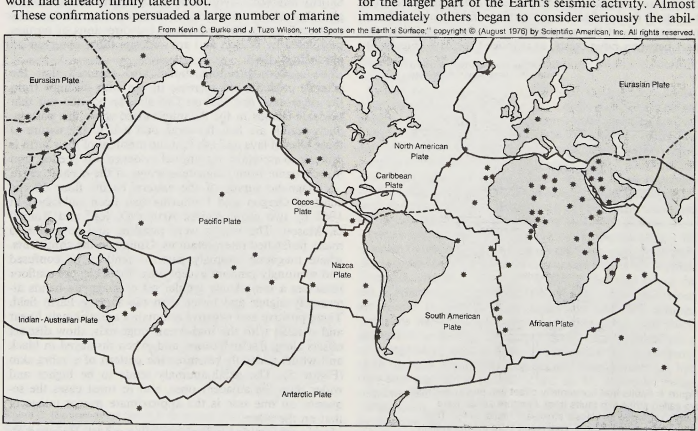


Figure 6: Principal plates that make up the Earth's lithosphere. Very small plates ("microplates") have been omitted. The stars indicate intraplate volcanoes, many of which are thought to be hot spots.

ity of the theory to explain mountain building and sea-level changes.

Only a few years later, details of the processes of plate movement and of boundary interactions, along with much of the plate history of the Cenozoic era (the past 66.4 million years), had been worked out. Yet, the driving forces—notwithstanding a brief flurry of discussion around 1970—remained mysterious and continue as such. The vast accumulation of data bearing on plate history and plate processes has yielded surprisingly little information about what happens beneath them. Pull by the subducting slab, push at the spreading ridge, convection in the asthenosphere, and even tidal forces have been considered, but in every case the evidence has to be admitted as inconclusive. Many favour convection, but, if this indeed is the driving force, the flow pattern at depth is clearly not reflected in the surface movements of the plates, constrained as they are by each other.

PLATE TECTONICS AS AN EXPLANATION FOR EARTH PROCESSES

Since its inception in 1967–68, plate tectonics has had a pervasive impact on the Earth sciences. If it has not fully lived up to the proud name "new global tectonics," it has nevertheless exerted enormous influence by clarifying areas of obscurity, reconciling seemingly conflicting evidence, unifying to a remarkable degree events occurring in distant parts of the globe, establishing new pathways toward knowledge, and opening the door to new subjects of investigation, all the while raising a myriad of new and important issues. A few examples may suffice to illustrate the dimensions of the revolution it has wrought.

Pangaea and its breakup. Magnetic anomalies, transform faults, and hot spot trails permit rigorous geometric reconstructions of past plate shapes, configurations, and movements. As previously noted, reasonable agreement on such reconstructions can be achieved as far back as the late Cretaceous, but beyond that time little oceanic crust remains. Much effort in reconstructing the geography of the past 100 million years or so has shown that, though some reconstructions are straightforward, others imply that the strict application of the assumption that plates are not internally deformed is not justified. Such is the case, for example, for the Pacific basin as opposed to the simple history of the Atlantic.

Reconstructions of ancient geography

When dealing with geologic time beyond the Cretaceous, other means have to be applied in reconstructing ancient geography. Geologic data, for instance, can be used in the time-honoured manner to determine the proper fit of continents to make Pangaea. Only minor controversy continues here, such as in the vexing case of the position of Madagascar or of the many fragments that now constitute the Mediterranean, southeastern Europe, and the Middle East. Old collision boundaries are marked by sutures—*i.e.*, zones of deformation that betray points of contact between formerly separate continents even after much erosion. Good examples include the Urals between the old central European and Siberian continental blocks, and the Indus suture north of the Himalayas between India and Asia. Some sutures, on the other hand, are well concealed, as, for example, the one in northern Florida and southern Georgia that marks a collision between North America and northwestern Africa during the Paleozoic.

Having reconstructed Pangaea and identified the pieces that came together to form it through largely traditional means, investigators are able to trace continental migrations using paleomagnetic data to obtain their paleolatitudes and orientations for appropriate instants in time. What cannot be determined are their paleolongitudes, so that knowledge about the widths of intervening oceans remains limited. Paleolongitudes can be estimated only by relatively crude means such as faunal and floral affinities. Accordingly, Earth scientists will never know as much about the geography of the Precambrian, Paleozoic, or early Mesozoic as they do about the last 100 million years of Earth history. Yet, enough information has become available to permit some undoubtedly flawed but fascinating and useful reconstructions.

A widely used Paleozoic reconstruction by a group of geologists at the University of Chicago shows that throughout the Paleozoic, and probably for some time before then, Gondwana existed, consisting of South America, Africa, and Australia. It was surrounded by an ever-shifting array of landmasses, some large—about the size of North America, China, or Siberia—and others small, though numerous and destined eventually to become parts of Europe and Asia.

Paleoclimates and ancient oceans. For Gondwana itself, geologic data suffice to recreate in broad outline form not only its coasts, plains, and mountain ranges but also its deserts, polar ice caps, and tropical jungles. These reflect broad climatic zones, which were themselves controlled mainly by insolation, seasonality, and the rotating of the Earth. Large landmasses may sometimes have modified climate to produce monsoons. Given the positions and sizes of continents, paleoclimatic zones may be inferred and then checked against climatic conditions deduced from the fossil record of sedimentary rocks. The results show a satisfactory correspondence between expectations and record, much better than the one obtained by plotting paleoclimatic data on the present-day configuration of continents. They leave little doubt that it is better to accept that the continents have drifted about than to assume that they have remained stationary.

The supercontinent Pangaea was completely surrounded by a world ocean extending from pole to pole and spanning 80 percent of the circumference of the Earth at the equator. The equatorial current system, driven by the trade winds, resided in warm latitudes much longer than it does today, and its waters were therefore warmer. The gyres that occupy most of the Southern and Northern hemispheres were also warmer, and consequently the temperature gradient from the equator to the poles was much reduced in comparison to the present.

Early in the Mesozoic, the equatorial current became circumglobal when the Tethys seaway split Gondwana from its northern counterpart, Laurasia. The equatorial surface waters, now able to circumnavigate the world one or more times, became even warmer; but whether the return flows continued to keep the high latitudes warm or diminished in strength and therefore somewhat increased the latitudinal temperature gradients is not certain. Nevertheless, in the middle and later Mesozoic the Arctic and Antarctic surface water temperatures were at or above 10° C (50° F), and the polar regions were warm enough to support forests. As the dispersal of continents following the breakup of Pangaea continued, however, the surface circulation of the oceans became much more complex. During the middle Cenozoic, the northward drift of Australia and South America created a new circumglobal seaway around Antarctica that remained centred on the South Pole. A vigorous circum-Antarctic current developed, isolating the southern continent from the warmer waters to the north. At the same time, the equatorial current system became blocked, first in the Indo-Pacific region, next in the Middle East and eastern Mediterranean, then at Gibraltar, and finally, about 5 million years ago, by the emergence of the Isthmus of Panama. As a result, the equatorial waters were heated less and the mid-latitude ocean gyres were not as effective in keeping the high latitudes warm. Because of this, an ice cap began to form on Antarctica some 20 million years ago and grew to roughly its present size about 5 million years later. This ice cap cooled the waters of the adjacent ocean to such a low temperature that the waters sank and initiated the north-directed abyssal flow that marks the present deep circulation. The Quaternary Ice Age arrived in full when the first ice caps appeared in the Northern Hemisphere about 2 million years ago.

The dispersal of the continents and its impact on oceanic current systems

There is no certainty that the changing configuration of continents and oceans can be held solely responsible for the onset of the Quaternary Ice Age, even if such factors as the drift of continents across the latitudes (with the associated changes in vegetation) and reflectivity for solar heat are included. There can be little doubt, however, that it was a major contributing factor and that recognition of its role has profoundly altered all concepts of paleoclimatology.

Plate tectonics and mountain building. Subduction and

The significance of plate convergence in orogeny

continental collision raise mountain ranges. Consequently, the implications of plate tectonics for the processes of mountain building have attracted much attention. One of the earliest to apply the new theory was the Cambridge geologist John Dewey, who analyzed the Appalachian and Alpine orogenies. Many other researchers have subsequently undertaken similar work in the Mediterranean system and the American Cordilleran ranges, as well as in the Appalachians.

The collisions that gradually joined North America to Pangaea during the Paleozoic exemplify the role of plate convergence in mountain building. Some 500 million years ago, a subduction zone existed along what is now the eastern seaboard of North America. During the course of the next 67 million years or so—the Ordovician period—this region underwent a major phase of mountain building. When it was over, the direction of sediment transport had reversed, indicating that a large source area had appeared to the east of the subduction zone. The thick sequence of deposits from this source in eastern North America bears witness to its size, which is incommensurable with the island arcs that normally accompany a subduction zone. Subduction, sedimentation, and volcanism then continued until a small continent comprising what is today northwestern Europe collided with the northeastern tip of North America. This collision raised a major chain of folded mountains, the remains of which are now found from New England through eastern Canada and Scotland into Norway. This Caledonian range shed sediments to both sides, forming the Catskill delta in eastern North America and the Old Red Sandstone in northwestern Europe. Subsequently, the southern edge of North America collided with South America, and a little later Africa arrived, producing the southern Appalachians. Miscellaneous fragments, today part of central Europe and the western Mediterranean, filled in the gap and completed this portion of Pangaea.

This account fits the plate model fairly well except for two things: (1) the occurrence of paroxysms of mountain building when no continental collision can be assumed and (2) the need for sizable landmasses, presumably small continents, floating offshore to provide a necessary sediment source. Both problems appear quite often wherever plate movements are used to explain the details of the formation of major mountain ranges. Yet, the model itself is simple. It allows for three possible cases: ocean–ocean collisions, ocean–continent collisions, and continent–continent collisions. In the first two situations, oceanic crust is subducted either under other oceanic crust or under a continent—a steady process presumably accompanied by an equally steady scraping off and deforming of sediments and basaltic crust at the plate edge and by volcanism farther away. The New Hebrides trench and island arcs in the southwestern Pacific exemplify an ocean–ocean collision, while the Andes mountain ranges of South America represent a long-lasting collision of oceanic crust and continental crust at the leading edges of plates.

Even in these cases the process appears to be more complicated than the theory suggests, demanding modifications that, though eminently plausible and often proven, are ad hoc and thus diminish its value as a unifying principle. The South American Andes, for example, have had an episodic history not readily matched to the evidently uneventful spreading of the seafloor in the adjacent Pacific. The volcanism of the segments of the Cascade Range in Oregon and Washington in the northwestern United States is only a few million years old in its present form, whereas subduction along the Pacific coast dates back much farther than that. All along the west coast of North America, mountain building has occurred in distinct major phases, ultimately caused in all probability by changes in the rate and direction of plate movements that complicate the basically simple subduction process.

One of the complicating factors is much in evidence in the coast ranges that extend from California to southern Alaska. Large sections of these mountains fit poorly with the geology of the surrounding terrain and are thought to have originated somewhere in the ancient Pacific. Perhaps they once were oceanic plateaus, island arcs, or other thick pieces of oceanic crust that would not subduct but instead became lodged against the continental edge, and so disturbed the processes of subduction and mountain building. Some of these "exotic terranes" have been shown by their paleomagnetic properties to have come a long way from their points of origin, perhaps as far as the Southern Hemisphere.

The Indus-Himalayan-Tibetan region in northern India and China represents a continent–continent collision that began 40 million years ago when all the oceanic crust between northward-drifting India and south-central Asia had been consumed and the continent-bearing edges of the two plates met. The suture of this collision, the site of the former subduction zone, lies inconspicuously in the high plateau region north of the Himalayas. This lofty mountain range itself formed long after the collision as India—its momentum not fully spent even today—drove another 2,000 kilometres into the underbelly of Asia. As a result, Asian lithosphere was displaced laterally along enormous transcurrent faults, which are the sites of the numerous devastating earthquakes in Iran, Afghanistan, and western China (Figure 7). The Himalayas themselves are the product of the northward, low-angle underthrusting of the upper slice of the Indian-Australian Plate under the Eurasian Plate, making it clear that a continent–continent collision involves a good deal more than the crumpling of opposing front edges.

Plate tectonics and life. Inevitably the continuous rearrangement over time of the size and shape of ocean basins and continents, followed by changes in ocean circulation and climate, have had a major impact on the development of life on Earth. Active interest in these aspects of the Earth science revolution has lagged behind that in other areas, even though as early as 1970 the American geologists James W. Valentine and Eldridge M. Moores

From T.H. van Andel, *New Views on an Old Planet* (1985); Cambridge University Press

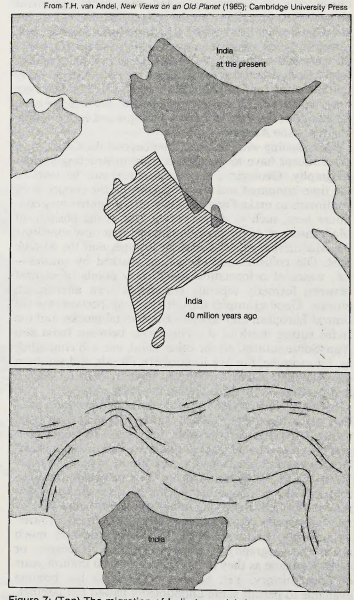


Figure 7: (Top) The migration of India toward Asia, and (bottom) the earthquake faults that result from India's continuing penetration into the Asian continent.

attempted to show that the diversity of life increased as continents fragmented and dispersed and diminished when they were joined together.

The effect of plate activity on the evolution of life

Subsequently, however, the study of plate activity as a force in the evolution of life has leaped forward. A few simple examples of such research must suffice to illustrate the impact of the plate theory. Toward the end of the Paleozoic, during the Permian period (about 286 to 245 million years ago), there was a drastic drop in the variety of animal forms inhabiting the shallow seas around Pangaea. Well over half of the total number of known families became extinct. This drop can be attributed in large part simply to the decrease in biogeographic variety that marks a world consisting of a single continent rather than one comprising many widely dispersed landmasses. Other factors, such as a sharp decrease in the area of shallow-water habitats or a change in ocean fertility due to upwelling, have also been invoked. Moreover, the extinction had a complex history. High latitudes were affected first as a result of the waning of the Permian ice age when the South Pole slipped beyond the southern edge of Pangaea. The equatorial and subtropical zones appear to have been affected somewhat later by a global cooling. On the other hand, the extinctions were not felt so strongly on the continent itself. Instead, the vast semiarid and arid lands that emerged on so large a continent, the shortening of its moist coasts, and the many mountain ranges remaining from the collisions that led to the formation of the supercontinent provided strong incentives for evolutionary adaption to dry or high-altitude environments.

The impact of plate movements and interactions on life is perhaps most clearly demonstrated by what happens when continents diverge or collide. When the Atlantic Ocean began to open during the middle Mesozoic, the similarity between the faunas of opposite shores gradually decreased in almost linear fashion—the greater the distance, the smaller the number of families in common. The difference increased more rapidly in the South Atlantic than in the North Atlantic, where a land connection between Europe and North America persisted until well after the middle Cenozoic. The inverse, the effect of a collision between two hitherto separate landmasses, is illustrated by the consequences of the Pliocene emergence of the Isthmus of Panama. In South America a highly specialized fauna had evolved, rich in marsupials but with few predators. After the emergence of the isthmus had made it possible for land animals to cross, numerous herbivores migrated from north to south. They adapted well to the new environment and were more successful than the local fauna in competing for food. The invasion of highly adaptable carnivores from the north contributed to the extinction of no fewer than four orders of South American land mammals. Only a few species, notably the armadillo and the opossum, managed to migrate in the opposite direction. Ironically, many of the invading northerners, such as the llama and tapir, subsequently became extinct in their country of origin and found their last refuge in the south.

Early plate activity. Whatever the forces may be that drive the plates, they consume energy. By far the largest part of this energy is derived from the decay of radioactive isotopes within the Earth, and the energy flow has therefore declined through the 4.5 billion years of the Earth's history—rapidly at first and then at a slowly diminishing though not negligible rate. Accordingly, it is quite likely that the behaviour of the lithospheric plates on the early, more energetic Earth was different from what it is today, and what prevails at present will certainly differ from what will prevail in the future. A thickening lithosphere, a decreasing heat flow, a temperature gradient that decreases with depth within the Earth, and enlarging—though perhaps less vigorous—convection cells in the mantle have all been postulated as unidirectional changes that affected the behaviour of the lithosphere. It is possible, for example, that the initial plates were too small, too hot, and hence too light to be subducted. In this case, the first subduction would mark the coming of age of classical plate tectonics, and, indeed, clear evidence is lacking for subduction until rather late in the Precambrian.

The evidence that bears on the existence, nature, and movements of possible lithospheric plates during the first several billion years of Earth history is very limited. The continental nuclei of the early and middle Precambrian seem to have been small and might be regarded as small plates on a more vigorously convecting mantle, though admittedly other explanations are equally possible. These nuclei are thought to have been embedded in strongly deformed complexes of sediments and basic igneous rocks, the greenstone belts reminiscent of the sutures that mark the closure of ancient oceans. In most cases, however, paleomagnetic data do not leave room for the existence of sizable oceanic areas between such nuclei. Investigators are thus forced to contemplate the possibility that, in the Precambrian, intensive deformation took place within plates, perhaps commensurate with the postulated thinner lithosphere and higher flow of heat toward the surface. On the other hand, current knowledge of this long but obscure portion of Earth history is so deficient that some geologists have emphatically denied that there might have been a remote past to which classical plate tectonics could not be applied.

DISSENTING OPINIONS AND UNANSWERED QUESTIONS

The dissenters. Scientific revolutions as far-reaching in their consequences as the plate tectonics revolution cannot be expected to be easily, or ever completely, accepted. Nevertheless, once the theory had fully emerged, acceptance was quick and widespread and by the late 1960s its influence in the West was pervasive. Such was, however, not the case in the Soviet Union, a country located largely in the continental interior far from present-day plate boundaries. Soviet scientists viewed as central to any issue of global tectonics the vertical movements of continental interiors, phenomena not satisfactorily dealt with by the plate tectonics theory. A leading spokesman for the Soviet position, the academician Vladimir Vladimirovich Belousov, strongly defended a model of the Earth that postulated stationary continents affected almost exclusively by vertical motions. The model, however, only vaguely defined the forces supposedly responsible for the motions. In recent years, a younger generation of Soviet geologists has very gradually come to regard plate tectonics as an attractive theory and a viable alternative to the concepts of Belousov and his followers.

Early opposition to plate tectonics

Opposition to plate tectonics was by no means limited to the Soviet Union. Critics were heard—in fact, some quite loudly—elsewhere as well. Sir Harold Jeffreys continued his lifelong rejection of continental drift on grounds that his estimates of the properties of the mantle indicated the impossibility of plate movements. He did not, in general, deign to take note of the mounting geophysical and geologic arguments that were in favour of a mobile outer shell of the Earth.

Others proffered different explanations of the accumulating evidence, most of which were rather farfetched, as, for instance, the suggestion that new crust was formed at trenches and destroyed on mid-ocean ridges. Still others, notably the American geologists A.A. Meyerhoff and Howard A. Meyerhoff, attempted to assemble data that contradicted the theory and thereby show that the supporting evidence was wrong, insufficient, or simply misconstrued. Demonstrating a remarkable command of often quite obscure literature, they issued a series of negative commentaries in the early 1970s; but they failed to convince the majority of their colleagues, partly because they did not offer alternative explanations for the evidence.

The only serious alternative had been proposed in 1958 by the Australian geologist S. Warren Carey in the form of a new version of an old idea—namely, the expanding Earth model. Carey accepted the existence and early Mesozoic breakup of Pangaea and the subsequent dispersal of its fragments and formation of new ocean basins, but he attributed it all to the expansion of the Earth, the planet presumably having had a much smaller diameter in the late Paleozoic. In his view, the continents represented the pre-expansion crust, and the enlarged surface was to be entirely accommodated within the oceans. This model accounted for a spreading ocean floor and for the young age of the oceanic crust; however, it failed to deal adequately

Carey's version of the expanding Earth model

with the evidence for subduction and compression. Carey's model also did not explain why the process should not have started until some four billion years after the Earth was formed, and it lacked a reasonable mechanism for so large an expansion. Finally, it disregarded the evidence for continental drift before the existence of Pangaea.

Unanswered questions. As the philosopher Thomas J. Kuhn has pointed out, science does not always advance in the gradual and stately fashion commonly attributed to it. Major breakthroughs often come from a leap forward that is at least in part intuitive and may fly in the face of conventional wisdom and widely accepted evidence while strict requirements for verification and proof are temporarily relaxed. Revolutions thus often become widely accepted before the verdict from rigorous analysis of evidence is completely in. Such was certainly the case with the geologic revolution, which also confirms Kuhn's view that a new paradigm is unlikely to supersede an existing one until there is little choice but to acknowledge that the conventional theory has failed. Thus, while Wegener did not manage to persuade the world, the successor theory was readily embraced 40 years later, even though it remained open to much of the same criticism that had caused the downfall of continental drift. What is the state of the new paradigm? Is it likely to suffer sooner rather than later the same fate that inevitably awaits all scientific theories?

In 1974, almost alone among the doubters who tried to discredit the new theory with contrary evidence, the American geologist John C. Maxwell, in a closely reasoned paper, enumerated all the points on which he believed plate tectonics had failed to offer an explanation. Many of these points have since been resolved, but more than a few remain to suggest that the theory, though in essence valid, may be incomplete.

The greatest successes of plate tectonics have been achieved in the ocean basins where additional decades of effort have confirmed its postulates and enabled investigators to construct a credible history of past plate movements. Inevitably in less rigorous form, the reconstruction of early Mesozoic and Paleozoic continental configurations has provided a powerful tool with which to resolve many important questions. On the other hand, the new paradigm has proved less useful in deciphering mountain-building processes or in offering explanations for the complex history of sea-level fluctuations. The American geologist L.L. Sloss has devoted a great deal of effort to demonstrating that continents do indeed rise and fall in unison, but the possible mechanisms for such a process remain elusive.

Where plate boundaries adjoin continents, matters often become very complex and have demanded an ever denser thicket of ad hoc modifications and amendments to the theory and practice of plate tectonics in the form of microplates, obscure plate boundaries, and exotic terranes. A good example is the Mediterranean, where the collisions between Africa and a swarm of microcontinents have produced a tectonic nightmare that is far from resolved. More disturbingly, some of the present plate boundaries, especially in the eastern Mediterranean, appear to be so diffuse and so anomalous that they cannot be compared to the three types of plate boundaries of the basic theory.

There is further evidence, as held by the American geophysicist Thomas H. Jordan, that the base of the plates extends far deeper into the asthenosphere below the continents than below the oceans. How much of an impediment this might be for the free movement of plates and how it might affect their boundary interactions remain open questions. Others have postulated that the lower layer of the lithosphere peels off and sinks late in any collision sequence, producing high heat flow, volcanism, and an upper lithospheric zone vulnerable to contraction by thrusting.

It is understandable that any simple global tectonic model would work better in the oceans, which, being young, retain a record of only a brief and relatively uneventful history. On the continents, almost four billion years of growth and deformation, erosion, sedimentation, and igneous intrusion have produced a complex imprint that, with its intricate zones of varying strength, must directly affect the application of plate forces. Seismic reflection studies of the deep structure of the continents have demonstrated just how complex the events that form the continents and their margins may have been, and their findings sometimes are difficult to reconcile with the accretionary structures one would expect to see as a result of subduction and collision.

Notwithstanding these cautions and the continuing lack of an agreed-upon driving mechanism for the plates, one cannot help but conclude that the plate tectonics revolution has been fruitful and has immensely advanced scientific understanding of the Earth. Like all paradigms in science, it will most likely one day be replaced by a better one; yet there can be little doubt that, whatever the new theory may state, continental drift will be part of it.

BIBLIOGRAPHY. J. TUZO WILSON (ed.), *Continents Adrift and Continents Aground* (1976), contains an excellent and readable set of articles on the plate tectonics revolution drawn from *Scientific American*, many written by its protagonists, with fine introductions by the editor. Similarly, ALLAN COX (ed.), *Plate Tectonics and Geomagnetic Reversals* (1973), offers a well-chosen selection of the original classical papers that produced this revolution in the Earth sciences, from Holmes's work in the early 1900s to contributions in the late 1960s. An excellent explanation of the new Earth science is SEIYA UYEDA, *The New View of the Earth: Moving Continents and Moving Oceans* (1978), which discusses plate theory and its application to the study of the Earth's surface structures. Scholarly, but written for a lay audience, is TJEERD H. VAN ANDEL, *New Views on an Old Planet: Continental Drift and the History of Earth* (1985), an application of plate theory to the climatic, oceanographic, and geographic history of the Earth, and the relation of the theory to the history of life. The history of ideas pertaining to continental drift and plate tectonics has been thoughtfully analyzed in ANTHONY HALLAM, *A Revolution in the Earth Sciences: From Continental Drift to Plate Tectonics* (1973). The same subject is reviewed from a somewhat different vantage point—that of a science writer—in WALTER SULLIVAN, *Continents in Motion: The New Earth Debate* (1974). A summary of the revolution, mostly in a critical vein by many of its principal opponents, is CHARLES F. KAHLE (ed.), *Plate Tectonics: Assessments and Reassessments* (1974), a bit dated but a good substantive statement on the subject. More technical, though not forbiddingly so, are three books rich in detail and substance: PETER J. WYLLIE, *The Way the Earth Works: An Introduction to the New Global Geology and Its Revolutionary Development* (1976); ROBERT H. DOTT, JR., and ROGER L. BATTEN, *Evolution of the Earth*, 3rd ed. (1981), for the impact of the plate theory on research in Earth history; and STEPHEN STANLEY, *Earth and Life Through Time* (1985), a newer and more advanced treatment.

(T.H.v.A.)

Plato and Platonism

Plato was the second of the great trio of ancient Greeks—Socrates, Plato, and Aristotle—who between them laid the philosophical foundations of Western culture. Building on the life and thought of Socrates, Plato developed a profound and wide-ranging system of philosophy. His thought has logical, epistemological, and metaphysical aspects; but its underlying motivation is ethical. It sometimes relies upon conjectures and myth, and it is occasionally mystical in tone; but fundamentally Plato is a rationalist, devoted to the proposition that reason must be followed wherever it leads. Thus the core of Plato's philosophy is a rationalistic ethics.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, Part Ten, Division V, especially Sections 51 and 53.

This article deals with the man, his works and influence, and the subsequent history of Platonism. It is divided into the following sections:

Plato and his thought 893
- Life 893
  - The Academy and Sicily
  - Formative influences
- General features of the dialogues 894
  - Order of composition
  - Persons of the dialogues
  - Thought of the earlier and later dialogues
  - Doctrine of Forms
  - Socrates and Plato
- The earlier dialogues 895
  - Dialogues of search
  - Ethical and political dialogues
  - Metaphysical foundation of Plato's doctrine: *Phaedo*
  - Aesthetic and mystical dialogues
  - The *Republic*
  - Dialogues of critical reconstruction
- The later dialogues 898

Platonism after Plato 899
- Greek Platonism from Aristotle through middle Platonism: its nature and history 899
- Neoplatonism: its nature and history 900
  - Plotinus and his philosophy
  - The later Neoplatonists
- Platonism in the world of revealed religions 903
  - Early Jewish Platonism
  - Ancient and medieval Christian Platonism
  - Islāmic and medieval Jewish philosophy
- Platonism from the Renaissance to modern times 905
- Evaluation of Platonism 906

Major works 906

Bibliography 906

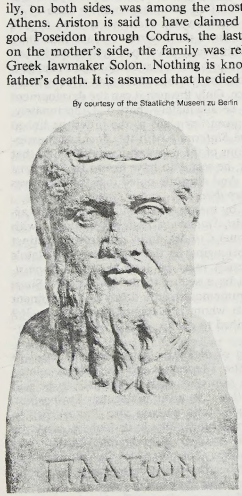
Plato and his thought

LIFE

Plato was born, the son of Ariston and Perictione, in Athens, or perhaps in Aegina, in about 428 BC, the year after the death of the great statesman Pericles. His family, on both sides, was among the most distinguished in Athens. Ariston is said to have claimed descent from the god Poseidon through Codrus, the last king of Athens; on the mother's side, the family was related to the early Greek lawmaker Solon. Nothing is known about Plato's father's death. It is assumed that he died when Plato was a boy. Perictione apparently married as her second husband her uncle Pyrilampes, a prominent supporter of Pericles; and Plato was probably brought up chiefly in his house. Critias and Charmides, leaders among the extremists of the oligarchic terror of 404, were, respectively, cousin and brother of Perictione; both were friends of Socrates, and through them Plato must have known the philosopher from boyhood.

Distinguished family

By courtesy of the Staatliche Museen zu Berlin



Plato. Roman herm probably copied from a Greek original, 4th century BC. In the Staatliche Museen, Berlin.

His own early ambitions—like those of most young men of his class—were probably political. A conservative faction urged him to enter public life under its auspices, but he wisely held back. He was soon repelled by its members' violent acts. After the fall of the oligarchy, he hoped for better things from the restored democracy. Eventually, however, he became convinced that there was no place for a man of conscience in Athenian politics. In 399 BC the democracy condemned Socrates to death, and Plato and other Socratic men took temporary refuge at Megara with Eucleides, founder of the Megarian school of philosophy. The next few years are said to have been spent in extensive travels in Greece, in Egypt, and in Italy. Plato himself (if the *Seventh Letter* is his; see below) states that he visited Italy and Sicily at the age of 40 and was disgusted by the gross sensuality of life there but found a kindred spirit in Dion, brother-in-law of Dionysius I, the ruler of Syracuse.

The Academy and Sicily. In about 387 Plato founded the Academy as an institute for the systematic pursuit of philosophical and scientific teaching and research. He presided over it for the rest of his life. The Academy's interests were not limited to philosophy in a narrow sense but also extended to the sciences: there is evidence that Plato encouraged research in such diverse disciplines as mathematics and rhetoric. He himself lectured (on at least one occasion he gave a celebrated public lecture "On the Good"), and he set problems for his pupils to solve. The Academy was not the only such "school" in Athens—there are traces of tension between the Academy and the rival school of Isocrates.

Political activity in Sicily

The one outstanding event in Plato's later life was his intervention in Syracusan politics. On the death of Dionysius I in 367, Dion conceived the idea of bringing Plato to Syracuse as tutor to his brother-in-law's successor, Dionysius II, whose education had been neglected. Plato was not

optimistic about the results; but because both Dion and Archytas of Tarentum, a philosopher-statesman, thought the prospect promising, he felt bound to risk the adventure. The plan was to train Dionysius II in science and philosophy and so to fit him for the position of a constitutional king who might hold Carthaginian encroachment on Sicily at bay. The scheme was crushed by Dionysius' natural jealousy of the stronger Dion, whom he drove into virtual banishment. Plato later paid a second and longer visit to Syracuse in 361–360, still in the hope of effecting an accommodation; but he failed, not without some personal danger. Dion then captured Syracuse by a coup de main in 357, but he was murdered in 354. Plato himself died in 348/347.

Of Plato's character and personality little is known, and little can be inferred from his writings. But it is worth recording that Aristotle, his most able pupil, described Plato as a man "whom it is blasphemy in the base even to praise," meaning that Plato was so noble a character that bad men should not even speak about him.

Influence of the Academy

To his readers through the ages Plato has been important primarily as one of the greatest of philosophical writers; but to himself the foundation and organization of the Academy must have appeared to be his chief work. The *Seventh Letter* contrasts the impact of written works with that of the contact of living minds as a vehicle of philosophy, and it passes a comparatively unfavourable verdict on written works. Plato puts a similar verdict into the mouth of Socrates in the *Phaedrus.* He perhaps intended his dialogues in the main to interest an educated outside world in the more serious and arduous labours of his school.

All of the most important mathematical work of the 4th century was done by friends or pupils of Plato. The first students of conic sections, and possibly Theaetetus, the creator of solid geometry, were members of the Academy. Eudoxus of Cnidus—author of the doctrine of proportion expounded in Euclid's *Elements,* inventor of the method of finding the areas and volumes of curvilinear figures by exhaustion, and propounder of the astronomical scheme of concentric spheres adopted and altered by Aristotle—removed his school from Cyzicus to Athens for the purpose of cooperating with Plato; and during one of Plato's absences he seems to have acted as the head of the Academy. Archytas, the inventor of mechanical science, was a friend and correspondent of Plato.

Nor were other sciences neglected. Speusippus, Plato's nephew and successor, was a voluminous writer on natural history; and Aristotle's biological works have been shown to belong largely to the early period in his career immediately after Plato's death. The comic poets found matter for mirth in the attention of the school to botanical classification. The Academy was particularly active in jurisprudence and practical legislation. As Plutarch testifies,

> Plato sent Aristonymus to the Arcadians, Phormion to Elis, Menedemus to Pyrrha. Eudoxus and Aristotle wrote laws for Cnidus and Stagirus. Alexander asked Xenocrates for advice about kingship; the man who was sent to Alexander by the Asiatic Greeks and did most to incite him to his war on the barbarians was Delios of Ephesus, an associate of Plato.

The Academy survived Plato's death. Though its interest in science waned and its philosophical orientation changed, it remained for two and a half centuries a focus of intellectual life. Its creation as a permanent society for the prosecution of both humane and exact sciences has been regarded—with pardonable exaggeration—as the first establishment of a university.

Formative influences. The most important formative influence to which the young Plato was exposed was Socrates. It does not appear, however, that Plato belonged as a "disciple" to the circle of Socrates' intimates. The *Seventh Letter* speaks of Socrates not as a "master" but as an older "friend," for whose character Plato had a profound respect; and he has recorded his own absence (through indisposition) from the death scene of the *Phaedo.* It may well be that his own vocation to philosophy dawned on him only afterward, as he reflected on the treatment of Socrates by the democratic leaders. Plato owed to Socrates his commitment to philosophy, his rational method, and his concern for ethical questions. Among other philosophical influences the most significant were those of Heraclitus and his followers, who disparaged the phenomenal world as an arena of constant change and flux, and of the Pythagoreans, with whose metaphysical and mystical notions Plato had great sympathy.

Plato had family connections with Pyrilampes, a Periclean politician, and with Critias, who became one of the most unscrupulous of the Thirty Tyrants who briefly ruled Athens after the collapse of the democracy.

Political setting of Plato's life

Plato's early experiences covered the disastrous years of the Deceleian War, the shattering of the Athenian empire, and the fierce civil strife of oligarchs and democrats in the year of anarchy, 404–403. He was too young to have known anything by experience of the imperial democracy of Pericles and Cleon or of the tide of the Sophistic movement. It is certainly not from memory that he depicted Protagoras, the earliest avowed professional Sophist, or Alcibiades, a brilliant but unreliable Athenian politician and military commander. No doubt these early experiences helped to form the political views that were later expounded in the dialogues.

GENERAL FEATURES OF THE DIALOGUES

The canon and text of Plato was apparently fixed at about the turn of the Christian Era. By reckoning the *Letters* as one item, the list contained 36 works, arranged in nine tetralogies. None of Plato's works has been lost, and there is a general agreement among modern scholars that a number of small items—*Alcibiades I, Alcibiades II, Theages, Erastae, Clitopho, Hipparchus,* and *Minos*—are spurious. Most scholars also believe that the *Epinomis,* an appendix to the *Laws,* was written by the mathematician Philippus of Opus. The *Hippias Major* and the *Menexenus* are regarded as doubtful by some, though Aristotle seems to have regarded them as Platonic. Most of the 13 *Letters* are certainly later forgeries. About the authenticity of the *Seventh Letter,* which is by far the most important from the biographical and the philosophical points of view, there exists a long and unsettled controversy.

Order of composition. Plato's literary career extended over the greater part of a long life. The *Apology* was probably written in the early 380s. The *Laws,* on the other hand, was the work of an old man, and the state of its text bears out the tradition that Plato never lived to give it its final revision. Since there is no evidence that Plato began his career with a fully developed system, and since there is every reason to believe that his thoughts changed, the order in which the various dialogues were written takes on importance. Only through it can the development of Plato's thought be adequately charted. Unfortunately, Plato himself has given few clues to the order: he linked the *Sophist* and the *Statesman* with the *Theaetetus* externally as continuations of the conversation reported in that dialogue. Similarly, he seems to have linked the *Timaeus* with the *Republic.* And Aristotle noted that the *Laws* was written after the *Republic.*

Stylometric studies of the dialogues

Modern scholars, by the use of stylistic criteria, have argued that the *Sophist, Statesman, Philebus, Timaeus* (with its fragmentary sequel *Critias*), and *Laws* form a distinct linguistic group, belonging to the later years of Plato's life. The whole group must be later than the *Sophist,* which professes to be a sequel to the *Theaetetus.* Since the *Theaetetus* commemorates the death of the eminent mathematician after whom it is named (probably in 369 BC), it may be ascribed to *c.* 368, the eve of Plato's departure for Syracuse.

The earlier group of dialogues is generally believed to have ended with the *Theaetetus* and the closely related *Parmenides.* Apart from this, perhaps all that can be said with certainty is that the great dialogues, *Symposium, Phaedo,* and *Republic* (and perhaps also *Protagoras*), in which Plato's dramatic power was at its highest, mark the culmination of this first period of literary activity. The later dialogues are often thought to lack the dramatic and literary merits of the earlier but to compensate for this by an increased subtlety and maturity of judgment.

Persons of the dialogues. One difficulty that initially besets the modern student is that created by the dramatic form of Plato's writings. Since Plato never introduced

himself into his own dialogues, he is not formally committed to anything asserted in them. The speakers who are formally bound by the utterances of the dialogues are their characters, of whom Socrates is usually the protagonist. Since all of these are real historical persons, it is reasonable to wonder whether Plato is reporting their opinions or putting his own views into their mouths, and, more generally, to ask what was his purpose in writing dialogues.

Dialogue form

Some scholars have suggested that Plato allowed himself to develop freely in a dialogue any view that interested him for the moment without pledging himself to its truth. Thus Plato can make Socrates advocate hedonistic utilitarianism in the *Protagoras* and denounce it in the *Gorgias*. Others argue that some of Plato's characters, notably Socrates and Timaeus, are "mouthpieces" through whom he inculcates tenets of his own without concern for dramatic or historical propriety. Thus it has often been held that the theory of Ideas, the doctrine of recollection, and the notion of the tripartite soul were originated by Plato after the death of Socrates and consciously fathered on the older philosopher.

Thought of the earlier and later dialogues. There are undeniable differences in thought between the dialogues that are later than the *Theaetetus* and those that are earlier. But there are no serious discrepancies of doctrine between individual dialogues of the same period. Plato perhaps announced his own personal convictions on certain doctrines in the second group of dialogues by a striking dramatic device. In the *Sophist* and *Statesman* the leading part is taken by a visitor from Elea and in the *Laws* by an Athenian. These are the only anonymous, indeed almost certainly the only imaginary, personages of any moment in the whole of Plato's writings. It seems likely, therefore, that these two characters were left anonymous so that the writer could be free to use them as mouthpieces for his own teaching. Plato thus took on himself the responsibility for the logic and epistemology of the *Sophist* and of the *Statesman* and for the ethics and the educational and political theory of the *Statesman* and of the *Laws*.

Plato's changing views

Doctrine of Forms. There is a philosophical doctrine running through the earlier dialogues that has as its three main features the theory of knowledge as recollection, the conception of the tripartite soul, and, most importantly, the theory of Forms. The theory that knowledge is recollection rests on the belief that the soul is not only eternal but also preexistent. The conception of the tripartite soul holds that the soul consists of reason, appetite, and spirit (or will). Each part serves a purpose and has validity, but reason is the soul's noblest part; in order for man to achieve harmony, appetite and spirit must be subjected to the firm control of reason. The theory of Forms has as its foundation the assumption that beyond the world of physical things there is a higher, spiritual realm of Forms, or Ideas, such as the Form of Beauty or Justice. This realm of Forms, moreover, has a hierarchical order, the highest level being that of the Form of the Good. Whereas the physical world, perceived with the senses, is in constant flux and knowledge derived from it restricted and variable, the realm of Forms, apprehensible only by the mind, is eternal and changeless. Each Form is the pattern of a particular category of things in this world; thus there are Forms of man, stone, shape, colour, beauty, and justice. Yet the things of this world are only imperfect copies of these perfect Forms.

In the *Phaedo* Socrates is made to describe the theory of Ideas as something quite familiar that he has for years constantly canvassed with his friends. In the dialogues of the second period, however, these tenets are less prominent, and the most important of them all, the theory of Forms, is in the *Parmenides* subjected to a searching set of criticisms. The question thus arises as to whether Plato himself had two distinct philosophies, an earlier and a later, or whether the main object of the first group of dialogues was to preserve the memory of Socrates, the philosophy there expounded being, in the main, that of Socrates—coloured, no doubt, but not consciously distorted, in its passage through the mind of Plato. On the second view, Plato had no distinctive Platonic philosophy until a late period in his life.

Socrates and Plato. It may be significant that the only dialogue later than the *Theaetetus* in which Socrates takes a leading part is the *Philebus*, the one work of the second group that deals primarily with the ethical problems on which the thought of Socrates had concentrated. This is usually explained by supposing that Plato was unwilling to make Socrates the exponent of doctrines that he knew to be his own property. It would, however, be hard to understand such misgivings if Plato had already been employing Socrates in that very capacity for years. It is notable, too, that Aristotle, who apparently knew nothing of an earlier and a later version of Platonism, attributed to Plato a doctrine that is quite unlike anything to be found in the first group of dialogues. It was also the view of Neoplatonic scholars that the theory of Ideas of the great earlier dialogues really originated with Socrates; and the fact that they did not find it necessary to argue the point may show that this had been the standing tradition of the Academy.

Few modern scholars, however, support this view. The differences between the early and late periods are not as great as they have sometimes been represented: although Plato's thought developed from the early to the late dialogues, it underwent no sudden dislocation. The ideas of the early period may have been inspired by Socrates, but they were Plato's own—for example, the theory of Forms could not have arisen with Socrates. Plato nevertheless attributed it to him because he saw it as the theoretical basis of what Socrates did teach.

THE EARLIER DIALOGUES

Three main areas of thought

In the *Republic,* the greatest of all the dialogues that precede the *Theaetetus,* there are three main strands of argument deftly combined into an artistic whole—the ethical and political, the aesthetic and mystical, and the metaphysical. Other major dialogues belonging to this period give special prominence to one of these three lines of thought: the *Phaedo* to the metaphysical theme; the *Protagoras* and the *Gorgias* to the ethical and political; the *Symposium* and the *Phaedrus* to the aesthetic. But it should be noted that Plato's dialogues are not philosophical essays, let alone philosophical treatises, and they do not restrict themselves to a single topic or subject.

Dialogues of search. The shorter dialogues, dealing with more special problems, generally of an ethical character, mostly conform to a common type: a problem in moral philosophy, often that of the right definition of a virtue, is propounded, a number of tentative solutions are considered, and all are found to be vitiated by difficulties that cannot be dispelled. The reader is left, at the end of the conversation, aware of his ignorance of the very things that it is most imperative for a man to know. He has formally learned nothing but has been made alive to the confusions and fallacies in what he had hitherto been content to take as knowledge. The dialogues are "aporetic" and "elenctic": they pose puzzles (*aporiai* in Greek) without solving them, and Socrates' procedure consists in the successive refutation (*elenchos*) of the various views presented by his interlocutors.

The effect of these dialogues of search is thus to put the reader in tune with the spirit of Socrates, who had said that the one respect in which he was wiser than other men was in his keen appreciation of his own ignorance of the most important matters. The reader learns the meaning of Socrates' ruling principle that the supreme business of life is to "tend" the soul and his conviction that "goodness of soul" means knowledge of good and evil. The three dialogues directly concerned with the trial of Socrates have a further purpose. They are intended to explain to a puzzled public, as a debt of honour to his memory, why Socrates thought it a matter of conscience neither to withdraw from danger before his trial, nor to make a conciliatory defense, nor, after conviction, to avail himself of the opportunity of flight.

The *Apology,* or *Defense,* purports to give Socrates' speeches at his trial for impiety. In the *Crito* Socrates, in the condemned cell, explains why he will not try to escape paying the death penalty; the dialogue is a consideration of the source and nature of political obligation. The *Eu-*

thyphro is represented as taking place just before Socrates' trial. Its subject is the virtue of "piety," or the proper attitude for men to take toward the gods. The *Hippias Major* propounds the question "What is the 'fine' (or 'beautiful')?" The *Hippias Minor* deals with the paradox that "wrongdoing is involuntary." The *Ion* discredits the poets, who create not "by science" but by a nonrational inspiration. The *Menexenus,* which professes to repeat a funeral oration learned from Aspasia, Pericles' mistress, is apparently meant as a satire on the patriotic distortion of history. The *Charmides, Laches,* and *Lysis* are typical dialogues of search. The question of the *Charmides* is what is meant by *sōphrosunē,* or "temperance," the virtue that is shown in self-command, in dutiful behaviour to parents and superiors, in balance, and in self-possession amid the turns of fortune. It seems that this virtue can be identified with the self-knowledge that Socrates had valued so highly. The *Laches* is concerned with courage, the soldier's virtue; and the *Lysis* examines in the same tentative way friendship, the relation in which self-forgetting devotion most conspicuously displays itself.

Themes of the shorter dialogues

The question of whether words have meaning by nature or by convention is considered in the *Cratylus*—whether there is some special appropriateness of the sounds or forms of words to the objects they signify, or whether meaning merely reflects the usage of the community. Plato argues that, since language is an instrument of thought, the test of its rightness is not mere social usage but its genuine capacity to express thought accurately. The dialogue *Euthydemus* satirizes the "eristics"—those who try to entangle a person in fallacies because of the ambiguity of language. Its more serious purpose, however, is to contrast this futile logic chopping with the "protreptic," or hortatory, efforts of Socrates, who urges that happiness is guaranteed not by the possession of things but by the right use of them—and particularly of the gifts of mind, body, and fortune.

Ethical and political dialogues. The *Gorgias,* the *Protagoras,* and the *Meno,* like several of the lesser dialogues, give prominence to ethical and political themes. The *Gorgias* begins ostensibly as an inquiry into the nature and worth of rhetoric, the art of advocacy professed by Gorgias, and develops into a plea of sustained eloquence and logical power for morality—as against expediency—as the sovereign rule of life, both private and public. It ends with an imaginative picture of the eternal destinies of the righteous and of the unrighteous soul.

Gorgias holds that rhetoric is the queen of all "arts." If the statesman skilled in rhetoric is clever enough, he can, though a layman, carry the day even against the specialist. Socrates, on the other hand, declares that rhetoric is not an art but a mere "knack" of humouring the prejudices of an audience. There are two arts conducive to health of soul, those of the legislator and of the judge. The Sophist counterfeits the first, the orator the second, by taking the pleasant instead of the good as his standard. The orator is thus not the wise physician of the body politic but its toady. This severe judgment is disputed by Polus, an ardent admirer of Gorgias, on the ground that the successful orator is virtually the autocrat of the community, and to be such is the summit of human happiness because he can do whatever he likes.

Disputes as to the nature of rhetoric

Socrates rejects this view. He does so by developing one of the "Socratic paradoxes": to suffer a wrong is an evil, but to inflict one is much worse. Thus if rhetoric is of real service to men, it should be most of all serviceable to an offender, who would employ it to move the authorities to inflict the penalties for which the state of his soul calls. All of this is in turn denied by Callicles, who proceeds to develop the extreme position of an amoralist. It may be a convention of the herd that unscrupulous aggression is discreditable and wrong, but "nature's convention" is that the strong are justified in using their strength as they please, while the weak "go to the wall." To Socrates, however, the creators of the imperialistic Athenian democracy were no true statesmen: they were the domestic servants of the democracy for whose tastes they catered; they were not its physicians. That would be a condition like that of the Danaids of mythology, who are punished in Hades by being set to spend eternity in filling leaking pitchers. A happy life consists not in the constant gratification of boundless desires but rather in the measured satisfaction of wants that are tempered by justice and *sōphrosunē.*

The *Meno* is nominally concerned with the question of what virtue is and whether it can be taught. But it is further interesting for two reasons: it states clearly the doctrine that knowledge is "recollection"; and it introduces as a character the democratic politician Anytus, the main author of the prosecution of Socrates.

Whether virtue can be taught depends on what virtue is. But the inquiry into virtue is difficult—indeed, the very possibility of inquiry is threatened by Meno's paradox concerning the quest for knowledge. If a person is ignorant about the subject of his inquiry, he could not recognize the unknown, even if he found it. If, on the other hand, the person already knows it, inquiry is futile because it is idle to inquire into what one already knows. But this difficulty would vanish if the soul were immortal and had long ago learned all truth, so that it needs now only to be reminded of truths that it once knew and has forgotten. To advance this argument, Socrates shows that a slave boy who has never studied geometry can be brought to recognize mathematical truths. He produces the right answer "out of himself." In general, knowledge is "recollection." Socrates next produces the hypothesis that virtue is knowledge and infers that it is teachable. But if virtue is knowledge, there must be professional teachers of it. Anytus insists that the Sophists, who claim to be such professionals, are mischievous impostors; and even the "best men" have been unable to teach it to their own sons. The *Meno* ends with a distinction between knowledge and true belief, and with the suggestion that virtue comes not by teaching but by divine gift.

Whether virtue can be taught

The *Protagoras* gives the most complete presentation of the main principles of Socratic morality. In this dialogue Socrates meets the eminent Sophist Protagoras, who explains that his profession is the "teaching of goodness"—*i.e.,* the art of making a success of one's life and of one's city. Socrates urges, however, that both common opinion and the failure of eminent men to teach "goodness" to their sons suggest that the conduct of life is not teachable. But the problem arises as to whether the various commonly recognized virtues are really different or all one. Protagoras is ultimately ready to identify all of the virtues except courage with wisdom or sound judgment. Socrates then attempts to show that, even in the case of courage, goodness consists in the fact that, by facing pain and danger, one escapes worse pain or danger. Thus all virtues can be reduced to the prudent computation of pleasures and of pains. Here, then, is a second "Socratic paradox": no one does wrong willingly—wrongdoing is a matter of miscalculation. It is a puzzling feature of this argument that Socrates appears to embrace a form of hedonism.

Metaphysical foundation of Plato's doctrine: "Phaedo." In the works so far considered, the foundation of a Socratic moral and political doctrine is laid, which holds that the great concern of man is the development of a rational moral personality and that this development is the key to man's felicity. Success in this task, however, depends on rational insight into the true scale of good. The reason men forfeit felicity is that they mistake apparent good for real. If a man ever knew with assurance what the Good is, he would never pursue anything else; it is in this sense that "all virtue is knowledge." The philosophical moralist, who has achieved an assured insight into absolute Good, is thus the only true statesman, for he alone can tend to the national character. These moral convictions have a metaphysical foundation and justification. The principles of this metaphysics are expounded more explicitly in the following dialogues, in which a theory of knowledge and of scientific method is also discernible.

Need for insight into Good

The object of the *Phaedo* is to justify belief in the immortality of the soul by showing that it follows from a fundamental metaphysical doctrine (the theory of Ideas, or the doctrine of Forms), which seems to afford a rational clue to the structure of the universe. Socrates' soul is identical with Socrates himself: the survival of his soul is the survival of Socrates—in a purified state. For his life has

Arguments for immortality of the soul

been spent in trying to liberate the soul from dependence on the body. In life, the body is always interfering with the soul's activity. Its appetites and passions interrupt the pursuit of wisdom and goodness.

There are four arguments for thinking that the soul survives death.

First, there is a belief that the soul has a succession of many lives. The processes of nature in general are cyclical; and it is reasonable to suppose that this cyclicity applies to the case of dying and coming to life. If this were not so, if the process of dying were not reversible, life would ultimately vanish from the universe.

Second, the doctrine that what men call "learning" is really "recollection" shows, or at least suggests, that the soul's life is independent of the body.

Third, the soul contemplates the Forms, which are eternal, changeless, and simple. The soul is like the Forms. Hence it is immortal.

The fourth argument is the most elaborate. Socrates begins by recalling his early interest in finding the causes of being and change and his dissatisfaction with the explanations then current. He offers instead the Forms as causes. First, and safely, he says that something becomes, say, hot simply by participating in Heat. Then, a little more daringly, he is prepared to say that it becomes hot by participating in Fire, which brings Heat with it. Now if Fire brings Heat, it cannot accept Cold, which is the opposite of Heat. All this is then applied to the soul. Human beings are alive by participating in Life—and, more particularly, by having souls that bring Life with them. Since the soul brings Life, it cannot accept Death, the opposite of Life. But in that case the soul cannot perish and is immortal. (For further discussion of the theory of Forms, see METAPHYSICS.)

Aesthetic and mystical dialogues. Both the *Symposium* and the *Phaedrus* present the Forms in a special light, as objects of mystical contemplation and as stimuli of mystical emotion.

The immediate object of the *Symposium,* which records several banquet eulogies of *erōs* (erotic love), is to find the highest manifestation of the love that controls the world in the mystic aspiration after union with eternal and supercosmic beauty. It depicts Socrates as having reached the goal of union and puts the figure of Alcibiades, who has sold his spiritual birthright for the pleasures of the world, in sharp opposition to him.

Ascent to the supreme Form of Good

The main argument may be summarized thus: *Erōs* is a reaching out of the soul to a hoped-for good. The object is eternal beauty. In its crudest form, love for a beautiful person is really a passion to achieve immortality through offspring by that person. A more spiritual form is the aspiration to combine with a kindred soul to give birth to sound institutions and rules of life. Still more spiritual is the endeavour to enrich philosophy and science through noble dialogue. The insistent seeker may then suddenly descry a supreme beauty that is the cause and source of all of the beauties so far discerned. The philosopher's path thus culminates in a vision of the Form of the Good, the supreme Form that stands at the head of all others.

Though the immediate subject of the *Phaedrus* is to show how a truly scientific rhetoric might be built on the double foundation of logical method and scientific study of human passions, Plato contrives to unite with this topic a discussion of the psychology of love, which leads him to speak of the Forms as the objects of transcendent emotion and, indeed, of mystical contemplation. The soul, in its antenatal, disembodied state, could enjoy the direct contemplation of the Forms. But sense experience can suggest the Form of Beauty in an unusually startling way: through falling in love. The unreason and madness of the lover mean that the wings of his soul are beginning to grow again; it is the first step in the soul's return to its high estate.

The "Republic." In the *Republic* the immediate problem is ethical. What is justice? Can it be shown that justice benefits the man who is just? Plato holds that it can. Justice consists in a harmony that emerges when the various parts of a unit perform the function proper to them and abstain from interfering with the functions of any other part. More specifically, justice occurs with regard to the individual, when the three component parts of his soul—reason, appetite, and spirit, or will—each perform their appropriate tasks; with regard to society, justice occurs when its component members each fulfill the demands of their allotted roles. Harmony is ensured in the individual when the rational part of his soul is in command; with regard to society, when philosophers are its rulers because philosophers—Platonic philosophers—have a clear understanding of justice, based on their vision of the Form of the Good.

In the ethical scheme of the *Republic* three roles, or "three lives," are distinguished: those of the philosopher, of the votary of enjoyment, and of the man of action. The end of the first is wisdom; of the second, the gratification of appetite; and of the third, practical distinction. These reflect the three elements, or active principles, within a man: rational judgment of good; a multitude of conflicting appetites for particular gratifications; and spirit, or will, manifested as resentment against infringements both by others and by the individual's own appetites.

This tripartite scheme is then applied to determine the structure of the just society. Plato develops his plan for a just society by dividing the general population into three classes that correspond to the three parts of man's soul as well as to the three lives. Thus there are: the statesmen; the general civilian population that provides for material needs; and the executive force (army and police). These three orders correspond respectively to the rational, appetitive, and spirited elements. They have as their corresponding virtues wisdom, the excellence of the thinking part; temperance, that of the appetitive part (acquiescence of the nonrational elements to the plan of life prescribed by judgment); and courage, that of the spirited part (loyalty to the rule of life laid down by judgment). The division of the population into these three classes would not be made on the basis of birth or wealth but on the basis of education provided for by the state. By a process of examination each individual would then be assigned to his appropriate rank in correspondence with the predominant part of his soul.

The state ordered in this manner is just because each of the elements vigorously executes its own function and, in loyal contentment, confines itself within its limits. Such a society is a true aristocracy, or rule of the best. Plato describes successive deviations from this ideal as timocracy (the benign military state), oligarchy (the state dominated by merchant princes, a plutocracy), and democracy (the state subjected to an irresponsible or criminal will).

The training of the philosophical rulers would continue through a long and rigorous education because the vision of the Good requires extensive preparation and intellectual discipline. It leads through study of the exact sciences to that of their metaphysical principles. The central books of the *Republic* thus present an outline of metaphysics and a philosophy of the sciences. The Forms appear in the double character of objects of all genuine science and formal causes of events and processes. Plato expressly denied that there can be knowledge, in the proper sense, of the temporal and mutable. In his scheme for the intellectual training of the philosophical rulers, the exact sciences—arithmetic, plane and solid geometry, astronomy, and harmonics—would first be studied for 10 years to familiarize the mind with relations that can only be apprehended by thought. Five years would then be given to the still severer study of "dialectic." Dialectic is, etymologically, the art of conversation, of question and answer; and according to Plato, dialectical skill is the ability to pose and answer questions about the essences of things. The dialectician replaces hypotheses with secure knowledge, and his aim is to ground all science, all knowledge, on some "unhypothetical first principle."

The Form of the Good

This principle is the Form of the Good, which, like the Sun in relation to visible things, is the source of the reality of all things, of the light by which they are apprehended, and also of their value. As in the *Symposium,* the Good is the supreme beauty that dawns suddenly upon the pilgrim of love as he draws near to his goal.

Dialogues of critical reconstruction. The two works that

probably anticipate the dialogues of Plato's old age, the *Parmenides* and *Theaetetus,* display a remarkable difference of tone, clearly the result of a period of fruitful reconstruction.

The theory expounded in the *Phaedo* and *Republic* does not allow enough reality to the sensible world. These dialogues suppose that an entity capable of being sensed is a complex that participates in a plurality of Forms: what else it may be they do not say. Clearly, however, the relation between a thing and a Form (*e.g.,* beauty), which has been called participation, needs further elucidation. In these dialogues truths of fact, of the natural world, have not yet had their importance recognized.

Plato clearly had an external motive for the reexamination of his system as well. The *Parmenides,* the *Theaetetus,* and the *Sophist* all reveal a special interest in the Eleatic philosophy, of which Parmenides was the chief representative. The doctrine of his friend Eucleides of Megara, like that of Parmenides, was that phenomena which can be apprehended by the senses are illusions with no reality at all. Continued reflection on this problem led straight to the discussion of the meaning of the copula "is" and the significance of the denial "is not," which is the subject of the *Sophist.*

Formally the *Parmenides* leads to an impasse. In its first half the youthful Socrates expounds the doctrine of the participation of things in Forms as the solution of the problem of the "one and many." ("How can this, that, and the other cat all be one thing—*e.g.,* black?" "Each distinct cat participates in the unique Form of Blackness.") Parmenides raises what appear to be insoluble objections and hints that the helplessness of Socrates under his criticism arises from insufficient training in logic.

In the second half Parmenides gives an example of the logical training that he recommends. He takes for examination his own thesis, "The one is," and constructs upon it as basis an elaborate set of contradictions.

Objections to the doctrine of Forms

The Eleatic objections to the doctrine of participation are, first, that it does not really reconcile unity with plurality since it leads to a perpetual regress. It says that the many things that have a common predicate, or characteristic, participate in, or imitate, a single Form. But the Form itself also admits of a common predicate, and therefore a second Form must exist, participated in alike by the sensible things and the first Form, and so on, endlessly. Second, a graver difficulty is that the relations between Forms must belong to the realm of Forms, and those between sensible things to the realm of things. Thus men, belonging to the second, can know nothing of the true realities, the Forms. Scholars disagree over the precise interpretation of these objections. They also disagree about how Plato should have reacted to them—and about how he did react.

The *Theaetetus* is a discussion of the question of how knowledge should be defined. It is remarkable that the dialogue treats knowledge at length without making any reference to the Forms or to the mythology of recollection. It remains to this day one of the best introductions to the problem of knowledge. The main argument is as follows:

Efforts at defining knowledge

It seems plausible to say that knowledge is perception, which appears to imply that "What seems to me is so to me; what seems to you is so to you" (Protagoras). This relativistic doctrine is, rather oddly, claimed by Plato to be equivalent to the view held by the late 6th-century-BC Greek philosopher Heracleitus that "everything is always and in all ways in flux." But these views imply that there is no common perceived world and therefore nothing of certainty can be said or thought at all.

As for the thesis that knowledge is perception, one must first distinguish what the soul perceives through bodily organs from what it apprehends by itself without organs—such as number, sameness, likeness, being, and good. But because all knowledge involves truth and therefore being, perception, which cannot grasp being, is not identical with knowledge.

Is knowledge, then, true belief? The reference to true belief leads Plato into a discussion of false belief, for which he can discover no satisfactory analysis. False belief is belief in what is not, and what is not cannot be believed. But the example of verdicts in the law courts is enough to show that there can be true belief without knowledge.

Finally, is knowledge true belief together with an "account"? The concept of an account (*logos*) is not a simple one. No satisfactory definition of knowledge emerges, and the dialogue ends without a conclusion.

Because Plato's argument nowhere appeals to his favourite doctrine of Forms and because the dialogue ends so inconclusively, some scholars have suggested that Plato wanted to show that the problem of knowledge is insoluble without the Forms.

THE LATER DIALOGUES

Formally the important dialogues the *Sophist* and the *Statesman* are closely connected, both being ostensibly concerned with a problem of definition. The real purpose of the *Sophist,* however, is logical or metaphysical; it aims at explaining the true nature of negative predication, or denials that something is so. The object of the *Statesman,* on the other hand, is to consider the respective merits of two contrasting forms of government, personal rule and constitutionalism, and to recommend the second, particularly in the form of limited monarchy. The *Sophist* thus lays the foundations of all subsequent logic, the *Statesman* those of all constitutionalism. A second purpose in both dialogues is to illustrate the value of careful classification as a basis for scientific definition.

The *Sophist* purports to investigate what a Sophist really is. The definitions all lead to such notions as falsity, illusion, nonbeing. But these notions are puzzling. How can there be such a thing as a false statement or a false impression? For the false means "what is not," and what is not is nothing at all and can neither be uttered nor thought. Plato argues that what is not in some sense also is, and that what is in some sense is not; and he refutes Parmenidean monism by drawing the distinction between absolute and relative nonbeing. A significant denial, *A* is not *B,* does not mean that *A* is nothing, but that *A* is other than *B;* every one of the "greatest kinds," or most general, features of reality—being, identity, difference, motion, and rest—is other than every other feature. Motion, say, is other than rest; and thus motion is not rest—but it does not follow that motion is not. The true business of dialectic is to treat the Forms themselves as an interrelated system, with relations of compatibility and incompatibility among themselves.

Absolute versus relative nonbeing

In the *Statesman* the conclusion is reached that government by a benevolent dictator is not suitable to the conditions of human life because his direction is not that of a god. The surrogate for direction by a god is the impersonal supremacy of inviolable law. Where there is such law, monarchy is the best and democracy the least satisfactory form of constitution; but where there is no law, this situation is inverted.

The *Philebus* contains Plato's ripest moral psychology. Its subject is strictly ethical—the question of whether the Good is to be identified with pleasure or with wisdom. Under the guidance of Socrates a mediating conclusion is reached: the best life contains both elements, but wisdom predominates.

Philosophically most important is a classification adopted to determine the formal character of the two claimants to recognition as the Good. Everything real belongs to one of four classes: (1) the infinite or unbounded, (2) the limit, (3) the mixture (of infinite and limit), (4) the cause of the mixture. It emerges that all of the good things of life belong to the third class, that is, are produced by imposing a definite limit upon an indeterminate continuum.

The *Timaeus* is an exposition of cosmology, physics, and biology. Timaeus first draws the distinction between eternal being and temporal becoming and insists that it is only of the former that one can have exact and final knowledge. The visible, mutable world had a beginning; it is the work of God, who had its Forms before him as eternal models in terms of which he molded the world as an imitation. God first formed its soul out of three constituents: identity, difference, being. The world soul was placed in the circles of the heavenly bodies, and the circles were animated with movements. Subsequently the various

The forming of the world

subordinate gods and the immortal and rational element in the human soul were formed. The human body and the lower components of its soul were generated through the intermediacy of the "created gods" (*i.e.,* the stars).

The *Timaeus* combines the geometry of the Pythagoreans with the biology of Empedocles by a mathematical construction of the elements, in which four of the regular solids—cube, tetrahedron, octahedron, and icosahedron—are assumed to be the shapes of the corpuscles of earth, fire, air, and water. (The fifth, the dodecahedron, comprises the model for the whole universe.)

Among the important features of the dialogue are its introduction of God as the "demiurge"—the intelligent cause of all order and structure in the world of becoming—and the emphatic recognition of the essentially tentative character of natural science. It is also noteworthy that, though Plato presents a corpuscular physics, his metaphysical substrate is not matter but *chôra* (space). The presence of space as a factor requires the recognition, over and above God or mind, of an element that he called *anankê* (necessity). The activity of the demiurge ensures that the universe is in general rational and well-ordered, but the brute force of material necessity sets limits to the scope and efficacy of reason. The details of Plato's cosmology, physiology, and psychophysics are of great importance for the history of science but metaphysically of secondary interest.

The practical level of the *Laws*

The *Laws,* Plato's longest and most intensely practical work, contains his ripest utterances on ethics, education, and jurisprudence, as well as his one entirely nonmythical exposition of theology. The immediate object is to provide a model of constitution making and legislation to assist in the actual founding of cities. The problem of the dialogue is thus not the construction of an ideal state as in the *Republic* but the framing of a constitution and code that might be successfully adopted by a society of average Greeks. Hence the demands made on average human nature, though exacting, are not pitched too high; and the communism of the *Republic* is dropped.

Purely speculative philosophy and science are excluded from the purview of the *Laws,* and the metaphysical interest is introduced only so far as to provide a basis for a moral theology. In compensation the dialogue is exceptionally rich in political and legal thought and appears, indirectly, to have left its mark on the great system of Roman jurisprudence.

In the ethics of the *Laws* Plato is rigid and rigorous—for example, homosexuality shall be completely suppressed and monogamous marriage with strict chastity shall be the rule. (In the *Republic* the guardian class enters into temporary unions or "sacred marriages," with a community of wives and children, to foster a concern for the common good.) In politics, Plato favours a mixed constitution, one with elements of democratic freedom and autocratic authoritarianism, and he suggests a system for securing both genuine popular representation and the proper degree of attention to personal qualifications. The basis of society is to be agriculture, not commerce. What amounts to a tax of 100 percent is to be levied on incomes beyond the statutory limits. Education is regarded as the most important of all the functions of government. The distinction between the sexes is to be treated as irrelevant.

Careful attention is to be paid to the right utilization of the child's instinct for play and to the demand that the young shall be taught in institutions where expert instruction in all of the various subjects is coordinated. Members of the supreme council of the state shall be thoroughly trained in the supreme science, which "sees the one in the many and the many in the one"; *i.e.,* in dialectic. In the *Laws* Plato instituted regulations which would ensure that trials for serious offenses would take place before a court of highly qualified magistrates and would proceed with due deliberation. Also, provision was made for appeals, and a foundation was laid for a distinction between civil and criminal law.

Plato's natural theology

The *Laws* also creates a natural theology. There are three false beliefs, Plato holds, that are fatal to moral character: atheism, denial of the moral government of the world, and the belief that divine judgment can be bought off by offerings. Plato claims that he can disprove them all. His refutation of atheism turns on the identification of the soul with the "movement which can move itself." Thus all motion throughout the universe is ultimately initiated by souls. It is then inferred from the regular character of the great cosmic motions and their systematic unity that the souls which originate them form a hierarchy with a best soul, God, at their head. Since some motions are disorderly, there must be one soul that is not the best, and there may be more. (There is no suggestion, however, that there is a worst soul, a devil.) The other two heresies can be similarly disposed of. Plato thus becomes the originator of the view that there are certain theological truths that can be strictly demonstrated by reason; *i.e.,* of philosophical theology. Plato goes on to enact that the denial of any of his three propositions shall be a grave crime.

The *Laws* strikes many readers as a dull and depressing work. Its prose lacks the sparkle of the early dialogues; and Socrates, the hero of those works, would not have been tolerated under a government of the repressively authoritarian style that the *Laws* recommends. (J.B./Ed.)

Platonism after Plato

The term Platonism can be applied to any philosophy that derives its ultimate inspiration from Plato. Though there was in antiquity a tradition about Plato's "unwritten doctrines" (much discussed by German scholars since 1959), Platonism then and later was based primarily on a reading of the dialogues. But these can be read in many different ways, often very selectively, and it may be that all that the various kinds of Platonism can be said to have in common is an intense concern for the quality of human life—always ethical, often religious, and sometimes political, based on a belief in unchanging and eternal realities, independent of the changing things of the world perceived by the senses. Platonism sees these realities both as the causes of the existence of everything in the universe and as giving value and meaning to its contents in general and the life of its inhabitants in particular. It is this belief in absolute values rooted in an eternal world that distinguishes Platonism from the philosophies of Plato's immediate predecessors and successors and from later philosophies inspired by them—from the immanentist naturalism of most of the pre-Socratics (who interpreted the world monistically in terms of nature as such), from the relativism of the Sophists, and from the correction of Platonism in a this-worldly direction carried out by Plato's greatest pupil, Aristotle.

GREEK PLATONISM FROM ARISTOTLE THROUGH MIDDLE PLATONISM: ITS NATURE AND HISTORY

Plato as construed by Aristotle

Since Plato refused to write his own metaphysics, knowledge of its final shape has to be derived from hints in the dialogues and statements by Aristotle and, to a far lesser extent, other ancient authorities. According to these, Plato's doctrine of Forms was, in its general character, highly mathematical, the Forms being somehow identified with, or explained in terms of, numbers. Here may be seen the influence of the Pythagoreans, though, as Aristotle says, the details of Plato's views on the mathematical constituents of being were not the same as theirs. In addition Aristotle states that Plato introduced a class of "mathematicals," or "intermediates," positioned between sensible objects and Forms. These differ from sensible objects in being immaterial (*e.g.,* the geometer's triangles *ABC* and *XYZ*) and from the Forms in being plural, unlike the Triangle itself. (H.J.Bl./Ed.)

Aristotle himself had little use for this sort of mathematical metaphysics and rejected Plato's doctrine of transcendent eternal Forms altogether. Something of Platonism, nonetheless, survived in Aristotle's system in his beliefs that the reality of anything lay in a changeless (though wholly immanent) form or essence comprehensible and definable by reason and that the highest realities were eternal, immaterial, changeless self-sufficient intellects which caused the ordered movement of the universe. It was the desire to give expression to their transcendent perfection that kept the heavenly spheres rotating. Man's intellect at

its highest was akin to them. This Aristotelian doctrine of Intellect (nous) was easily recombined with Platonism in later antiquity.

The Academy through the Hellenistic Age

Aristotle, however, was not reacting only against Plato but also against Plato's associates and immediate successors as head of the Academy, namely Plato's nephew Speusippus (*c.* 410–339 BC) and Xenocrates (396–314 BC). Speusippus, in particular, accented the mathematical tendencies of the late Plato and abolished Forms in favour of numbers. He also posited different principles for different sorts of entities and so was accused by Aristotle of breaking the connections in reality. Xenocrates identified Forms and numbers and began the long process of finding firm doctrines in Plato by laying down that Forms were only of those things that exist in nature. Xenocrates was also the first, as far as is known, to turn his attention to what continued to be a subject of controversy throughout the history of Platonism, namely whether the account of creation offered in the *Timaeus* was to be taken as chronological or merely expository. He took the latter view, which turned out to be the most favoured one in antiquity; Aristotle was on the other side. Whether Xenocrates' three successors as head of the Academy (Polemon, Crates, and Crantor) developed Platonism is uncertain. Crantor (*c.* 330–270 BC) was allegedly the first to write commentaries on Plato, particularly on the *Timaeus.* After Crantor the Academy was preoccupied for about two centuries with the serious questioning of man's claims to knowledge. This began with Arcesilaus (316/315–*c.* 241 BC), who is described as the founder of the Middle Academy. There was a genuine desire to recover the critical, questioning, and agnostic attitude of the Socrates of Plato's early dialogues as well as philosophical exasperation with the dogmatism of some of the contemporary Hellenistic philosophers, especially the Stoics. It is likely that Arcesilaus was influenced to some extent by Pyrrhon (*c.* 360–*c.* 272 BC), founder of the tradition to which the name Skeptic was applied in antiquity. The Skeptical Academics denied that certainty on any subject was possible and worked out a sophisticated theory of probability as a guide to practical decision making. Their critical dialectic and probability theory were best expounded by Carneades (214/213–129/128 BC). Though he wrote nothing, he was regarded as the founder of the New Academy. A return to dogmatic and positive philosophical teaching was effected by Philo of Larissa (died *c.* 79 BC) and his pupil Antiochus of Ascalon, who was head of the school in 79–78 BC.

Transition to Neo-platonism

The next important phase of Platonism, Middle Platonism or pre-Neoplatonism, was significant through the influence that it exerted in more than one direction. In the direction of Jewish culture (further described in a later section), it formed the Greek philosophical background of the efforts of Philo Judaeus (Philo of Alexandria) to create a philosophical system on the basis of the Old Testament heritage. Though the origins of Middle Platonism are obscure, its main direction became clear in the 1st century AD. It seems to have been linked from the beginning with the closely related revival of Pythagoreanism (a philosophy holding that reality is number, and sometimes showing, after the revival, a tendency to superstitious occultism). The somewhat Platonized Stoicism of Poseidonius (*c.* 135–*c.* 51 BC), whose dualism of matter and reason enhanced the roles of emotion and will, may have influenced its beginnings, as did the Stoicized Platonism of Antiochus; and Stoic influence, especially in the ethical field, remained important in its later developments. There was also a strong Aristotelian influence, though a minority of 2nd-century Platonists, notably Atticus and, to a lesser extent, Gaius Calvenus Taurus, objected to certain Aristotelian doctrines. Atticus was particularly offended by Aristotle's failure to provide for providence. The general characteristics of this revised Platonic philosophy (and the closely related Neo-Pythagoreanism) were the recognition of a hierarchy of divine principles with stress on the transcendence of the supreme principle, which was already occasionally called "the One"; the placing of the Platonic Forms in the divine mind; a strongly otherworldly attitude demanding a "flight from the body," an ascent of the mind to the divine and eternal; and a preoccupation with the problem of evil, attributed either to an evil world soul or to matter. The best known of the Middle Platonists is the biographer and essayist Plutarch of Chaeronea (*c.* AD 46–120). More important philosophically were other 2nd-century figures: Gaius and two men possibly influenced by him, Albinus and Apuleius (better known as author of the prose narrative *The Golden Ass*); Atticus; and Numenius of Apamea. It was from the thought of these and other Middle Platonists, combined with his own reading of Alexander and other Peripatetic commentators on Aristotle, that the foremost Neoplatonist, Plotinus, started constructing his own interpretation of Platonism, which was both profoundly original and firmly rooted in an established school tradition.

NEOPLATONISM: ITS NATURE AND HISTORY

Neoplatonism is the modern name given to the form of Platonism developed by Plotinus in the 3rd century AD and modified by his successors. It came to dominate the Greek philosophical schools and remained predominant until the teaching of philosophy by pagans ended in the second half of the 6th century AD. It represents the final form of pagan Greek philosophy. It was not a mere syncretism (or combination of diverse beliefs) but a genuine, if one-sided, development of ideas to be found in Plato and earlier Platonism—though it incorporated important Aristotelian and Stoic elements as well. There is no real evidence for Oriental influence. A certain Gnostic (relating to intuitive knowledge acquired by privileged individuals and immune to empirical verification) tone or colouring sometimes may be discerned in the thought of Plotinus. But he was consciously a passionate opponent of Gnosticism, and in any case there was often a large element of popular Platonism in the Gnostic systems then current. Moreover, the theosophical works of the late 2nd century AD known as the *Chaldean Oracles,* which were taken as inspired authorities by the later Neoplatonists, seem to have been a hodgepodge of popular Greek religious philosophy.

Core concepts of Neo-platonism

Neoplatonism began as a complex (and in some ways ambiguous) philosophy and grew vigorously in a variety of forms over a long period; it is therefore not easy to generalize about it. But the leading ideas in the thought of philosophers who can properly be described as Neoplatonists seem always to have included the following:

1. There is a plurality of levels of being, arranged in hierarchical descending order, the last and lowest comprising the physical universe, which exists in time and space and is perceptible to the senses.

2. Each level of being is derived from its superior, a derivation that is not a process in time or space.

3. Each derived being is established in its own reality by turning back toward its superior in a movement of contemplative desire, which is implicit in the original creative impulse of outgoing that it receives from its superior; thus the Neoplatonic universe is characterized by a double movement of outgoing and return.

4. Each level of being is an image or expression on a lower level of the one above it. The relation of archetype and image runs through all Neoplatonic schemes.

5. Degrees of being are also degrees of unity; as one goes down the scale of being there is greater multiplicity, more separateness, and increasing limitation—until the atomic individualization of the spatiotemporal world is reached.

6. The highest level of being, and through it all of what in any sense exists, derives from the ultimate principle, which is absolutely free from determinations and limitations and utterly transcends any conceivable reality, so that it may be said to be "beyond being." Because it has no limitations, it has no division, attributes, or qualifications; it cannot really be named, or even properly described as being, but may be called "the One" to designate its complete simplicity. It may also be called "the Good" as the source of all perfections and the ultimate goal of return, for the impulse of outgoing and return that constitutes the hierarchy of derived reality comes from and leads back to the Good.

7. Since this supreme principle is absolutely simple and undetermined (or devoid of specific traits), man's knowl-

edge of it must be radically different from any other kind of knowledge. It is not an object (a separate, determined, limited thing) and no predicates can be applied to it; hence it can be known only if it raises the mind to an immediate union with itself, which cannot be imagined or described.

Plotinus and his philosophy. As far as is known, the originator of this distinctive kind of Platonism was Plotinus (AD 205–270). He had been the pupil at Alexandria of a self-taught philosopher called Ammonius, who also taught the Christian Origen and the latter's pagan namesake, and whose influence on his pupils seems to have been deep and lasting. But Ammonius wrote nothing; there are few reports of his views, and these are unreliable so that nothing is actually known about his thought. A number of distinguished scholars have made attempts to reconstruct it, but their speculations go far beyond the evidence. Plotinus must thus be regarded as the first Neoplatonist, and his collected works, the *Enneads* (Greek *enneas*, "set of nine"—six sets of nine treatises each, arranged by his disciple Porphyry), are the first and greatest collection of Neoplatonic writings.

Plotinus, like most ancient philosophers from Socrates on, was a religious and moral teacher as well as a professional philosopher engaged in the critical interpretation of a long and complicated school tradition. He was an acute critic and arguer, with an exceptional degree of intellectual honesty for his, or any, period; philosophy for him was not only a matter of abstract speculation but also a way of life in which, through an exacting intellectual and moral self-discipline and purification, those who are capable of the ascent can return to the source from which they came. His written works explain how from the eternal creative act—at once spontaneous and necessary—of that transcendent source, the One, or Good, proceeds the world of living reality, constituted by repeated double movements of outgoing and return in contemplation; and this account, showing the way for the human self—which can experience and be active on every level of being—to return to the One, is at the same time an exhortation to follow that way.

Metaphysics: system of emanation

Plotinus always insisted that the One, or Good, is beyond the reach of thought or language; what he said about this supreme principle was intended only to point the mind along the way to it, not to describe or define it. But though no adequate concept or definition of the Good is possible, it was, nonetheless, for Plotinus a positive reality of superabundant excellence. Plotinus often spoke of it in extremely negative language, but his object in doing so was to stress the inadequacy of all of man's ways of thinking and speaking to express this supreme reality or to clarify the implications of the claim that the Good is absolutely one and undetermined, the source of all defined and limited realities.

The original creative or expressive act of the One is the first great derived reality, nous (which can be only rather inadequately translated as "Intellect" or "Spirit"); from this again comes Soul, which forms, orders, and maintains in being the material universe. It must be remembered that, to Plotinus, the whole process of generation is timeless; Nous and Soul are eternal, while time is the life of Soul as active in the physical world, and there never was a time when the material universe did not exist. The "levels of being," then, though distinct, are not separate but are all intimately present everywhere and in everyone. To ascend from Soul through Intellect to the One is not to travel in space but to awake to a new kind of awareness.

Intellect, Soul, world

Intellect for Plotinus is at one and the same time thinker, thought, and object of thought; it is a mind that is perfectly one with its object. As object, it is the world of Forms, or Ideas, the totality of real being in the Platonic sense. These Forms, being one with Intellect and therefore with each other, are not merely objects but are living, thinking subjects, each not only itself but, in its contemplation, the whole. They are the archetypes and causes of the necessarily imperfect realities on lower levels, souls and the patterns or structures that make bodies what they are. Men at their highest are intellects, or souls perfectly conformed to Intellect; they become aware of their intellectual nature when, passing not only beyond sense perception but beyond the discursive reasoning characteristic of the life of Soul, they immediately grasp eternal realities.

Soul for Plotinus is very much what it was for Plato, the intermediary between the worlds of Intellect and Sense and the representative of the former in the latter. It is produced by Intellect, as Intellect is by the One, by a double movement of outgoing and return in contemplation, but the relationship between the two is more intimate and the frontier less clearly defined. For Plotinus, as for Plato, the characteristic of the life of the Soul is movement, which is the cause of all other movements. The life of the Soul in this movement is time, and on it all physical movement depends. Soul both forms and rules the material universe from above; and in its lower, immanent phase, which Plotinus often calls nature, it acts as an indwelling principle of life and growth and produces the lowest forms, those of bodies. Below these lies the darkness of matter, the final absence of being, the absolute limit at which the expansion of the universe—from the One through diminishing degrees of reality and increasing degrees of multiplicity—comes to an end. Because of its utter negativity, such matter is for Plotinus the principle of evil; and although he does not really believe it to be an independent principle forming, with the Good, a dualism, his language about it often has a strongly dualistic flavour.

He was not, however, really dualistic in his attitude toward the material universe. He strongly maintained its goodness and beauty as the best possible work of Soul. It is a living organic whole, and its wholeness is the best possible (though very imperfect) reflection on the space–time level of the living unity in diversity of the world of Forms in Intellect. It is held together in every part by a universal sympathy and harmony. In this harmony external evil and suffering take their place as necessary elements in the great pattern, the great dance of the universe. Evil and suffering can affect men's lower selves but can only exceptionally, in the thoroughly depraved, touch their true, higher selves and so cannot interfere with the real well-being of the philosopher.

Ethics and religion: purification

As souls within bodies, men can exist on any level of the soul's experience and activity. (The descent of souls into bodies is for Plotinus—who had some difficulty in reconciling Plato's various statements on this point—both a fall and a necessary compliance with universal law.) Man can ascend through his own intellect to the level of universal Soul, become that whole that he already is potentially, and, in Soul, attain to Intellect itself; or he can isolate himself on the lower level, shutting himself up in the experiences, desires, and concerns of his lower nature. Philosophical conversion—the beginning of the ascent to the One—consists precisely in turning away, by a tremendous intellectual and moral effort, from the life of the body, dominating and rising above its desires, and "waking to another way of seeing, which everyone has but few use." This, Plotinus insisted, is possible while one is still in an earthly body and without neglecting the duties of one's embodied state. But the body and bodily life weight a man down and hamper him in his ascent. Plotinus' language when speaking of the body and the senses in this context is strongly dualistic and otherworldly. Platonists in general think much more dualistically about their own bodies than about the material universe as a whole. The physical world is seen positively as a noble image of the intelligible; the individual, earthly, animal body, on the contrary, tends to be regarded negatively as a hindrance to the intellectual and spiritual life.

When a man's philosophical conversion is complete and he has become Intellect, he can rise to that mystical union in which the One manifests his continual presence, carried on the surging current of the impulse of return to the source (in its strongest and final flow), the pure love of Intellect for the Good from which it immediately springs. There is no consciousness of duality in that union; the individual is not aware of himself; but neither is he destroyed or dissolved into the One—because even in the union he is still Intellect, though Intellect "out of itself," transcending its normal nature and activity. This mystical union for Plotinus was the focus of much of his effort and, for those of similar inclination, the source of the

continuing power of his teaching. Philosophy for him was religion, the effort to actualize in oneself the great impulse of return to the Good, which constitutes reality on all its levels; and religion for him was philosophy. There was no room in his thought and practice for special revelation, grace, and repentance in the Christian sense, and little for external rites or ceremonies. For him the combination of moral purification and intellectual enlightenment, which only Platonic philosophy as he understood it could give, was the only way to union with the Good.

Porphyry and Iamblichus

The later Neoplatonists. Porphyry (*c.* AD 234–*c.* 305), a devout disciple of Plotinus and a careful editor of his works, occupied a special position in the development of later Neoplatonism. In some ways his thought paralleled that of the later pagan Neoplatonists, but in others it quite opposed them. The most distinctive features of his thought seem to have been an extreme spiritualism, an insistence, even sharper than that of Plotinus, on the "flight from the body" and—more philosophically important—a greater sympathy with the less sharply defined vertical hierarchies of the Platonists who had preceded Plotinus. Porphyry did not always clearly distinguish the One from Intellect. On the other hand one may see in him the beginnings of the late Neoplatonic tendency to structure reality in both vertical and "horizontal" triads. Thus Being, Life, and Intellect are phases in the eternal self-determination of the ultimate reality. This triad became one of the most important elements in the complex metaphysical structures of the later Neoplatonists. But perhaps Porphyry's most important and influential contribution was the incorporation into Neoplatonism of Aristotle's logic, in particular the doctrine of the categories, with the characteristic Neoplatonic interpretation of them as terms signifying entities. Also of interest is his declaration of ideological war against the Christians, whose doctrines he attacked on both philosophical and exegetical grounds in a work of 15 books entitled *Against the Christians.*

Iamblichus (*c.* AD 250–*c.* 330) seems to have been the originator of the type of Neoplatonism that came to dominate the Platonic schools in the 5th and 6th centuries AD. This kind of Neoplatonism sharpened and multiplied the distinctions between the levels of being. The basic position underlying its elaborations is one of extreme philosophical Realism: it is assumed that the structure of reality corresponds so exactly to the way in which the mind works that there is a separate real entity corresponding to every distinction that it can make. In the fully developed late Neoplatonic system the first principle of reality, the ultimate One, was removed to an altogether ineffable transcendence, mitigated by two factors: the presence of the expressions or manifestations of its unifying power, the "henads"—identified with the gods of paganism—at every level of reality; and the possibility of return to absolute unification through the henad with which one is linked. Below the One a vast structure of triads, or trinities, reached down to the physical world; this was constructed by combining Plotinus' vertical succession of the levels of Being, Intellect, and Soul (much complicated by internal subdivision and the interposition at every stage of mediating hypostases, or underlying orders of nonmaterial reality) with another horizontal triadic structure, giving a timeless dynamic rhythm of outgoing and return, such as that already encountered in Porphyry.

Nearly all of Iamblichus' works have been lost, and his thought must be recovered from other sources. At present the main authority for this type of Platonism, and also for some of the later Neoplatonists, is Proclus (AD 410–485). Proclus appears to have codified later Platonism, but it is often impossible to tell which parts of his thought are original and which derive from his teachers Plutarch (see below) and Syrianus on the one hand and Porphyry and Iamblichus, from whom he quotes copiously but not always identifiably, and other earlier Platonists on the other hand. A carefully argued summary of the basic metaphysics of this kind of Neoplatonism may be found in Proclus' *Elements of Theology,* which exhibits the causal relationships of the several hierarchies that constituted his intelligible universe.

This later Neoplatonism aspired to be not only a complete and coherent metaphysical system but also a complete pagan theology, which is perhaps best seen in Proclus' *Platonic Theology.* The maintenance and defense of the old religion in a world more and more intolerantly dominated by its triumphant rival, Christianity, was one of the main concerns of the Platonists after Plotinus. By the study and sometimes forced exegesis of Aristotle and then Plato, culminating in the *Timaeus* and *Parmenides,* of which they offered a variety of highly metaphysical interpretations totally unacceptable to Plato scholars, they believed it possible to arrive at a complete understanding of divine truth. This truth they held to be cryptically revealed by the gods themselves through the so-called theologians—the inspired authors of the Orphic poems and of the *Chaldean Oracles,* published in the second half of the 2nd century AD. Porphyry first gave some guarded and qualified recognition to them, but they were inspired scripture to Iamblichus, who wrote a work of at least 28 books on the subject, and his successors. Their view of the human soul was a humbler one than that of Plotinus. It was for them a spiritual being of lower rank, which had descended altogether into the material world, while for Plotinus a part remained above; they could not therefore aspire, like Plotinus, through philosophy alone, to that return to and unification with the divine that remained for them the goal of human life. Help from the gods was needed, and they believed that the gods in their love for men had provided it, giving to all things the power of return in prayer and implanting even in inanimate material things—herbs and stones and the like—sympathies and communications with the divine, which made possible the secret rites of theurgy, through which the divine gave the needed spiritual help by material means. Theurgy, though its procedures were generally those of late Greek magic, was thus not thought of merely as magic; in fact a higher and more intellectual theurgy was also practiced. The degree of attention paid to external rites varied considerably from philosopher to philosopher; there seem to have been men even in the last generation of pagan Neoplatonists who had little use for or interest in such things and followed a mystical way much like that of Plotinus.

Neoplatonism as a pagan theology

Neoplatonic schools

The different schools of late Neoplatonism seem to have differed less from each other than has sometimes been supposed. The school of Pergamum, founded by Aedesius, a pupil of Iamblichus, made perhaps the least contribution to the philosophical development of Neoplatonism, but it was not entirely given over to theurgy. Its greatest convert was the emperor Julian, called by Christians the "Apostate"; in that capacity he achieved great notoriety, but philosophically he is of no importance. By the end of the 4th century AD the Platonic Academy at Athens had been reestablished and had become an institute for Neoplatonic teaching and research following the tradition of Iamblichus. It was particularly fervent and open in its paganism and attracted Christian hostility. Though maintaining itself for a surprisingly long time against this hostility, it eventually yielded to it and was probably closed by Justinian in AD 529. In the interim, however, it had produced the greatest and most influential systematic expositor of later Neoplatonism, Proclus (see above). The head of the school at the time of its closing, Damascius, was also a notable philosopher. Another centre of Neoplatonism flourished at Gaza during the 5th and early 6th centuries; it was already Christian in its inspiration, though some of its members studied with the pagan Ammonius (see below). The school of Alexandria in the 5th and 6th centuries does not seem to have differed very much from that of Athens, either in its fundamental philosophical outlook or in the main outline of its doctrines. In fact there was much interchange between the two. The Athenian Syrianus taught the Alexandrian Hermias, whose son Ammonius was taught by Proclus. Ammonius (died *c.* 520) was the most influential of the Alexandrian Platonists. His expositions of Aristotle were published mainly in the commentaries of the Christian heretic John Philoponus (late 5th to mid-6th century). Simplicius, the other great Aristotelian commentator, worked at Athens but, like Damascius, had studied with Ammonius. The Alexandrian concentration on Aristotle, which produced

a vast body of learned but Neoplatonically coloured commentary on his treatises, has often been attributed to Christian pressure and attempts to compromise with the church; it may equally well have been due to the quality and extent of Proclus' published work on Plato. Though Philoponus' later philosophical work contains important Christian modifications, an openly pagan (and very inferior) philosopher, Olympiodorus, was still teaching at Alexandria well into the second half of the 6th century. Finally, in the 7th century, under Heraclius, after philosophical teaching had passed peacefully into Christian hands, the last known Alexandrian philosopher, the Christian Stephanus, was called to teach in the University of Constantinople.

PLATONISM IN THE WORLD OF REVEALED RELIGIONS

Early Jewish Platonism. Well before the beginning of the Christian Era, Jews with some Greek education had begun to make casual use of popular Greek philosophy in expounding their revealed religion: there are traces of this in the wisdom literature of the Old Testament. In Paul's speech to the Areopagus in Acts 17, commonplaces of Stoic philosophy were employed for apologetic purposes.

Philo of Alexandria

But, as far as is known, the first Jew who was really well-read in Greek philosophy and used it extensively in the exposition and defense of his traditional religion was Philo Judaeus (Philo of Alexandria [*c.* 15 BC–after AD 45]), an older contemporary of St. Paul. Philo expressed his philosophical religion in the form of lengthy allegorical commentaries on the Jewish Scriptures, especially on Genesis. In these he showed to his own satisfaction that the ancient revelation given to Moses accorded with the teaching of the best Greek philosophers, which, in his view, was later and derivative. The Greek philosophy that he preferred and found to be most in accordance with revelation was an early form of Middle Platonism. Philo was neither approved of nor read by later orthodox Jews, but his influence on Greek-speaking and Greek-educated Christians from the 2nd century AD was great; and in important ways he determined the tone of their religious speculation.

Ancient and medieval Christian Platonism. Like Philo, the Christian Platonists gave primacy to revelation and regarded Platonic philosophy as the best available instrument for understanding and defending the teachings of Scripture and church tradition. But, also like Philo, they did not believe that truth could conflict with truth and were confident that all that was rationally certain in Platonic speculation would prove to be in perfect accordance with the Christian revelation. Their unhistorical approach and unscholarly methods of exegesis of texts, both pagan and Christian, facilitated this confidence. The general attitude of Christian Platonists was one of relatively moderate and humane otherworldliness (the cruder sorts of Christian otherworldliness and hatred of the body seem to derive from non-Platonic and non-Greek sources). They stressed the transcendence of God though, by insisting that it is a transcendence that is also the deepest immanence, they acknowledged his intimate presence within the world as well. They took a dualistic view of soul and body (though accepting bodily resurrection) and emphasized the primacy of the spiritual, while insisting on the goodness of God's material creation.

Patristic Platonism. From the middle of the 2nd century AD Christians who had some training in Greek philosophy began to feel the need to express their faith in its terms, both for their own intellectual satisfaction and in order to convert educated pagans. The philosophy that suited them best was Platonism. Though Stoicism had exerted a considerable influence on Christian ethical thinking (which has persisted to modern times), Stoic corporealism—the belief that God and the soul are bodies of a subtle and peculiar kind—repelled most Christians, and Stoic pantheism was incompatible with Christianity. The Platonism that the first Christian thinkers knew was of course Middle Platonism, not yet Neoplatonism. Its relatively straightforward theism and high moral tone suited their purposes excellently; and the influence of this older form of Platonism persisted through the 4th century and beyond, even after the works of Plotinus and Porphyry began to be read by Christians.

Justin, Origen, the Cappadocians

The first Christian to use Greek philosophy in the service of the Christian faith was Justin Martyr (martyred *c.* 165), whose passionate rejection of Greek polytheism, combined with an open and positive acceptance of the essentials of Platonic religious philosophy and an unshakable confidence in its harmony with Christian teaching, was to remain characteristic of the Christian Platonist tradition. This was carried on in the Greek-speaking world by Clement of Alexandria (*c.* 150–*c.* 215), a persuasive Christian humanist, and by the greatest of the Alexandrian Christian teachers, Origen (*c.* 185–254). Although Origen was consciously more hostile to and critical of Platonic philosophy than either Justin or Clement, he was, nonetheless, more deeply affected by it. He produced a synthesis of Christianity and late Middle Platonism of remarkable originality and power, which is the first great Christian philosophical theology. In spite of subsequent condemnations of some of his alleged views, his influence on Christian thought was strong and lasting. The Greek philosophical theology that developed during the Trinitarian controversies over the relationships among the persons of the Godhead, which were settled at the ecumenical councils of Nicaea (325) and Constantinople (381), owed a great deal to Origen on both sides, orthodox and heretical. Its most important representatives on the orthodox side were the three Christian Platonist theologians of Cappadocia, Basil of Caesarea (*c.* 329–379), Gregory of Nazianzus (*c.* 330–*c.* 389), and Basil's brother Gregory of Nyssa (*c.* 335–*c.* 394). Of these three, Gregory of Nyssa was the most powerful and original thinker (as well as the closest to Origen). He was the first great theologian of mystical experience, at once Platonic and profoundly Christian, and he exerted a strong influence on later Greek Christian thought.

Pseudo-Dionysius and others

At some time between the period of the Cappadocian Fathers and the early years of the 6th century, a new turn was given to Christian Platonism by the remarkable writer who chose to publish his works under the name of St. Paul's convert at Athens, Dionysius the Areopagite. The kind of Platonism that the Pseudo-Dionysius employed for his theological purposes was the 5th-century Neoplatonism that is best represented by Proclus (see above). Almost everything about this mysterious author is vigorously disputed by scholars. But there can be no doubt about the influence that his system of the hierarchic universe exerted upon later Christian thought; his vision of man's ascent through it—carried up by divine love, to pass beyond all hierarchy and all knowledge into the darkness of the mystical union with God—had its impact both in the East, where one of the greatest of Greek Christian Platonist thinkers, Maximus the Confessor (*c.* 580–662), was deeply influenced by the Dionysian writings and commented extensively upon them, and in the West, where they became known and were translated into Latin in the 9th century. In the Latin West there was more than one kind of Christian Platonism. An impressive and extremely difficult philosophical theology, employing ideas approximating Porphyry's version of Neoplatonism to explain and defend the orthodox doctrine of the Trinity, was produced in the second half of the 4th century by the rhetorician and grammarian Marius Victorinus. A strong and simple Platonic theism and morality, which had a great influence in the Middle Ages, was nobly expressed in the final work of the last great philosopher-statesman of the ancient world, Boethius (*c.* 470–524). This was the *Consolation of Philosophy,* written in prison while its author was under sentence of death. Boethius was also influential in the medieval West through his translations of Aristotle's logical works, especially the *Categories* together with Porphyry's *Isagoge* ("Introduction"), on which he in turn produced two commentaries. But the Christian Platonism that had the widest, deepest, and most lasting influence in the West was that of St. Augustine of Hippo (354–430).

Augustinian Platonism. Each of the great Christian Platonists understood Platonism and applied it to the understanding of his faith in his own individual way, and of no one of them was this truer than of Augustine with

his extremely strong personality and distinctive religious history. Augustine's thought was not merely a subspecies of Christian Platonism but something unique—Augustinianism. Nonetheless, the reading of Plotinus and Porphyry (in Latin translations) had a decisive influence on his religious and intellectual development, and he was more deeply and directly affected by Neoplatonism than any of his Western contemporaries and successors.

Teachings of Augustine

In his anthropology Augustine was firmly Platonist, insisting on the soul's superiority to and independence of the body. For him, as for Plotinus and Porphyry, it was axiomatic that body could not act on soul, for soul was superior in the hierarchy of reality, and the inferior cannot act on the superior. This affected both his ethical doctrine and his epistemology. On the other hand, he differed from the philosophers who influenced him in his insistence that not only man but higher spiritual beings as well are mutable and peccable, liable to sin and fall, and in his consequent stress on the necessity of divine grace. His crucial doctrine that man's destiny is determined by the right direction of love, though profoundly original, was a development rather than a contradiction of Platonism. His very original theology of history and his view of human society, however, owed little to Plotinus and Porphyry, whose interests lay elsewhere.

In his epistemology Augustine was Neoplatonic, especially in the subjectivity of his doctrine of illumination—in its insistence that in spite of the fact that God is exterior to man, men's minds are aware of him because of his direct action on them (expressed in terms of the shining of his light on the mind, or sometimes of teaching) and not as the result of reasoning from sense experience. For a Platonist, as has been said, body cannot act upon soul. Sense experience, therefore, though genuinely informative on its own level, cannot be a basis for metaphysical or religious thinking. This must be the result of the presence in the soul of higher realities and their action upon it. In Plotinus the illumination of the soul by Intellect and the One was the permanent cause of man's ability to know eternal reality; and Augustine was at this point very close to Plotinus, though for him there was a much sharper distinction between Creator and creature, and the personal relationship between God and the soul was much more strongly stressed.

In his theology, insofar as Augustine's thought about God was Platonic, he conformed fairly closely to the general pattern of Christian Platonism; it was Middle Platonic rather than Neoplatonic in that God could not be the One beyond Intellect and Being but was the supreme reality in whose creative mind were the Platonic Forms, the eternal patterns or regulative principles of all creation. Perhaps the most distinctive influence of Plotinian Neoplatonism on Augustine's thinking about God was in his Trinitarian theology. He started with the unity of God and continually insisted upon it, unlike Greek Christian thinkers, who started with the Three Persons perfectly united; and because he thought that something like the Christian doctrine of the Trinity was to be found in Plotinus and Porphyry, he tended to regard it as a philosophical doctrine and tried to make philosophical sense of it to a greater extent than the Greek Fathers did. His last and most important and influential attempt to do so was in his treatise *On the Trinity,* with its discovery of analogies to the divine mystery in the self-directed, internal activities of the soul.

Medieval Platonism. With the gradual revival of philosophical thinking in the West that began in the Carolingian period (late 8th–9th centuries), the history of Platonism becomes extremely complex. Only a sketch distinguishing the main streams of a more or less Platonic tradition is given here.

Influence of the *Timaeus*

In the 4th century the Christian exegete Calcidius (Chalcidius) prepared a commentary on Plato's *Timaeus,* which exerted an important influence on the medieval interpretation of the *Timaeus.* A Christian Platonic theism of the type of which Boethius is the finest example thus arose; based on a reading of the *Timaeus* with Christian eyes, it continued to have a strong influence in the Middle Ages, especially in the earlier period. This kind of theism, issuing in a strongly positive view of God's creation and a nobly austere but humane view of man's duty and destiny, was particularly apparent in the Christian humanism of the School of Chartres (12th century).

The widest, deepest, and most persistent Christian Platonist influence in the Latin West was that of Augustine (see above *Augustinian Platonism*). Augustinianism in a variety of forms—often stiffened, exaggerated, or distorted—persisted throughout the Middle Ages and survived the "recovery of Aristotle" (see below). In the later Middle Ages Augustine's influence was particularly strong in the Franciscan school, though not confined to it. But the greatest and most influential of medieval thinkers deeply influenced by Augustine was Anselm of Canterbury (1033/34–1109), the originator (probably on the basis of suggestions in Augustine) of the still much discussed "ontological argument" for the existence of God (see PHILOSOPHIES OF THE BRANCHES OF KNOWLEDGE: *Philosophy of religion*) and a philosopher whose humility, openness, and readiness to consider objections had a genuinely Socratic quality.

John Scotus Erigena

One of the boldest and most original thinkers of medieval Europe was John Scotus Erigena (810–*c* 877), who introduced to the West the Greek Christian Platonist tradition (see above *Patristic Platonism*), as it had been developed by Gregory of Nyssa, the Pseudo-Dionysius, and Maximus the Confessor. His views were much disapproved of by the Western church; and his great philosophical work, the *Periphyseon* (usually known as *De divisione naturae* [*On the Division of Nature*]), was not much read and ceased to be copied after his condemnation in 1210. But a considerable part of the text circulated in the form of anonymous glosses to the Latin translations of the Pseudo-Dionysius (of which the first adequate translation was by Erigena himself); and in this way his thought influenced both the tradition of Western mysticism, which derived from the Pseudo-Dionysius, and 13th-century Scholasticism, for which St. Paul's supposed disciple was still a major authority.

The Platonism in "Aristotelian" Scholasticism

There is no more superficial and misleading generalization in the history of philosophy than that which sharply opposes "Christian Platonism" and "Christian Aristotelianism." To be sure, the recovery of the authentic thought of Aristotle through Latin translations of his works in the 12th and 13th centuries was indeed a major event in the history of philosophy. But Platonism and Aristotelianism have never been tidily separated in the history of European thought. There was already a strong Aristotelian element in Middle Platonism and Neoplatonism. Byzantine theologians (in the East) from the 6th century AD onward were as Aristotelian as anybody in western Europe in the 13th century. Thirteenth-century "Aristotelian" Scholastics, though much preoccupied with the new translations of Aristotle and their philosophical and theological implications, were still deeply influenced by Augustine, Boethius, and the Pseudo-Dionysius (with glosses derived from Erigena). And the Islāmic philosophy, to be mentioned below, with which they had to grapple, was as much Neoplatonist as it was Aristotelian. Further, they also were influenced by Latin translations of two pseudo-Aristotelian works in Arabic, based on Neoplatonic sources (see below) as well as by those of some of the shorter works of Proclus (see above). It has been said that "Aquinas is closer to Plotinus than to the real Aristotle," and there is some truth in this judgment.

Islāmic and medieval Jewish philosophy. After the Muslim conquest of Syria and Egypt, there began a great work of translation of the texts that had been studied in the late Greek philosophical schools—including a number of dialogues of Plato and Neoplatonic treatises, as well as the works of Aristotle and a number of the Alexandrian Neoplatonist commentaries on them. The translations—partly from Greek, partly from Syriac versions of the Greek texts—were made between about 800 and 1000. On the basis of these translated texts an impressive development of Islāmic theology and philosophy took place, strongly influenced by Neoplatonism, though Aristotelian influence also became increasingly important. An interesting feature of this Islāmic philosophy, which distinguished it from the familiar Neoplatonism, was the reappearance in al-Fārābī and Averroës of an interest in the political and social side

of Plato's thought. The tradition may be seen in four great Muslim philosophers, the Arab al-Kindī (*c.* 800–870), the Turk al-Fārābī (*c.* 878–*c.* 950), and two who deeply influenced the medieval West, Avicenna (Ibn Sīnā, 980–1037) from Persia and Averroës (Ibn Rushd, 1126–98) from Muslim Spain. Of these, Avicenna was perhaps the more Platonist, and Averroës, whose fame and influence rested primarily on his commentaries on Aristotle, was the more Aristotelian although the latter's commentaries were written on the basis of Greek ones, some of whose authors had used them as a vehicle for Neoplatonism. Medieval Jewish philosophy, which also developed within this Muslim intellectual tradition, reflected—at least in its earlier phases—strong Neoplatonic influence. This is especially true of the thought of the early figure Isaac Israeli (mid-9th–mid-10th century), whose Platonism was pervasive, though derivative and less than fully coherent, and the first great Jewish philosopher of Muslim Spain, Avicebron (Ibn Gabirol, *c.* 1022–*c.* 1058/70), whose Platonism may have been derived from Israeli's. Avicebron's *Fons vitae* (*Fountain of Life*) was also a major influence on scholastic philosophers.

PLATONISM FROM THE RENAISSANCE TO MODERN TIMES

From the 15th century onward the dialogues of Plato and a large number of Middle Platonist and Neoplatonist works, above all the *Enneads* of Plotinus, became available in the original Greek in western Europe. As a result of this new acquaintance with the original texts, Platonic influences on Renaissance and post-Renaissance thought became even more complex and difficult to recognize than those on medieval thought. Older Neoplatonically influenced traditions (notably Augustinianism) persisted, and new ones developed from the direct reading of the Neoplatonic texts. And, at least from the time of Leibniz (1646–1716), European thinkers realized that the Neoplatonic interpretation of Plato was in some ways a distorted and one-sided one; hence they sometimes developed their own allegedly more authentic understandings of Plato on the basis of direct readings of such of his varied works as they found to be philosophically congenial. Only a few of the more interesting Platonic influences can be indicated here.

The persistence of Christian Platonism

In spite of its deep influence on Greek Christian thinkers, Platonism was regarded with profound suspicion by the Byzantine Orthodox Church. The suspicion reflected its association in the Byzantine ecclesiastical mind with the militant paganism of the Athenian Neoplatonists (see above). Nonetheless, it survived in the Byzantine world—generally underground but with an overt revival in the 11th century, in which the most notable figures were the broadly erudite Michael Psellus, who did much to enhance the prestige of philosophy, and his rival, the syncretistic Aristotelian commentator John Italus. In the following century Eustratius, metropolitan of Nicaea, and Michael of Ephesus continued the tradition of writing Neoplatonic commentary on Aristotle, plugging some of the gaps left by the Alexandrian commentators. In the 15th century the last known Byzantine philosopher, George Gemistus Plethon, a passionate pagan Platonist in the manner of Proclus, traveled to Italy (1438–39) and persuaded Cosimo de' Medici to sponsor a Platonic Academy at Florence, of which the greatest figures to emerge were its founder, Marsilio Ficino (1433–99), who translated all of Plato and Plotinus into Latin, the first complete version of either in a Western language, and the humanist Pico della Mirandola (1463–94), author of the influential *Oration on the Dignity of Man*. Ficino's *Platonic Theology: On the Immortality of Souls* contains not only Platonic and Neoplatonic philosophy but also elements drawn from medieval Aristotelianism, Cicero, Augustine, and Italian humanist writers. In spite of the paganism of Plethon, the Platonism of the Florentine Academy was a Christian one of a humane and liberal kind. This was probably at least partly due to the influence in Italy of Nicholas of Cusa (1401–64), who worked out his own very original version of Christian Platonism, influenced by the Pseudo-Dionysius, Erigena, and the German mystical tradition (as in Meister Eckehart).

The influence of the Platonism of the Florentine Academy was quite extensive; it may be seen not only in the writings of later Italian philosophers but also in the iconography of Italian Renaissance painting and in 16th-century French literature and was particularly marked in England. Perhaps the most impressive development of this post-Renaissance movement lay in the works of the Cambridge Platonists (late 17th century). Since their time a tradition of liberal Christian Platonism has persisted in England. Moreover, there have been other notable traditions of Platonically influenced Christian thought in Europe. One that deserves to be better known is that of the outstanding French philosopher of "action" Maurice Blondel (1861–1949), who found a prominent place in his system for the formation of ideas—interpreted as an important species of action that faithfully reflects the eternal order of reality. Blondel's philosophy has had a widespread influence, mainly among Catholic philosophers dissatisfied with Neoscholasticism. Another French philosopher much influenced by Platonism, in its Plotinian form, was Henri Bergson (1859–1941), whose thought attracted much attention during and just after his lifetime but has been largely neglected since.

Influences on German Idealism

The rediscovery of Proclus by the great German Idealist G.W.F. Hegel (1770–1831) had an important influence on his thought and so on the whole history of 19th-century Idealist philosophy. His contemporary F.W.J. von Schelling (1775–1854) was also strongly influenced by Neoplatonism, in his case that of Plotinus. Idealism, however, should not be interpreted as revived Neoplatonism, nor Neoplatonism as an anticipation of Idealism. But the historical influence of Neoplatonism on Idealist thought is indisputable. There was a strong reaction against Hegel's influence in some quarters, and this reaction led to a corresponding depreciation of Neoplatonism though the tradition of Idealism continued in the work of F.H. Bradley and John Ellis McTaggart in England and Josiah Royce in the United States. But 20th-century continental European philosophers and scholars were, until the 1960s, readier than English-speaking ones to take a serious interest in Neoplatonism. The latter, with some notable exceptions, maintained a hostile attitude toward that philosophy which they wrongly regarded not only as "decadent" but also as "mystical," and thus outside the true tradition of Greek philosophy.

The influence of the sort of Christian Platonism mentioned above on English literature, and especially on English poetry, has been wide and deep. But there has also been a strongly anti-Christian Neoplatonic influence, that of Thomas Taylor "the Platonist" (1758–1835), who published translations of Plato, Aristotle, and a large number of Neoplatonic works in the late 18th and early 19th centuries. Taylor was as militant in his pagan Platonism as was Gemistus Plethon. His ideas had a strong influence on the English Romantics. In the poetry of William Blake, who eventually succeeded in reconciling Taylor's paganism with his own very original version of Christianity, much of the symbolism is Neoplatonic. The Platonism of the English Romantic poets Coleridge and Shelley also derives from Taylor, although both were able to read the original texts. Taylor also deeply influenced Emerson and his circle in America. Later, in the early 20th century, the influence of Taylor's writings was again apparent in the Irish poet and dramatist William Butler Yeats, who in his later poems made use of Stephen MacKenna's then new translation of Plotinus.

The foremost process philosopher (an adherent of a view emphasizing the elements of becoming, change, and novelty in experienced reality), Alfred North Whitehead (1861–1947), perhaps because of his original and abiding concern with mathematical philosophy, was interested in Plato (though not, apparently, in the Neoplatonists); and his reading of the *Timaeus* in particular contributed something to the metaphysical system of his last period and especially to his concept of a God who does not timelessly transcend process but is in some way involved in it. Whitehead is an excellent example of a Platonically influenced thinker whose development of Plato's own thought proceeded along lines completely opposed to Neoplatonism.

EVALUATION OF PLATONISM

The essential point at issue between Platonists and their opponents through the centuries has been the existence (in some sense) of a spiritual or intelligible reality that is independent of the world, and is the ultimate origin of both existence and values. This is a very rough generalization that does not apply to the Skeptical Academy (see above *Greek Platonism from Aristotle through Middle Platonism*) or do full justice to the thought of the modern skeptical Platonist George Santayana. Platonists have understood this central doctrine in a great variety of ways and defended it with a great variety of arguments. But whenever it has been strongly held, it seems to have been by a faith depending on some sort of experience rather than simply on the conclusion of an argument. Its opponents have generally followed the lines of attack laid down by Aristotle (and to some extent anticipated by Plato in the first part of his *Parmenides*), that the doctrine involves the duplication of reality and the postulation of entities for the existence of which no sufficient evidence or arguments can be offered and the relationship of which to the world of sense experience cannot be intelligibly stated. The argument continues and will perhaps never finally be settled, but there can be no doubt about the central importance of Platonism in the history of European thought.

(A.H.A./H.J.Bl.)

MAJOR WORKS

WORKS: Plato's works are here listed in their traditional order, certain spurious items being omitted: *Euthyphrōn* (*Euthyphro*); *Apologia Sōkratous* (*Apology*); *Critōn* (*Crito*); *Phaedōn* (*Phaedo*); *Cratylos* (*Cratylus*); *Theaetētos* (*Theaetetus*); *Sophistēs* (*Sophist*); *Politikos* (*Statesman*); *Parmenidēs*; *Philēbos* (*Philebus*); *Symposion* (*Symposium*); *Phaedros* (*Phaedrus*); *Alkibiadēs* (*Alcibiades*); *Hipparchos* (*Hipparchus*); *Erastai* (*Lovers*); *Charmidēs*; *Lachēs*; *Lysis*; *Euthydēmos* (*Euthydemus*); *Prōtagoras*; *Gorgias*; *Menōn* (*Meno*); *Hippias Meizōn* (*Hippias Major*); *Hippias Elattōn* (*Hippias Minor*); *Iōn*; *Menexenos* (*Menexenus*); *Politeia* (*Republic*); *Timaeos* (*Timaeus*); *Critias*; *Nomoi* (*Laws*); and *Epinomis*.

TEXTS: The standard Greek text is the edition by Ioannes (John) Burnet, *Platonis Opera*, 5 vol. (1900–07, reprinted 1973 from various printings). All of the major works were translated into English by Benjamin Jowett, and the latest revisions of the Jowett translations are still the best available English versions of Plato. The Loeb Classical Library also contains the whole of Plato (in Greek and English); and several of the dialogues are published in the Penguin Classics series, including *Philebus*, trans. by Robin A.H. Waterfield (1982); *Protagoras and Meno*, trans. by W.K.C. Guthrie (1956, reissued 1966); and *Plato's Symposium*, trans. by W. Hamilton (1956).

RECOMMENDED LATER EDITIONS: Numerous English translations are available, including David Gallop (trans.), *Phaedo* (1975, reprinted 1983); John McDowell (trans.), *Theaetetus* (1973); Francis Macdonald Cornford (trans.), *Plato's Theory of Knowledge: The Theaetetus and the Sophist of Plato* (1935, reprinted 1973); J.B. Skemp (trans.), *Statesman* (1957, reissued 1977); R.E. Allen (trans.), *Plato's Parmenides* (1983); J.C.B. Gosling (trans.), *Philebus* (1975); R. Hackforth (trans.), *Phaedo* (1955, reprinted 1972); C.C.W. Taylor (trans.), *Protagoras* (1976); Terence Irwin (trans.), *Gorgias* (1979); R.W. Sharples (ed. and trans.), *Meno* (1985); Paul Woodruff (trans.), *Hippias Major* (1982); and James Adam (ed.), *The Republic*, 2nd ed., 2 vol. (1963, reprinted 1969).

BIBLIOGRAPHY

Plato. *Life and thought:* There are several good brief introductions to Plato, including C.J. ROWE, *Plato* (1984); J.E. RAVEN, *Plato's Thought in the Making: A Study of the Development of His Metaphysics* (1965, reprinted 1985); G.M.A. GRUBE, *Plato's Thought* (1935, reprinted 1980); G.C. FIELD, *The Philosophy of Plato*, 2nd ed. (1969, reprinted 1978); and A.E. TAYLOR, *Plato, the Man and His Work*, 7th ed. (1960, reprinted 1969). On Plato's life, see also G.C. FIELD, *Plato and His Contemporaries: A Study in Fourth-Century Life and Thought* (1930, reprinted 1975); ALICE SWIFT RIGINOS, *Platonica: The Anecdotes Concerning the Life and Writings of Plato* (1976); and the old but still valuable work by GEORGE GROTE, *Plato, and the Other Companions of Sokrates*, new ed., 4 vol. (1888, reprinted 1974). On the history of the Academy, see HAROLD CHERNISS, *The Riddle of the Early Academy* (1945, reprinted 1980); and JOHN GLUCKER, *Antiochus and the Late Academy* (1978).

A full and scholarly account of Plato's philosophy can be found in W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 4, *Plato: The Man and His Dialogues: Earlier Period* (1975), and vol. 5, *The Later Plato and the Academy* (1978), while vol. 3, *The Fifth-Century Enlightenment* (1969), contains a full account of what is known about Socrates. Other general accounts include PAUL FRIEDLÄNDER, *Plato*, 3 vol. (1958–69; originally published in German, 2nd ed., 1954–60); I.M. CROMBIE, *An Examination of Plato's Doctrines*, 2 vol. (1962–63; reissued 1979); and J.C.B. GOSLING, *Plato* (1973, reissued 1983). GILBERT RYLE, *Plato's Progress* (1966), is idiosyncratic. Much contemporary scholarly work has appeared in articles. GREGORY VLASTOS, *Platonic Studies*, 2nd ed. (1981), contains several classic papers. There are two useful anthologies: GREGORY VLASTOS (ed.), *The Philosophy of Socrates* (1971, reprinted 1980), and *Plato*, 2 vol. (1970–71, reprinted 1978).

On Plato's ethics, see in particular JOHN GOULD, *The Development of Plato's Ethics* (1955, reprinted 1972); PAMELA HUBY, *Plato and Modern Morality* (1972); and TERENCE IRWIN, *Plato's Moral Theory: The Early and Middle Dialogues* (1977, reissued 1979). On his political theory, see R.H.S. CROSSMAN, *Plato Today*, rev. 2nd ed. (1959, reissued 1971); K.R. POPPER, *The Open Society and Its Enemies*, vol. 1, *The Spell of Plato*, 5th ed. (1966, reprinted 1971); RONALD B. LEVINSON, *In Defense of Plato* (1953, reissued 1970); RENFORD BAMBROUGH (ed.), *Plato, Popper and Politics: Some Contributions to a Modern Controversy* (1967); and ROBERT W. HALL, *Plato* (1981). See also MARY MARGARET MACKENZIE, *Plato on Punishment* (1981); and RICHARD KRAUT, *Socrates and the State* (1984). For Plato's views on aesthetics, see IRIS MURDOCH, *The Fire and the Sun: Why Plato Banished the Artists* (1977); and JULIUS MORAVCSIK and PHILIP TEMKO (eds.), *Plato on Beauty, Wisdom, and the Arts* (1982).

For Plato's view of the soul, see T.M. ROBINSON, *Plato's Psychology* (1970). On the physical theory of the *Timaeus*, see GREGORY VLASTOS, *Plato's Universe* (1975); and on his attitude to science, see JOHN P. ANTON (ed.), *Science and the Sciences in Plato* (1980).

On epistemology, see W.F.R. HARDIE, *A Study in Plato* (1936); NORMAN GULLEY, *Plato's Theory of Knowledge* (1962, reprinted 1973); W.G. RUNCIMAN, *Plato's Later Epistemology* (1962); NICHOLAS P. WHITE, *Plato on Knowledge and Reality* (1976); and JON MOLINE, *Plato's Theory of Understanding* (1981). The standard study of Plato's ideas on logic and dialectic is RICHARD ROBINSON, *Plato's Earlier Dialectic*, 2nd ed. (1953, reprinted 1984). See also JULIUS STENZEL, *Plato's Method of Dialectic* (1940, reprinted 1973; originally published in German, 2nd ed., 1931); and KENNETH M. SAYRE, *Plato's Analytic Method* (1969).

On metaphysics and the theory of Forms there is a comprehensive survey by W.D. ROSS, *Plato's Theory of Ideas* (1951, reissued 1976); and a useful collection of essays, R.E. ALLEN (ed.), *Studies in Plato's Metaphysics* (1965, reprinted 1968). See also FRIEDRICH SOLMSEN, *Plato's Theology* (1942, reissued 1967); ANDERS WEDBERG, *Plato's Philosophy of Mathematics* (1955, reprinted 1977); RENFORD BAMBROUGH (ed.), *New Essays on Plato and Aristotle* (1965); I.N. FINDLAY, *Plato: The Written and Unwritten Doctrines* (1974); and WILLIAM J. PRIOR, *Unity and Development in Plato's Metaphysics* (1985).

Commentaries: Among the more useful commentaries are R.E. ALLEN, *Plato's Euthyphro and the Earlier Theory of Forms* (1970), and *Socrates and Legal Obligation* (1980); A.D. WOOZLEY, *Law and Obedience: The Law of Plato's Crito* (1979); ROSAMUND KENT SPRAGUE, *Plato's Use of Fallacy: A Study of the Euthydemus and Some Other Dialogues* (1962); B.A.F. HUBBARD and E.S. KARNOFSKY, *Plato's Protagoras: A Socratic Commentary* (1982, reissued 1984); N.R. MURPHY, *The Interpretation of Plato's Republic* (1951, reprinted 1967); R.C. CROSS and A.D. WOOZLEY, *Plato's Republic: A Philosophical Commentary* (1964, reissued 1980); NICHOLAS P. WHITE, *A Companion to Plato's Republic* (1979); JULIA ANNAS, *An Introduction to Plato's Republic* (1981); and GLENN R. MORROW, *Plato's Cretan City: A Historical Interpretation of the Laws* (1960). (J.B.)

Platonism. There are few modern English translations of the basic works of Neoplatonism, though more will appear as a result of a project sponsored by the National Endowment for the Humanities. PLOTINUS, *The Enneads*, trans. by STEPHEN MACKENNA, 3rd ed. rev. by B.S. PAGE (1962), is not entirely satisfactory; it is being replaced by the Loeb Classical Library edition, A.H. ARMSTRONG (trans.), *Plotinus* (1966–), 5 vol. having appeared to 1986. Other basic works include PROCLUS, *The Elements of Theology*, trans. by E.R. DODDS, 2nd ed. (1963), *A Commentary on the First Book of Euclid's Elements*, trans. by GLENN R. MORROW (1970), and *Proclus: Alcibiades I*, trans. by WILLIAM O'NEILL, 2nd ed. (1971). French translations of Proclus' commentaries have been made by A.J. FESTUGIÈRE, *Commentaire sur le Timée*, 2 vol. (1966–68), and *Commentaire sur la République*, 3 vol. (1970). See also PROCLUS, *Théologie platonicienne*, trans. by H.D. SAFFREY and L.G. WESTERINK (1968–), 4 vol. having appeared to 1986; JULIANUS, *Oracles chaldaïques*, trans. by ÉDOUARD DES PLACES (1971); and IAMBLICHUS, *Les Mystères d'Egypte*, trans. by ÉDOUARD DES PLACES (1966). A good source of information on the Platonic and Neoplatonic

philosophers up to and including Anselm is *The Cambridge History of Later Greek and Early Medieval Philosophy,* ed. by A.H. ARMSTRONG (1967; reprinted with revised bibliographies, 1970). The bibliographies of this work include a list of the editions of ancient and medieval sources (complete and fragmentary), with the more important translations and modern works. PAUL SHOREY, *Platonism, Ancient and Modern* (1938), remains an excellent introduction, but much of the important work on the period between Plato and Plotinus is still confined to technical articles; this is also true of later Neoplatonism. J.N. FINDLAY, *Plato and Platonism: An Introduction* (1978), argues that Plato developed a complete metaphysical system. JOHN DILLON, *The Middle Platonists: 80 B.C. to A.D. 220* (U.K. title, *The Middle Platonists: A Study of Platonism, 80 B.C. to A.D. 220,* 1977), offers a clear and comprehensive account of its subject; it gives the background to the best general book on ancient Neoplatonism, R.T. WALLIS, *Neo-Platonism* (1972). Many of the important problems in Plotinus are discussed by J.M. RIST, *Plotinus: The Road to Reality* (1967, reprinted 1977). Certain key ideas are traced in RICHARD SORABJI, *Time, Creation, and the Continuum: Theories in Antiquity and the Early Middle Ages* (1983). JAMES A. COULTER, *The Literary Microcosm: Theories of Interpretation of the Later Neoplatonists* (1976), is a study of Neoplatonic literary theory. A collection of articles on pagan and early Christian Neoplatonism may be found in H.J. BLUMENTHAL and R.A. MARKUS (eds.), *Neoplatonism and Early Christian Thought* (1981); a similar collection, extending to modern times, is DOMINIC J. O'MEARA (ed.), *Neoplatonism and Christian Thought* (1982). For the earlier Judeo-Christian tradition, see ERWIN R. GOODENOUGH, *An Introduction to Philo Judaeus,* 2nd ed. rev. (1963); and HARRY CHADWICK, *Early Christian Thought and the Classical Tradition: Studies in Justin, Clement, and Origen* (1966, reprinted 1984). An excellent short summary of the thought of St. Augustine, with due attention to the Platonist elements, is Chadwick's *Augustine* (1986), and his *Boethius: The Consolations of Music, Logic, Theology, and Philosophy* (1981), gives an account of the further development of Western Neoplatonism and includes a study of the development of Neoplatonic logic. For medieval Platonism, see FRIEDRICH UEBERWEG, *Grundriss der Geschichte der Philosophie,* vol. 2, *Die patristische und scholastische Philosophie,* ed. by BERNHARD GEYER, 12th ed. (1951); ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955, reissued 1980), with valuable bibliographical material; DAVID KNOWLES, *The Evolution of Mediaeval Thought* (1962); GORDON LEFF, *Mediaeval Thought: St. Augustine to Ockham* (1958, reprinted 1983); and WERNER BEIERWALTES (ed.), *Platonismus in der Philosophie des Mittelalters* (1969), a collection of important articles. All these works to some extent cover the later medieval period. On Islāmic philosophy see the brief account in W. MONTGOMERY WATT, *Islamic Philosophy and Theology: An Extended Survey,* 2nd ed. (1985). There is no satisfactory longer treatment in English, but for Avicenna, see SOHEIL M. AFNAN, *Avicenna: His Life and Works* (1953, reprinted 1980). A sample of the Platonist contribution to Islāmic thought may be seen in FRANZ ROSENTHAL, *The Classical Heritage in Islam* (1975; originally published in German, 1965). The best survey of Platonism in medieval Jewish philosophy is GEORGES VAJDA, "Le Néoplatonisme dans la pensée juive du moyen âge," in G.E. WEIL (ed.), *Mélange Georges Vajda* (1982), pp. 407–422. For Neoplatonic movements in Jewish Hellenistic and medieval philosophy, see JULIUS GUTTMANN, *Philosophies of Judaism: The History of Jewish Philosophy from Biblical Times to Franz Rosenzweig,* trans. from Hebrew (1964, reissued 1973). D.P. WALKER, *The Ancient Theology: Studies in Christian Platonism from the Fifteenth to the Eighteenth Century* (1972), reviews the Christian apologetic tradition of the Renaissance.

On Byzantine Platonism, see J.M. HUSSEY, *Church & Learning in the Byzantine Empire, 867–1185* (1937, reissued 1963); and BASILE TATAKIS, *La Philosophie byzantine,* 2nd ed. (1959). A good short general account of Renaissance Platonism in English is that by FREDERICK C. COPLESTON, *History of Philosophy,* vol. 3, ch. 12 and 15 (1953); see also the essays in PAUL OSKAR KRISTELLER, *Renaissance Thought: The Classic, Scholastic, and Humanistic Strains* (1961, reprinted 1980). A good short introduction to Renaissance Platonism in England is ERNST CASSIRER, *The Platonic Renaissance in England* (1953, reissued 1970; originally published in German, 1932). GERALD R. CRAGG (ed.), *The Cambridge Platonists* (1968, reprinted 1985), is an excellent anthology, with good introductions and notes; see also C.A. PATRIDES (ed.), *The Cambridge Platonists* (1969, reissued 1980). On English Christian Platonism, see WILLIAM RALPH INGE, *The Platonic Tradition in English Religious Thought* (1926, reprinted 1977). On the influence of Thomas Taylor, see *Thomas Taylor, the Platonist: Selected Writings,* ed. by KATHLEEN RAINE and GEORGE MILLS HARPER (1969); and F.A.C. WILSON, *W.B. Yeats and Tradition* (1958). RICHARD D. MCKIRAHAN, JR., *Plato and Socrates: A Comprehensive Bibliography, 1958–1973* (1978), contains 4,600 unannotated entries. (A.H.A./H.J.Bl.)

Poisons and Poisoning

Although poisons have been the subject of practical lore since ancient times, their systematic study is often considered to have begun during the 16th century, when the German-Swiss physician and alchemist Paracelsus first stressed the chemical nature of poisons. It was Paracelsus who introduced the concept of dose and studied the actions of poisons through experimentation. It was not until the 19th century, however, that the Spaniard Matthieu Orfila, the attending physician to Louis XVIII, correlated the chemistry of a toxin with the biological effects it produces in a poisoned individual. Both concepts continue to be fundamental to an understanding of modern toxicology.

Poisoning involves four elements: the poison, the poisoned organism, the injury to the cells, and the symptoms and signs or death. These four elements represent the cause, subject, effect, and consequence of poisoning. To initiate the poisoning, the organism is exposed to the toxic chemical. When a toxic level of the chemical is accumulated in the cells of the target tissue or organ, the resultant injury to the cells disrupts their normal structure or function. Symptoms and toxic signs then develop, and, if the toxicity is severe enough, death may result.

This article considers humans as the primary subjects of poisoning. It first discusses the actions of poisons on the body and then examines principal types of synthetic and natural poisons.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 321, and the *Index*.

The article is divided into the following sections:

Nature of a toxic substance 908
- Definition of a poison
- Classification of a poison

Transport of chemicals through a cell membrane 909
Conditions of exposure 909
- Routes of exposure and absorption of chemicals
- Frequency of exposure
- Dose of exposure

Distribution of toxicants in the body 911
- Role of the lymphatics
- Role of the blood
- Role of tissue blood flow
- Role of protein binding
- Role of distribution barriers

Elimination of toxicants 912
Therapeutic, toxic, and lethal responses 913
- Morphological versus functional toxic responses
- Local versus systemic toxic responses
- Immediate versus delayed toxic responses
- Reversible versus irreversible toxic responses
- Chemically induced immune responses

Teratogenesis 915
Carcinogenesis 915
Mutagenesis 916
Types of poison 916
- Agricultural and industrial chemicals 916
  - Agricultural chemicals
  - Industrial chemicals
  - General air pollutants
- Drugs and health care products 919
  - Painkillers
  - Tranquilizers and sleeping pills
  - Antipsychotic drugs
  - Cold medications
  - Antiseptics
  - Vitamins and iron pills
  - Antidepressants
  - Drugs of abuse
  - Cardiovascular drugs
  - Antiasthmatics
- Poisons of biological origin 920
  - Importance to humans
  - Microbial toxins
  - Plant poisons (phytotoxins)
  - Animal poisons (zootoxins)
- Radiation 924
  - Radiation, radioactivity, and radioisotopes
  - Ionizing radiation
  - Nonionizing radiation

Bibliography 929

Nature of a toxic substance

Definition of a poison. A poison is a substance capable of producing adverse effects on an individual under appropriate conditions. The term "substance" is almost always synonymous with "chemical" and includes drugs, vitamins, pesticides, pollutants, and proteins. Even radiation is a toxic substance. Though not usually considered to be a "chemical," most radiations are generated from radioisotopes, which are chemicals. The term "adverse effects" above refers to the injury, such as structural damage to tissues. "Appropriate conditions" refers to the dosage of the substance that is sufficient to cause these adverse effects. The dose concept is important because according to it even a substance as innocuous as water is poisonous if too much is ingested. Whether a drug acts as a therapy or as a poison depends on the dose.

Classification of a poison. Poisons are of such diverse natures that they are classified by origin, physical form, chemical nature, chemical activity, target site, or use.

Classification based on origin. Poisons are of microbial, plant, animal, or synthetic origin. Microbial poisons are produced by microscopic organisms such as bacteria and fungi. Botulinus toxin, for example, is produced by the bacterium *Clostridium botulinum* and is capable of inducing weakness and paralysis when present in underprocessed, nonacidic canned foods or in other foods containing the spores. An example of a plant toxin is the belladonna alkaloid hyoscyamine, which is found in belladonna (*Atropa belladonna*) and jimsonweed (*Datura stramonium*).

Animal toxins

Animal poisons are usually transferred through the bites and stings of venomous terrestrial or marine animals, the former group including poisonous snakes, scorpions, spiders, and ants, and the latter group including sea snakes, stingrays, and jellyfish. Synthetic toxins are responsible for most poisonings. "Synthetic" refers to chemicals manufactured by chemists, such as drugs and pesticides, as well as chemicals purified from natural sources, such as metals from ores and solvents from petroleum. Synthetic toxins include pesticides, household cleaners, cosmetics, pharmaceuticals, and hydrocarbons.

Classification based on physical form. The physical form of a chemical—solid, liquid, gas, vapour, or aerosol—influences the exposure and absorbability.

Because solids are generally not well absorbed into the blood, they must be dissolved in the aqueous liquid lining the intestinal tract if ingested or the respiratory tract if inhaled. Solids dissolve at different rates in fluids, however. For example, compared with lead sulphate granules, granules of lead are practically nontoxic when ingested, because elemental lead is essentially insoluble in water, while lead sulphate is slightly soluble and absorbable. Even different-sized granules of the same chemical can vary in their relative toxicities because of the differences in dissolution rates. For example, arsenic trioxide is more toxic in the form of smaller granules than is the same mass of larger granules because the smaller granules dissolve faster.

A poison in a liquid form can be absorbed by ingestion or by inhalation or through the skin. Poisons that are gases at room temperature (*e.g.,* carbon monoxide) are absorbed mainly by inhalation, as are vapours, which are the gas phase of substances that are liquids at room temperature and atmospheric pressure (*e.g.,* benzene). Because organic liquids are more volatile than inorganic liquids, inhalation of organic vapours is more common. Although vapours are generally absorbed in the lungs, some vapours that are highly soluble in lipids (*e.g.,* furfural) are also absorbed through the skin.

Aerosols

Aerosols are solid or liquid particles small enough to remain suspended in air for a few minutes. Fibres and dust are solid aerosols. Aerosol exposures occur when aerosols are deposited on the skin or inhaled. Aerosol toxicity is usually higher in the lungs than on the skin. An example of a toxic fibre is asbestos, which can cause a rare form of lung cancer (mesothelioma).

Many liquid poisons can exist as liquid aerosols, although highly volatile liquids, such as benzene, seldom exist as aerosols. A moderately volatile liquid poison can exist as both an aerosol and as a vapour. Airborne liquid chemicals of low volatility exist only as aerosols.

Classification based on chemical nature. Poisons can be classified according to whether the chemical is metallic versus nonmetallic, organic versus inorganic, or acidic versus alkaline. Metallic poisons are often eliminated from the body slowly and accumulate to a greater extent than nonmetallic poisons and thus are more likely to cause toxicity during chronic exposure. Organic chemicals are more soluble in lipids and therefore can usually pass through the lipid-rich cell membranes more readily than can inorganic chemicals. As a result, organic chemicals are generally absorbed more extensively than inorganic chemicals. Classification based on acidity is useful because, while both acids and alkalis are corrosive to the eyes, skin, and intestinal tract, alkalis generally penetrate the tissue more deeply than acids and tend to cause more severe tissue damage.

Classification based on chemical activity. Electrophilic (electron-loving) chemicals attack the nucleophilic (nucleus-loving) sites of the cells' macromolecules, such as deoxyribonucleic acid (DNA), producing mutations, cancers, and malformations. Poisons also may be grouped according to their ability to mimic the structure of certain important molecules in the cell. They substitute for the cells' molecules in chemical reactions, disrupting important cellular functions. Methotrexate, for example, disrupts the synthesis of DNA and ribonucleic acid (RNA).

Other classifications. Unlike the classifications described above, there is usually no predictive value in classification by target sites or by uses. Such classifications are done, however, to systematically categorize the numerous known poisons. Target sites include the nervous system, the cardiovascular system, the reproductive system, the immune system, and the lungs, liver, and kidneys. Poisons are classified by such uses as pesticides, household products, pharmaceuticals, organic solvents, drugs of abuse, or industrial chemicals.

TRANSPORT OF CHEMICALS THROUGH A CELL MEMBRANE

In order for a poison to produce toxicity, a sufficient quantity of that chemical must be absorbed into the body. Because the chemical must pass through a number of cell membranes before it can enter the blood, the ability of the chemical to cross these lipid-rich membranes determines whether it will be absorbed, and that ability depends on the chemical's lipid solubility.

Membrane structure

The cell membrane, the most external layer of all animal cells, is composed of two layers of lipid molecules (the lipid bilayer). The lipid molecules each have a hydrophilic (water-loving, or polar) end and a hydrophobic (water-hating, or nonpolar) end. Because they are surrounded by an aqueous environment, lipid molecules of the cell membrane arrange themselves so as to expose their hydrophilic ends and protect their hydrophobic ends. Suspended randomly among the lipid molecules are proteins, some of which extend from the exterior surface of the cell membrane to the interior surface.

A chemical tends to dissolve more readily in a solvent of similar polarity. Nonpolar chemicals are considered lipophilic (lipid-loving), and polar chemicals are hydrophilic (water-loving). Lipid-soluble, nonpolar molecules pass readily through the membrane because they dissolve in the hydrophobic, nonpolar portion of the lipid bilayer. Although permeable to water (a polar molecule), the nonpolar lipid bilayer of cell membranes is impermeable to many other polar molecules, such as charged ions or those that contain many polar side chains. Polar molecules pass through lipid membranes via specific transport systems.

The four types of chemical transport systems through cell membranes are diffusion, facilitated diffusion, active transport, and pinocytosis.

As mentioned above, lipophilic, nonpolar chemicals dissolve in the lipid bilayer. Simultaneously, some of the molecules are leaving the lipid bilayer. The net result is that chemicals cross the membrane until the concentrations of chemical molecules on both sides of the membrane are equal and there is no net flow of molecules across the cell membrane (diffusion). Therefore, chemicals diffuse across the membrane only when a concentration gradient exists across the cell membrane. Diffusion is considered to be passive transport because no external energy is used. Polar molecules, such as water and small water-soluble molecules (*e.g.,* urea, chloride ions, sodium ions, and potassium ions), can diffuse across membranes through the water-filled channels created by membrane proteins. Large polar water-soluble chemicals, such as sugars, however, do not diffuse through the membrane.

Concentration gradient

Certain relatively large water-soluble molecules cross the cell membrane using carriers. Carriers are membrane proteins that complement the structural features of the molecules transported. They bind to the chemicals in order to move them across the cell membrane. Energy is consumed because the transport proceeds against the concentration gradient.

Active transport systems move chemicals essential to cellular functions through the membrane into the cell. Such essential chemicals include calcium ions, amino acids, carbohydrates, and vitamins. Because the structures of poisons usually are not similar to those of chemicals essential to cells, few poisons are absorbed by active transport. Active transport, however, is important in the elimination of organic acids, bases, and foreign compounds by the kidneys and liver.

Molecules of similar structure compete with one another in binding with the carrier molecule. Thus, the transport of one chemical can be inhibited by another chemical of similar structure, a phenomenon called competitive inhibition. The chemical being transported also competes with itself for a carrier molecule, so that only a limited amount of the chemical can be transported by the carrier protein during a specific time.

Transport systems that use carrier molecules but which do not require energy to proceed are called facilitated diffusion. A chemical first binds to the carrier protein in the cell membrane and then diffuses through the membrane. Because no energy is used, facilitated transport into the cell cannot proceed if the concentration of that chemical is greater inside the cell membrane than outside. The involvement of carriers means that the process is also subject to competitive inhibition and saturation.

Large molecules, such as proteins and solid particles, are often transported by pinocytosis. The cell membrane engulfs a particle or protein molecule outside the cell, and brings it into the cell. Although inefficient, pinocytosis operates in the slow absorption of proteins and particles in the intestine and respiratory tract.

CONDITIONS OF EXPOSURE

Figure 1 summarizes the conditions of exposure to toxicants.

Routes of exposure and absorption of chemicals. *Injection.* Although not a common route of exposure for poisons, injection is the only route in which the entire amount exposed is absorbed regardless of the chemical administered, because the chemical is introduced directly into the body. Chemicals may be injected intravenously

Types of injection

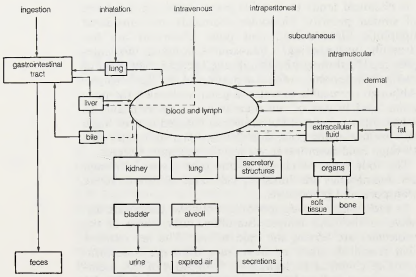


Figure 1: Routes of absorption, distribution, and excretion of toxicants in the human body.
From L.J. Casarett and J. Doull, *Toxicology: The Basic Science of Poisons* (1975); Macmillan Publishing Co., Inc.

(directly into a vein), intramuscularly (into a muscle), subcutaneously (under the skin), and intraperitoneally (within the membrane lining the organs of the abdomen).

Because the blood is the vehicle of chemical distribution in the body, intravenous injection is the most rapid method of introducing a chemical into the body. The almost instantaneous distribution, together with the irreversibility, makes intravenous injection a dangerous method of chemical exposure, with a fair chance of causing drug overdose if improperly administered.

Because there is a relatively large flow of blood to the skeletal muscles, chemicals are absorbed into the blood relatively rapidly after intramuscular injection. The slow absorption of a chemical into the blood after subcutaneous injection is probably due to the low blood flow in the subcutaneous tissues. Intraperitoneal injection is used only in biomedical research. Absorption is relatively rapid with intraperitoneal injection because of the rich blood supply to the abdomen.

Ingestion. Ingestion is the most common route of exposure to toxic chemicals. Most chemicals diffuse across the cell membrane in the nonionized form, so that the degree to which the chemical is ionized is important in determining whether a chemical is absorbed (see above *Transport of chemicals through a cell membrane*).

pH influence on ionization

Organic acids and bases dissociate into their ionized forms in response to the pH conditions of the environment. Organic acids are in their nonionized form in an acidic environment (such as the stomach), and they thus tend to diffuse across a membrane, whereas organic bases are nonionized and thus diffuse across a membrane in a basic environment (such as in the intestine).

Because the pH and surface areas differ in different segments of the gastrointestinal tract, chemical absorbabilities of these segments also differ. The major sites of absorption of ingested poisons are the stomach and the small intestine, with most of the absorption taking place in the latter. The intestine has a greater blood supply and a much larger surface area. Folds in the mucosa of the small intestine house numerous projections on the luminal surface, which increases the surface area of the 280-centimetre- (110-inch-) long small intestine to up to 2,000,000 square centimetres.

The pH on the mucosal surface of the small intestine is alkaline. Organic bases tend to be in the nonionized, lipid-soluble form and thus in general are absorbed there. The pH of the stomach contents is in the range of 1 to 2 (strongly acidic), and weak organic acids tend to be in the nonionized, lipid-soluble form. It might be expected that the poisons would be absorbed there, but, because the surface area of the stomach is much smaller than that of the small intestine, often the stomach contents (along with the poisons) are passed to the intestine before the chemicals are absorbed. The acidic environment of the stomach is the main reason for the poor absorption of organic bases by the stomach.

Topical (skin). The skin is composed of three layers of tissues—the epidermis, dermis, and subcutaneous tissues—and is an effective barrier to many substances. The outer skin layer is the epidermis, containing five layers of cells. The stratum corneum, which is the outermost epidermal layer, consists of dead cells and is the major barrier to chemical transfer through the skin. Although nonpolar chemicals cross the skin by diffusion through the stratum corneum, no active transport exists in the dead cells of this layer. The second layer, the dermis, is thicker and is composed of loosely packed connective tissue cells in a watery matrix of collagen and elastin fibres, as well as sweat glands, hair follicles, capillaries, and lymphatic vessels. After crossing the epidermis, chemical molecules are absorbed into the circulatory system via the capillaries. The capillaries drain into venules in the subcutaneous tissue.

The stratum corneum is not very permeable to water-soluble molecules and ions, although lipid-soluble molecules do cross it to a certain extent. The permeability is directly proportional to the lipid solubility of the chemical (*i.e.,* highly lipid-soluble chemicals are readily absorbed) and inversely proportional to the molecular weight of the chemical (*i.e.,* the rate of absorption increases as the molecular weight of the molecule decreases).

Rate of absorption

The rate of percutaneous absorption also varies with the thickness of the stratum corneum at different sites of the body. The rate of absorption is higher for skin on the forehead, axilla, back, and abdomen than for thicker regions like the plantar surface of the foot and the palm. The condition of the skin is also important. Percutaneous absorption is faster when the skin is moist rather than dry.

Solids are not absorbed through the skin because the skin is generally not covered with liquid and because pinocytosis does not operate in dead cells. Liquid chemicals penetrate the skin largely because of their lipid solubility. Gases and certain vapours can be absorbed through the skin also, although to a much lesser extent than via inhalation.

Inhalation. The absorption of inhaled gases and vapours differs from that of aerosols and thus will be discussed separately.

Because the same principles govern the absorption of gases and vapours, the word "gases" is used here to represent both gases and vapours. Absorption of inhaled gases takes place mainly in the lungs. Before the gases reach the lung, however, they pass through the nose, where highly water-soluble, or highly reactive, gas molecules are retained by mucosa.

Unlike intestinal and percutaneous absorption of chemicals, respiratory absorption of gases does not depend on the pH of the alveoli, because gas molecules are not ionized. It also does not depend on the lipid solubility of the gas molecules, for three reasons. First, the alveolar gas molecules are situated in close proximity to the capillaries. Second, the alveoli form a huge surface for gas absorption. Third, the time it takes for a unit of blood to go through the lungs is more than adequate for gas molecules to diffuse from the alveolar space to the blood.

Gas molecules move into the blood by partitioning, which is a gas-transfer process between two phases, such as between the air and the blood or the blood and the tissues. In partitioning, gas molecules move from a phase of high partial pressure to an adjacent phase of low partial pressure. When an individual first inhales the gas, the partial pressure of the gas is higher in the air than in the blood, driving gas molecules from the alveolar space to the blood. As more gas molecules are driven into the blood, the blood's partial pressure is raised. Eventually the partial pressure gradient between the air and blood dissipates and gas transfer stops; equilibrium is then reached, usually before the blood leaves the lungs.

The blood carries the gas molecules to the rest of the body, where the gas is transferred from the blood to the tissue until equilibrium is reached. The blood picks up more gas molecules in the lungs, and the process continues until the gas in each tissue of the body is in equilibrium with that of the blood entering the tissue. At this time, barring biotransformation, no further net absorption of gas takes place as long as the exposure concentration re-

mains constant. A person can breathe the gas forever and not absorb more, a unique characteristic of gas exposure.

The particle size and water solubility of an aerosol chemical are the important characteristics determining absorption of aerosols. For an aerosol to be absorbed, it must be inhaled and deposited on the respiratory tract. If not deposited, the aerosol particles are exhaled. Aerosols of less than 100 micrometres (0.004 inch) can be inhaled.

The aerosol size also determines the tendency of a particle to be deposited on a certain region of the respiratory tract. The larger aerosols (greater than five micrometres) tend to be deposited in the upper respiratory tract, while the smaller ones (less than five micrometres) have a greater chance of being deposited on deeper sites of the lung. The nose acts as a "scrubber" for larger aerosols and thus protects the lung from injury.

Once deposited, aerosol particles must dissolve in the liquid lining the respiratory tract in order to be absorbed. For most aerosols of poor water solubility, the particles are cleared from the respiratory tract by mechanical or cellular means. In the nasopharyngeal region, mechanical methods of clearance include sneezing and nose blowing for particles deposited on the anterior one-fifth of the nasal cavity. Particles deposited on the remaining portion of the nasal cavity and on the pharynx are removed by tiny hairs, called cilia, on the surface of these two regions, which beat almost continuously to move a covering layer of mucous toward the throat (mucociliary apparatus). Any particles deposited on the mucous are carried along and finally swallowed.

In the tracheobronchial region, mechanical clearance includes coughing and the mucociliary apparatus. The trachea, bronchi, and bronchioles, down to the terminal bronchioles, are covered with mucous and cilia. The mucociliary apparatus moves upward toward the larynx, where the respiratory tract joins the esophagus. The particles are eventually swallowed and may be absorbed by the gastrointestinal tract.

The alveolar region has the slowest rate of particle clearance in the entire respiratory system, unless the particles are water-soluble, in which case they are cleared readily by dissolution. Water-insoluble particles in the respiratory bronchioles and alveoli are removed by cellular means, principally by macrophages—scavenger cells that engulf cellular debris in the body by a process called phagocytosis. Once phagocytosed, macrophages that contain particles are removed by the mucociliary apparatus in the terminal bronchioles. Pinocytosis by the cells lining the alveoli probably move the free particles to the interstitial space, where they either enter the lymphatic capillaries and are carried to the bloodstream, or they undergo a long process of dissolution. It can take years for water-insoluble particles to dissolve, depending on the chemical, which is why water-insoluble particles deposited in the alveolar region tend to remain in the interstitial space for a long time and can cause serious harm.

Types of exposure

Frequency of exposure. The second important condition of exposure is frequency: acute (single exposure), subchronic (repeated exposures that in total last for no more than 10 percent of the lifetime of an individual), and chronic (repetitive exposures that last in total longer than 10 percent of the lifetime). The difference between the frequencies of exposure is the length of time a chemical is maintained in a target tissue. A single exposure of a poison at a certain dose may be sufficient to produce a toxic concentration in a target tissue, leading to the development of toxicity. Repetitive exposures at the same dose will then enhance the severity of the injury because of the presence of toxic levels of the chemical in the target tissue. The continuous presence of a toxic amount of poison may impair the ability of the damaged cells to carry out repair and thus prevent any chance of recovery. Consequently, a single dose that produces symptoms and toxic signs can lead to death if repeated over time. Repetitive exposures of some chemicals may also produce a different toxic effect than the acute exposure.

Toxic accumulation is one of the reasons repetitive exposures of a chemical produce toxicity while a single exposure may not. In a hypothetical case, as depicted in Figure 2, a concentration of more than 100 milligrams per gram in a target tissue is required for chemical A to cause toxic injury. If chemical A is administered at a dose that does not produce toxic levels in the tissue and the elimination of the chemical is essentially complete within 24 hours, repetitive exposures at the same dose once a day will not result in toxicity. With chemical A there will be no difference in toxicity between acute and repetitive exposures. Suppose, however, that there is a similar chemical, B, with a slower elimination rate so that chemical B is not completely eliminated from the target tissue within 24 hours. If the exposure to chemical B is carried out at the same dose as chemical A, the concentration of B in the target tissue will not return to zero after 24 hours. Consequently, daily exposures of B will cause the toxin to accumulate, so that the peak target concentration of B increases daily (Figure 2). Eventually, the toxic threshold is reached and injury will develop. Therefore, repetitive exposure can produce toxicity at a dose that does not result in injury if given only once.

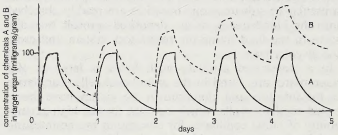


Figure 2: Accumulation of toxicants in the target organ during repetitive exposures.

Dose of exposure. The amount of chemical to which a person is exposed is extremely important. The chemical acts at a certain site, called the active site, triggering a biological response in a target tissue. Because the biological effect is caused by the presence of the chemical at the active site, the higher the concentration of the chemical at the site, the greater the response. This is the case with all known poisons, a phenomenon called the dose–response relationship.

Dose–response relationship

The dose–response curve is sigmoid, with the linear portion between approximately 16 percent and 84 percent. To compare the potency of chemicals causing similar responses, the dose that produces a biological response in 50 percent of the subject group is chosen, because it can be calculated with the least chance of error. If the biological response is mortality, the dose that kills 50 percent of the exposed population is known as the lethal dose 50, or LD50. Toxicity ratings for chemicals are based on their LD50s. The toxicity rating indicates the amount of chemical required to produce death, but it should be remembered that all chemicals can kill. Thus, all chemicals are toxic. More important than the toxicity of a chemical is its hazard or risk of usage, a concept that incorporates exposure to dosage. For example, botulinum toxin is not especially hazardous, even though it is supertoxic, because food is well-preserved, keeping the exposure or dose very low. In contrast, ethanol (alcohol) is hazardous even though it is not very toxic, because some people have a tendency to use it to excess.

DISTRIBUTION OF TOXICANTS IN THE BODY

Role of the lymphatics. After a chemical crosses the transport barrier at the portal of entry, it remains in the interstitial spaces, the spaces between cells that are filled with water and loose connective tissue. The absorbed chemical can gain entry into the bloodstream directly via the blood capillaries or indirectly via the lymphatic capillaries.

Lymphatic capillaries are minute vessels located in the interstitial spaces, with one end closed and the other end draining into larger lymphatic vessels. Just like blood capillaries, the walls of the lymphatic capillaries are composed of a thin layer of cells, the endothelial cells. Unlike the blood capillaries, however, the junctions between the endothelial cells of the lymphatic capillaries are much looser, and as a result lymphatic capillaries are much more porous than blood capillaries. Plasma proteins and excess

fluid in the interstitial spaces from blood capillaries enter the lymphatic capillaries and eventually flow back to the heart via the lymphatic system. Insoluble aerosols that cross the alveolar wall by pinocytosis may be absorbed into the circulatory system after first entering the porous lymphatic capillaries.

Role of the blood. The chemical is distributed via the blood to the various tissues of the body, where the chemical is transported across blood capillary walls. There are four types of blood capillary walls: tight, continuous, fenestrated, and discontinuous.

Types of capillary walls

Tight capillary walls are characterized by tight junctions between the endothelial cells, which prevent the diffusion of large molecules and impede that of hydrophilic molecules. The capillaries in the brain are typical of this type of capillary and form part of the blood–brain barrier.

In a continuous capillary wall, channels about five nanometres wide exist between endothelial cells, allowing most small molecules to pass through. Capillaries of this type are found in the skeletal and smooth muscles, connective tissue, lungs, and fat. Chemicals given by intramuscular or subcutaneous injection are readily absorbed into the bloodstream, as are deposited aerosols that dissolve in the fluid lining the respiratory system and cross the alveolar wall.

In a fenestrated capillary wall, holes as large as 100 nanometres are found in the endothelial cells. Capillaries in the intestine and glomeruli in the kidney have fenestrated capillary walls, which account for the high permeability of blood capillaries for absorption by the intestine and for filtration of the blood by the kidney.

The discontinuous capillary wall, the most porous of all capillaries, contains large gaps between the cells through which large molecules and even blood cells pass. This type of capillary is found in the reticuloendothelial system (including the liver, spleen, and bone marrow), which assists in the removal of aged blood cells.

The porous nature of capillaries in most tissues or organs means that a chemical in the bloodstream can be distributed almost freely to most tissues, except for organs with a barrier. The molecules diffuse from the blood to the interstitial spaces of the tissue and finally into the cells by either diffusion or active transport.

Rate of chemical accumulation

Role of tissue blood flow. The rate at which a chemical accumulates in a particular tissue is influenced by the blood flow to that tissue. The well-perfused organs—*i.e.,* organs that receive a rich blood supply relative to organ weight—include major organs like the liver, brain, and kidney. A middle group receives an intermediate blood supply and includes the skeletal muscle and skin. The poorly perfused group includes the fat and bone. As a chemical is distributed to the tissues by the bloodstream, the chemical concentrations in the well-perfused organs rapidly reach a steady state with the blood concentration while the concentrations of the chemical in the poorly perfused tissue lag behind.

Role of protein binding. The plasma contains many proteins, the most abundant being albumin. Some chemicals are known to bind to albumin. Because albumin is too large to cross the blood capillary wall, chemicals that are bound to this plasma protein are confined in the bloodstream and are not readily distributed to the tissues. Chemicals with a high affinity to bind with plasma proteins have lower concentrations in tissues than do chemicals that are not bound to plasma proteins.

Blood–brain barrier

Role of distribution barriers. There are barriers in certain organs that limit the distribution of some molecules. The blood–brain barrier consists of tight capillary walls with glial cells wrapped around the capillaries in the brain. Molecules must diffuse through two barriers to get from blood to the nerve cells of the brain. Despite the barrier, water, most lipid-soluble molecules, oxygen, and carbon dioxide can diffuse through it readily. It is slightly permeable to the ions of electrolytes, such as sodium, potassium, and chloride, but is poorly permeable to large molecules, such as proteins and most water-soluble chemicals. The blood–brain barrier is the reason the ions of some highly water-soluble metals, such as mercury and lead, are nontoxic to the brain of an adult. Children, however, are more sensitive to the toxicity of lead because the blood–brain barrier is less well developed in children.

The second distribution barrier is the blood–testis barrier, which limits the passage of large molecules (like proteins and polysaccharides), medium-sized molecules (like galactose), and some water-soluble molecules from blood into the seminiferous tubules of the testis. Water and very small water-soluble molecules, like urea, however, can pass through the barrier. The lumen of the seminiferous tubules is where sperm cells of more advanced stages develop. It is thought that the barrier protects the sperm cells.

The placental barrier between mother and fetus is the "leakiest" barrier and is a very poor block to chemicals. The placenta is composed of several layers of cells acting as a barrier for the diffusion of substances between the maternal and fetal circulatory systems. Lipid-soluble molecules, however, can cross readily, while the transfer of large-molecular-weight molecules is limited.

ELIMINATION OF TOXICANTS

An organism can minimize the potential damage of absorbed toxins by excreting the chemical or by changing the chemical into a different chemical (biotransformation), or by both methods. The body can excrete exogenous chemicals in the urine, bile, sweat, or milk; the lungs can excrete gases such as carbon monoxide.

Urinary excretion, the most common excretory pathway, takes place in the kidney, where the functional units are the glomerulus (a filter) and the renal tubule. The artery entering the glomerulus divides into capillaries, with fenestrated walls encased in the Bowman's capsule. Twenty percent of the blood is filtered through the holes in the capillary walls; molecules smaller than 60,000 molecular weight end up in the filtrate, while red blood cells, large proteins, and chemicals bound to plasma proteins are not filtered.

Chemical exchange can also take place along the renal tubule. As the filtrate flows down the renal tubule, essential molecules, such as amino acids and glucose, are reabsorbed by active transport in the first portion of the tubule (the proximal tubule). Chemicals in the filtrate are also reabsorbed by active transport if they structurally resemble these essential molecules. Unlike glomerular filtration, tubular resorption of a chemical is not influenced by whether or not it is bound to plasma proteins.

As the fluid flows down the renal tubule, water and some chemicals are reabsorbed from the tubular fluid into the blood by diffusion. The tubular fluid emerges from the kidney and is collected in the urinary bladder. Lipid-soluble chemicals are readily reabsorbed in the renal tubule, and only water-soluble chemicals are excreted in the urine to a significant extent.

The second major excretory route is the bile, which is formed in the liver and flows into the intestinal tract. The liver does not filter chemicals as does the kidney, but the liver does secrete chemicals into bile. Chemicals excreted in the bile are eventually eliminated in the feces.

Biliary excretion of a chemical does not necessarily result in the elimination of the chemical from the body. Bile is dumped into the small intestine; there is a chance that chemicals in the bile may be reabsorbed by the intestine and in turn reenter the liver via the portal vein. This cycling of a chemical, known as the enterohepatic cycle, can continue for a long time, keeping the chemical in the body.

During inhalation exposure, absorption of the gas continues until the partial pressure of the gas in the tissues is equal to that of the inspired gases in the lungs. As soon as the concentration of inspired gases decreases or the exposure terminates, respiratory excretion of the gas occurs. Because the partial pressure of the inspired gas is lower in the lungs than in blood, the blood releases some gas molecules into the alveolar space and these molecules are exhaled. The tissues lose gas molecules to the blood, which carries them to the lungs to be excreted.

The composition of sweat is similar to that of plasma except that sweat does not contain proteins. After secretion, the fluid moves through the sweat duct, where salt and water are reabsorbed. The exact mechanism of sweat

secretion is not known. It appears that sweat is a filtrate of plasma that contains electrolytes (such as potassium, sodium, and chloride) and metabolic wastes (like urea and lactic acid). Because sweat resembles a filtrate of plasma, water-soluble chemicals, like some drugs and metal ions, are found in sweat. Sweat is not a major route of excretion of chemicals, however.

Milk is a potential, albeit minor, route of chemical excretion, but more importantly it is a potential means of chemical exposure for breast-fed infants.

Most chemicals enter milk by diffusion. Therefore, only the nonionized, lipid-soluble forms of organic chemicals are found to a significant extent in milk. Chemicals with a molecular weight less than 200 and that are present in plasma not bound to proteins are more likely to be found in milk. Because the lipid content of milk is higher than that of plasma, highly lipid-soluble chemicals can exist in a more concentrated level in milk than in plasma. Therefore, milk can be a significant route of excretion for highly lipid-soluble chemicals in lactating women.

Biotransformation

Biotransformation, sometimes referred to as metabolism, is the structural modification of a chemical by enzymes in the body. Chemicals are biotransformed in several organs, including the liver, kidneys, lungs, skin, intestines, and placenta, with the liver being the most important. Chemicals absorbed in the gastrointestinal tract must pass through the liver, where they can be biotransformed and thus eliminated before being distributed to other parts of the body. This phenomenon is known as the first-pass effect. As a result, smaller amounts of certain chemicals are distributed throughout the body after oral administration than after other exposure routes, such as intravenous or intramuscular injections. Biotransformation of a chemical primarily facilitates its excretion into urine or bile; however, certain chemicals are biotransformed into more toxic forms and, as a result, biotransformation of chemicals is not always beneficial.

Biotransformation of exogenous chemicals (chemicals that are not naturally found in the body) generally occurs in two phases. In phase I, an exogenous molecule is modified by the addition of a functional group such as a hydroxyl, a carboxyl, or a sulfhydryl. This modification allows phase II, the conjugation, or joining, of the exogenous molecule with an endogenous molecule (one naturally found in the body), to take place. The major end product in most cases is a more water-soluble chemical that is easily excreted.

Phase I reactions can be classified as oxidation, reduction, or hydrolysis. Oxidation is carried out by cytochrome P-450 monooxygenases, mixed-function amine oxidases, and alcohol and aldehyde dehydrogenases. The reactions mediated by cytochrome P-450 monooxygenases can make the chemical less toxic or more toxic. The cytochrome P-450 enzymes can, for example, produce epoxides of some chemicals, which are very reactive and can attack important cellular molecules, such as DNA. The remaining phase I oxidative enzymes act on a narrow range of substrates.

In addition to the oxidation of a chemical, cytochrome P-450 monooxygenases can catalyze the reduction. Another group of enzymes that can carry out reduction is the aldehyde/ketone reductases. Each of the three groups of hydrolytic enzymes (epoxide hydrolases, esterases, and amidases, respectively) creates metabolites with a hydroxyl, carboxyl, or amino functional group.

In phase II reactions an altered exogenous chemical binds with an endogenous molecule, leading to the formation of a final product (the conjugate), which is usually much more water-soluble and easily excreted than the parent chemical. There are four types of parent compounds whose excretion can be enhanced by conjugation: glucuronic acid, glutathione, amino acids, or sulfate. The first two types are the most common phase II reactions.

Conjugation of glucuronic acid with a hydroxyl, carboxyl, amino, or sulfhydryl group leads to the formation of oxygen, nitrogen, or sulfur glucuronides, which are more easily excreted than glucuronic acid because they are more water soluble and because they contain a carboxyl group. Conjugation with glutathione also enhances excretion. Glutathione conjugation yields glutathione conjugates and mercapturic acid derivatives, which are excreted by the liver, kidney, or both.

Two types of conjugations, acetylations and methylation, do not enhance the excretion of the parent chemical. Acetylation and methylation decrease the water solubility of the parent chemical and mask the functional group of the parent chemical, preventing these functional groups from participating in conjugations that increase their excretion. Acetylation acts on chemicals with an amino group and may render them less toxic. Chemicals with an amino, hydroxyl, or sulfhydryl group can be methylated. Methylation is not as important a route of biotransformation for exogenous chemicals as it is for endogenous chemicals.

THERAPEUTIC, TOXIC, AND LETHAL RESPONSES

Because the response to a chemical varies with the dose, any substance can be a poison. Medicine can produce responses that are therapeutic (beneficial) or toxic (adverse), or even lethal. The sigmoid dose–response relationships for the therapeutic and lethal responses typically look like curves A and C, respectively, of Figure 3. If drug X has therapeutic, toxic, and lethal dose–response curves of A, B, and C, respectively, X is a very safe drug, since there is no overlap of the curves. For some medicinal agents, there is overlap of the therapeutic and lethal dose-response curves, so that a dose which causes a therapeutic response in some individuals can kill others. These agents, consequently, are not as safe.

A quantitative measurement of the relative safety of drugs is the therapeutic index, which is the ratio of the dose that elicits a lethal response in 50 percent of treated individuals (LD50) divided by the dose that elicits a therapeutic response in 50 percent of the treated individuals (TD50). For instance, the therapeutic index of drug X is 9,000 milligrams per kilogram divided by 30 milligrams per kilogram and is equal to 300. The larger the therapeutic index, the safer the drug. Diazepam and digoxin are examples of drugs with a large and a small therapeutic index, respectively.

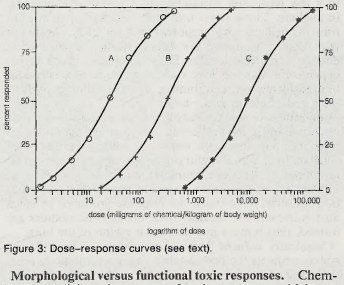


Figure 3: Dose–response curves (see text).

Morphological versus functional toxic responses. Chemicals can elicit various types of toxic responses, which can be classified by the nature of the response, the site of toxic action, the time it takes for the response to develop, and the chance of resolution of the response. The nature of the toxic response can be morphological (structural) or functional or both. In most cases, the chemical produces morphological changes in an organ, which in turn affects the function of the organ. In a small number of cases, the chemical produces functional changes in an organ without changing the structure of the organ.

Nature of response

Inhalation exposures to silica dust at a low concentration for 10 years or more can lead to chronic silicosis, a condition characterized by the formation in the lungs of silicotic nodules, which are egg-shaped lesions composed of layers of fibroblasts (reparative cells) and inflammatory cells surrounding a central silica particle. Such lesions can be considered a morphological toxic response; unless the silica exposure is prolonged, there will be little respiratory

impairment because the lungs and certain other organs have a large functional reserve. If the silica exposure is prolonged, however, the silicotic nodules coalesce (complicated silicosis), and the structure of the lungs is altered so drastically that they do not distend easily during inspiration. Oxygen exchange in the alveoli is impaired, causing such functional toxic responses as breathlessness, chest tightness, and coughing with sputum.

Malathion exposure, on the other hand, can lead to functional toxic responses without causing any morphological changes. Malathion does not alter the structure of tissues; rather, it inhibits an enzyme, acetylcholinesterase, which normally degrades acetylcholine, the neurotransmitter of the parasympathetic nervous system. Inhibition of this enzyme leads to an exaggeration of the actions of the parasympathetic nervous system, including sweating, secretion of saliva, adjustment of pupil size, and defecation. The end results are increased perspiration, increased salivation, tearing, blurred vision, abdominal cramping, diarrhea, and if severe enough, death from respiratory depression.

Local versus systemic toxic responses. Toxic responses are also classified according to the site at which the response is produced. The site of toxic response can be local (at the site of first contact or portal of entry of the chemical) or systemic (produced in a tissue other than at the point of contact or portal of entry).

An example of a local toxic effect is the tissue corrosion produced by strong acids (*e.g.,* sulfuric acid) and bases (*e.g.,* sodium hydroxide) in contact with tissues. If the exposure is external, skin burns result; if ingested, the acid or base causes serious local damage to the esophagus and stomach.

An example of a systemic toxicant is methanol, which is absorbed and biotransformed into formic acid. The acid is responsible for metabolic acidosis and optic nerve damage in the retina of the eye, leading to visual impairment, a systemic effect.

Immediate versus delayed toxic responses. Toxic responses may also be classified according to the time it takes for development of a toxic response. If it takes up to a few days after exposure, the response is considered immediate. There is no universal standard of minimum time for delayed toxic responses, but generally a response that takes more than a few days to develop is considered delayed. The time it takes for a systemic toxicant to act depends on many factors, such as the rates of absorption, biotransformation, distribution, and excretion, as well as the speed of action at the target site.

Reversible versus irreversible toxic responses. Toxic responses differ in their eventual outcomes; the body can recover from some toxic responses, while others are irreversible. Irritation of the upper respiratory tract by inhaled formaldehyde gas, for example, is rapidly reversible in that as soon as the inhalation exposure terminates, the irritation subsides. In contrast, the response produced by silica dust is irreversible because, once the silicotic nodules are formed, they remain in the alveolar region of the lung.

Chemically induced immune responses. The immune system protects the body against foreign substances, especially microbes and viruses. To be antigenic, a substance is usually both relatively large and foreign to the body. Large proteins are often strong antigens. Smaller chemicals can become antigenic by combining with proteins in chemicals called haptens.

The development of immunity toward an antigen is called sensitization. After exposure to an antigen, a combination of cellular and humoral immunity usually develops. Exposure routes that favour slow absorption into the bloodstream, such as percutaneous injection, often primarily elicit cellular immunity, while rapid routes of exposure, such as intravenous injection, favour the development of humoral immunity.

Cellular immunity utilizes phagocytes (such as macrophages, neutrophils, and eosinophils), which engulf antigens, and T-lymphocytes, which are thymus-derived, antigen-specific immune cells containing receptors specific for a special antigen. Cellular immunity is particularly important in defending the body against tumours and infections. Macrophages phagocytize antigens and secrete proteins (monokines) that regulate cells involved in immune responses. One monokine is interleukin-2, which stimulates an increase in the number of T-lymphocytes. The T-lymphocytes then develop surface receptors for specific antigens. Because T-lymphocytes survive for months or years, cellular immunity toward the antigen remains with the individual for a long time. If reexposed to the same antigen, the sensitized T-lymphocytes recognize the antigen and secrete their own proteins (lymphokines), which stimulate phagocytes to destroy the antigen. If an antigen is located on foreign or tumour cells, certain T-lymphocytes are transformed into cytotoxic T-lymphocytes, which destroy the target cells.

Humoral immunity utilizes antibodies, also known as immunoglobulins (Ig), produced by B-lymphocytes. B-lymphocytes are lymphocytes derived from the spleen, tonsils, and other lymphoid tissues. They become plasma cells, which make antibodies. There are five classes of antibodies: IgG, IgM, IgA, IgD, and IgE. IgG, IgM, and IgA are involved in humoral immunity, the function of IgD is not known, and IgE takes part in immediate hypersensitivity (see below). Classes of antibodies

Humoral immunity involves the inactivation, removal, or destruction of antigens. Antibodies can inactivate viruses by binding to them. With two antigen binding sites per protein unit, an antibody can also precipitate the antigen by cross-linking in a network formed with other antibodies. Because each IgM has five protein units, and thus five potential binding sites, IgM is particularly efficient in precipitating the antigen. After the antigen is precipitated, it can be removed by phagocytes. In addition, antigen binding by IgG or IgM activates a serum protein, called a complement, which can then initiate antigen precipitation, amplifying the inflammatory response. If the antigen is on the surface of certain cells, activated complement can also facilitate the lysis of these cells. IgG or IgM can also link the antigen to phagocytes or to killer cells, resulting in lysis of the cell by an unknown mechanism.

Although the immune system generally protects the body, it can respond in certain ways that are detrimental to some individuals. Allergy, or hypersensitivity, is a condition of increased reactivity of the immune system toward an antigen that leads to adverse effects. Substances that cause allergies are known as allergens. Hypersensitivity (allergy)

Confusion is sometimes caused by the terms hypersensitivity, hypersusceptibility, and idiosyncrasy. Hypersensitivity is a reaction to a chemical or substance in certain individuals and has a basis in the immune system. Hypersusceptibility is an increased predisposition of certain individuals to react to a chemical. Because of biological variability among humans, some individuals respond to a chemical at a dose too low to produce a similar effect in others. Idiosyncrasy is a genetically determined hypersusceptibility.

Allergic responses differ from the usual toxic responses in three ways. First, the allergic response does not occur during the first exposure to an allergen, but is evident only after at least one previous exposure. In rare occasions, an allergic response can occur on the first exposure to a chemical if the individual has already developed a hypersensitivity toward a closely related chemical. For example, people allergic to one kind of penicillin are usually allergic to other penicillins as well. Second, allergy is specific to both the allergen and the individual. Unlike in a toxic response, in which everyone exposed develops the response if a sufficient dose is administered, only a small fraction of the exposed population is sensitized by an allergen, regardless of the dose. Third, the amount of a chemical required to elicit an allergic response is usually much less than that required to produce a toxic response.

There are four types of hypersensitivities (allergies): immediate, cytotoxic, immune-complex, and delayed. Each differs from the others in the mechanism of induction and the responses produced. Immediate hypersensitivity is the most common form of allergy. Delayed hypersensitivity is the second most common, whereas cytotoxic and immune-complex hypersensitivities are relatively rare.

Immediate hypersensitivity, also called anaphylaxis, pro-



Queen hornet *(Vespa crabro).*



Black widow *(Latrodectus mactans).*

Poisonous plants and animals





Poison hemlock *(Conium maculatum).*

Toadstool *(Amanita pantherina).*



Laportea gigas, an Australian nettle tree, (left) entire tree, (below) underside of a leaf showing stinging hairs.



Plate 1: (Top left) S.C. Bisserat—Bruce Coleman Inc.; (top right) J.A.L. Cooke, (bottom left, bottom right) K. Gillett, World Life Research Institute, (centre left) Kitty Kahout from Root Resources—EB Inc., (centre right) L. Hugh Newman from The Natural History Photographic Agency—EB Inc.



Castor bean *(Ricinus communis)* with details of (left) seeds and (right) fruit.



Rosary pea *(Abrus precatorius)* with enlarged view of the poisonous seeds.

Poisonous plants



Larkspur *(Delphinium anthiscifolium)* with details of flower and fruit.

Monkshood *(Aconitum japonicum)* with details of tuberous root and flower.

Black nightshade *(Solanum nigrum)* with enlarged views of flower and fruit.







Belladonna *(Atropa belladonna)* showing details of (above) the flower, (below) the fruit, and (right) the root.



Marijuana *(Cannabis sativa)* with (left) male plant, (right) female plant.



Opium poppy *(Papaver somniferum)* with (left) mature fruit and seed and (right) detail of flower.

Jimson weed *(Datura stramonium)* with enlarged view of seeds.



Foxglove *(Digitalis purpurea)*.



Plate 2: B.W. Halstead, World Life Research Institute; artwork, (top right) Luon Kwan-Ming Li, (others) J. Fujishima

Plate 3: B.W. Halstead, World Life Research Institute; artwork, (top right) H. Matsuura, (others) J. Fujishima

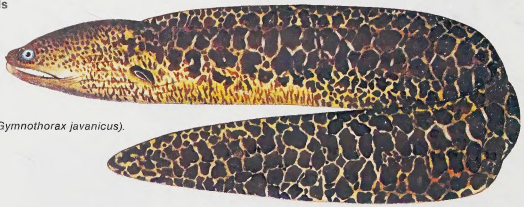


Crab *(Zozymus aeneus).*

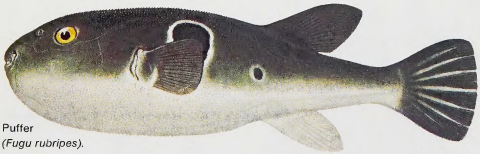


Lionfish *(Pterois volitans).*

Poisonous aquatic and land animals



Moray eel *(Gymnothorax javanicus).*



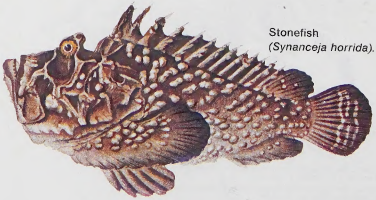
Puffer
(Fugu rubripes).



Blue crevally
(Caranx melampygus).

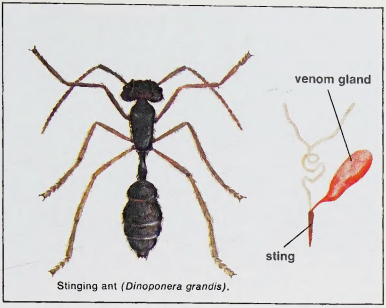


Red snapper
(Lutjanus bohar).

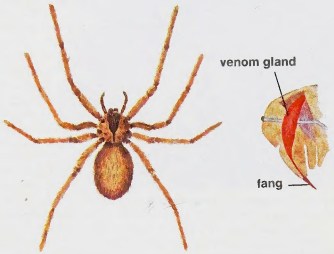


Stonefish
(Synanceja horrida).

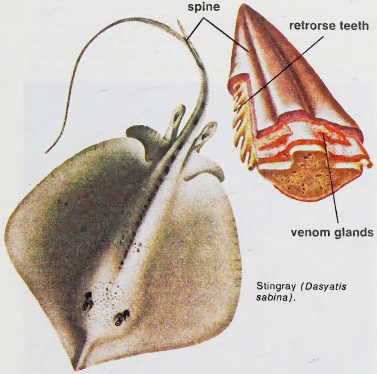
Plate 4: B.W. Halstead—World Life Research Institute: artwork, (centre right above) T. Kurnada, (centre right below) K. Tomita, (bottom right) S. Anita, (others) M. Shirao
Plate 5: B.W. Halstead—World Life Research Institute: artwork, (centre left) L. Barlow, (bottom right) L. Barlow and M. Shirao, (others) K. Fogassy



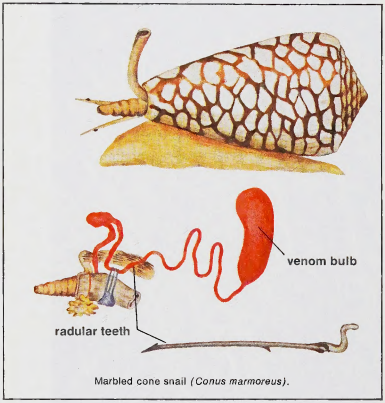
Stinging ant *(Dinoponera grandis)*.



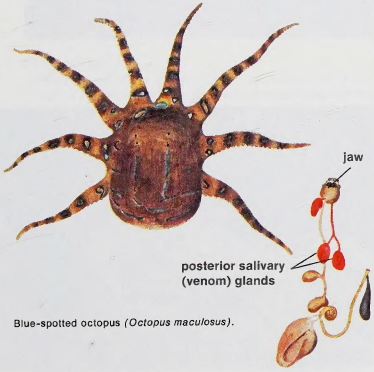
Brown recluse spider *(Loxosceles reclusa)*.



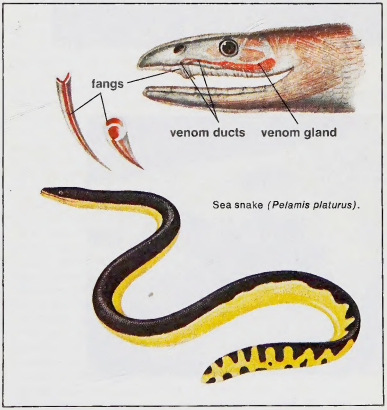
Stingray *(Dasyatis sabina)*.



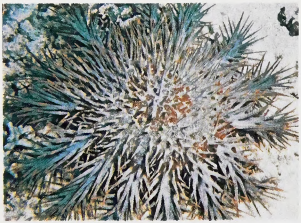
Marbled cone snail *(Conus marmoreus)*.



Blue-spotted octopus *(Octopus maculosus)*.



Sea snake *(Pelamis platurus)*.



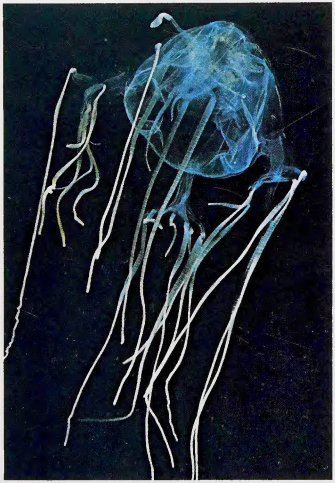
Crown of thorns starfish *(Acanthaster planci)*.



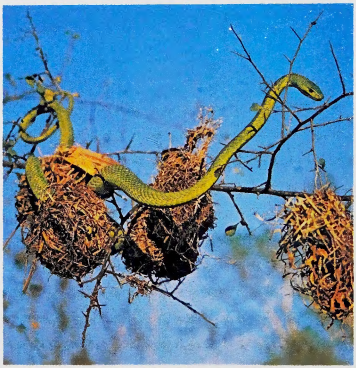
Red tide, Tampa Bay, Florida, showing fish kill and red coloration caused by dinoflagellates.

Poisonous aquatic and land animals

Sea wasp *(Chironex fleckeri)*.



Green mamba *(Dendroaspis viridis)*.



Kokoa frog or South American arrow poison frog (*Dendrobates* species).



Gila monster *(Heloderma suspectum)*.



Plate 6: (Top left, centre left) World Life Research Institute, (top left) K. Gillett, (top right, bottom left) The National Audubon Society Collection/Photo Researchers, (top right) R.F. Head, (bottom left) George Porter, (centre right) N. Myers—Bruce Coleman Inc., (bottom right) Richard Weymouth Brooks—Photo Researchers

Immediate hypersensitivity

duces IgE in response to an allergen that binds to the surface of mast cells or basophils. When reexposed to the allergen, the antigen-binding end of IgE on mast cells and basophils binds the allergen, triggering a release of anphylactic mediators from these cells. These mediators, such as histamine and serotonin, cause the contraction of certain smooth muscles (*e.g.,* those of the respiratory tract, leading to bronchoconstriction in asthmatic attacks), relaxation of blood vessels (*e.g.,* in the skin, resulting in redness, or in the whole body, causing a fall in blood pressure as in anaphylactic shock), and increased permeability of capillary walls (*e.g.,* in the skin, leading to local edema as seen in urticaria). The unique characteristic of immediate hypersensitivity is its rapid onset, with the response initiated within a few minutes of allergen exposure.

The anaphylactic mediators affect tissues differently. Thus, the allergic response depends on where the immune reaction takes place. In the skin, immediate hypersensitivity can result in skin eruptions or urticaria, characterized by wheals with redness. In the respiratory system, it can produce hay fever or asthma. In the gastrointestinal tract, allergic gastroenteritis, an inflammatory condition of the stomach and intestine, may result. Systemic anaphylaxis may involve the entire body, with shock as a key feature.

Other hypersensitivities

A second type of hypersensitivity is cytotoxic hypersensitivity, which has a gradual onset. After reexposure to an allergen, the allergen molecules attach to the surfaces of blood cells, forming an antigen new to the body. IgG or IgM binds to the new antigen on the blood cells, lysing blood cells via either complement fixation or antibody-dependent cell cytotoxicity. If the lysed cells are red blood cells, hemolytic anemia results. If platelets (the blood components intrinsic to blood clotting) are lysed, however, the blood clotting mechanism is impaired.

In a third type of allergy, immune-complex hypersensitivity, the allergen-IgG complex precipitates in tissues, resulting in inflammation via complement fixation. Immune-complex hypersensitivity in the kidney results in an inflammatory injury of the glomeruli (glomerulonephritis), and in the lung it leads to a pneumonia-like condition known as hypersensitivity pneumonitis.

Delayed hypersensitivity differs from other types in not involving humoral immunity. Upon reexposure to the allergen, sensitized T-lymphocytes release lymphokines, which trigger a series of inflammatory reactions. The inflammation leads to the development of allergic contact dermatitis in the skin and a chronic form of hypersensitivity pneumonitis in the lung. Symptoms of allergic contact dermatitis develop gradually, taking a day or two to reach maximum levels, which is the best way to distinguish allergic contact dermatitis from atopic dermatitis with similar symptoms. In contrast, the chronic form of hypersensitivity pneumonitis develops insidiously and not in a fixed time.

TERATOGENESIS

Teratogenesis is a prenatal toxicity characterized by structural or functional defects in the developing embryo or fetus. It also includes intrauterine growth retardation, death of the embryo or fetus, and transplacental carcinogenesis (in which chemical exposure of the mother initiates cancer development in the embryo or fetus, resulting in cancer in the progeny after birth).

Intrauterine human development has three stages: implantation, postimplantation, and fetal development. The first two stages are the embryonic stages and last through the first eight weeks after conception. The fetal stage begins in the ninth week and continues to birth.

Depending on the developmental stage, chemical exposure in the mother can result in different degrees of toxicity in the embryo or fetus. In the preimplantation period, a toxic chemical can kill some of the cells in the blastocyst, resulting in the death of the embryo. During the postimplantation period, chemical-induced cell death leads to one of two outcomes. If death is confined to those cells undergoing active cell division at the moment, the corresponding organs are affected, resulting in malformation. If the cell death is generalized without significant replication by the remaining cells to sustain life, the embryo dies. During the third, fetal, period, chemical injury can retard growth or, if severe enough, kill the fetus.

Critical periods

The genesis of a particular organ (organogenesis) occurs at a specific time during gestation and is not repeated. Because organogenesis is a tightly programmed sequence of events, each organ system has a critical period during which it is sensitive to chemical injury. Chemical exposure in a critical period is likely to produce malformations of that organ and not others; however, because there is some overlapping of critical periods of organ development and because chemicals frequently remain in the embryo for a period of time, malformations of more than one organ usually occur. Since organogenesis occurs mostly in the embryonic stages, chemical exposure in the first trimester should be minimized, if possible.

Mechanisms

Little is known about mechanisms of teratogenesis. It is thought that some teratogens produce malformations directly by killing the cells in the embryo. Teratogens can also produce malformations indirectly by causing maternal toxicity, resulting in oxygen or nutrient deficiency for the embryo. A few well-known examples are discussed below.

Thalidomide is a drug originally marketed to combat nausea and vomiting in pregnancy. It was discovered in the 1960s in West Germany to cause rare limb defects, among other congenital anomalies. The discoveries about thalidomide triggered legislation requiring teratogenicity testing for drugs.

Chronic alcohol ingestion during pregnancy is the most common cause of congenital problems in mental development. Ingestion of more than 30 millilitres (1 ounce) of ethyl alcohol per day during pregnancy can lead to the development of fetal alcohol syndrome, characterized by intrauterine growth retardation and subsequent learning disabilities, such as distractability, language disorders, and low IQ. Heavier consumption of alcohol, more than 60 millilitres per day, by a pregnant woman can result in malformations of the fetal brain and in spontaneous abortions.

Diethylstilbestrol (DES) is a drug used primarily from the 1940s to the '50s to prevent miscarriage. The drug is an example of a chemical that can produce transplacental carcinogenesis. It was discovered in the early 1970s that exposures to diethylstilbestrol before the ninth week of gestation could lead to the formation of rare vaginal and cervical cancers in female progenies.

CARCINOGENESIS

Carcinogens are chemicals that can produce tumours, abnormal tissue growths caused by a loss of control in cell replication. Most tumours are solid masses (*e.g.,* lung cancer), but some do not occur as tissue swellings (*e.g.,* leukemia).

Tumours may be benign or malignant. Benign tumours are to a certain degree controlled in their growth. As a result, benign tumours maintain some form of cellular organization and grow rather slowly over a period of years. In contrast, cell growth in malignant tumours is almost totally uncontrolled. Cells in malignant tumours grow very rapidly in a disoriented fashion.

Benign tumours are encapsulated by a fibrous layer and so do not invade surrounding tissue. Malignant tumours invade surrounding tissue. Thus, while a benign tumour grows at one site, a malignant tumour sends out cancerous cells via the blood and lymphatic system to distant sites of the body, spreading by a process known as metastasis. The invasion of surrounding tissues by a malignant tumour produces various symptoms.

Human carcinogenesis

Carcinogenesis is a complicated process in which many factors are known to play significant roles. Certain external environmental factors are important. For instance, cigarette smoking is known to cause predisposition to the development of lung cancer. A diet low in fibre content and high in fat is correlated with a high incidence of colorectal cancer. In addition, internal factors, such as hormonal imbalances and immunosuppression, can also increase the chance of developing tumours. Sensitivity to chemical carcinogens is known to be species-dependent. A chemical carcinogen may induce tumours in one animal species but not another, and a species that is sensitive to one carcinogen may be resistant to another. Known

human carcinogens include some anticancer drugs, aromatic (containing a benzene ring in its chemical structure) amino and nitro compounds, metals, radionuclides, and miscellaneous chemicals. In humans the respiratory tract is the most common target for chemical carcinogens, followed by the liver and the blood.

Although there have been many theories on the mechanism of chemically induced tumour formation, it is now thought that DNA is the target of most chemical carcinogens. The carcinogens interact with the DNA and interfere with its normal function. Because DNA controls cellular functions, when DNA is damaged, the cell presumably loses control and divides in a chaotic fashion. A clone of the parent cell is generated, and these cells maintain the chaotic replication, which ultimately leads to the formation of a tumour. In general it takes 10 to 20 years for the initial DNA damage in one cell to develop into a recognizable tumour.

Carcinogens that are thought to produce cancer in laboratory animals by altering the DNA are referred to as genotoxic carcinogens. They are either direct-acting or indirect-acting chemicals.

Direct-acting (reactive) genotoxic chemicals can themselves interact with DNA. Indirect-acting genotoxic carcinogens do not bind to DNA until they have been biotransformed in the body to reactive chemicals. Among the indirect-acting carcinogens, polycyclic aromatic hydrocarbons, nitrosamines, and nitrosonornicotine are found in cigarette smoke. Polycyclic aromatic hydrocarbons are also formed in charcoal-broiled meat. Nitrosamines can be formed by the nitrosation of nitrite-cured, protein-rich food, such as nitrite-cured meat and fish, in the intestine.

Chemicals that produce cancer by a mechanism other than by binding to DNA are known as epigenetic carcinogens. The mechanisms by which epigenetic carcinogens produce tumours are not known with certainty, but various theories have been proposed. Cytotoxins are thought to kill cells in the target organ. The cell death increases cell replication by the remaining cells, which somehow results in tumour development, possibly by stimulating the division of cells that have previously had their DNA damaged by a genotoxic carcinogen.

It has been proposed that hormones and chemicals which modify the activities of the endocrine system create a physiological imbalance in organs dependent for their functioning on a particular hormone. With the imbalance, the organ may lose its normal physiological control and tumour growth may occur. This may be the mechanism by which estrogens in postmenopausal women lead to development of uterine cancer and the reason antithyroid agents, such as 3-amniotriazole, produce thyroid tumours.

Chemicals that depress the immune system are thought to produce tumours by impairing cell-mediated immunity, which is important in the normal elimination of tumour cells. The development of tumours involves two main steps: initiation and promotion. Initiation is the creation by genotoxic carcinogens of a cell with abnormal DNA. After initiation, promoters stimulate the replication of these neoplastic cells and facilitate the development of the tumour. Initiators include genotoxic chemicals. Although promoters do not produce tumours directly, they are still considered carcinogens because they can lead to the development of tumours in concert with an initiator. Promoters include large chlorinated hydrocarbon molecules (*e.g.*, DDT, PCBs, TCDD, butylated hydroxy antioxidants, and saccharin) and tetradecanoyl phorbol acetate in croton oil.

Development of tumours

MUTAGENESIS

Mutagenesis is the alteration of genes. Substances able to produce mutations are naturally genotoxic substances. Once a gene is mutated in a cell, the altered gene can be passed on to daughter cells. The body has ways to repair some of these gene alterations so that the genetic damage does not always propagate.

The effect that a mutation has depends on the cell in which the mutation occurs. In the somatic cells of most organs, mutation either has no effect, causes one cell to die, or causes a cell to divide at an uncontrolled rate so that a tumour develops. If the mutation occurs in germ cells (egg and sperms), there may be detectable changes or birth defects, or stillbirth may result.

Types of poison

In regard to poisoning, chemicals can be divided into three broad groups: agricultural and industrial chemicals, drugs and health care products, and biological poisons—*i.e.*, plant and animal sources. These three groups, along with a fourth category, radiation, are discussed below.

AGRICULTURAL AND INDUSTRIAL CHEMICALS

Agricultural chemicals. The majority of agricultural chemicals are pesticides, which include insecticides, herbicides, fungicides, fumigants, and rodenticides.

Insecticides. The four main classes of insecticides are organophosphates, carbamates, chlorinated hydrocar-

Table 1: Agricultural Chemicals

chemicals	toxicity, symptoms, and signs
Insecticides	
Organophosphates (*e.g.*, malathion, parathion)	parasympathetic excess
Carbamates (*e.g.*, carbaryl, carbofuran)	parasympathetic excess
Chlorinated hydrocarbons	
DDT, methoxychlor	CNS stimulation, convulsions, nausea, vomiting
Chlordecone (Kepone)	nausea, vomiting
Insecticides from plants	
Pyrethrins	allergic contact dermatitis, asthma, CNS stimulation
Rotenone	irritation of skin, eyes, and lung; mild CNS stimulation; breast tumours in rats
Herbicides	
Chlorophenoxyacetic acids	
2,4-dichlorophenoxy-acetic acid (2,4-D)	nausea, vomiting, fatigue, diarrhea, muscle ache and twitches, peripheral nerve damage, convulsion, memory loss, colour visual disorder
2,4,5-trichlorophenoxy-acetic acid (2,4,5-T)	irritation to skin, eyes, and nose; teratogenic in animals
Bipyridinium compounds	
Paraquat	lung fibrosis; kidney and liver damage
Diquat	nosebleed, cough, fever, jaundice
Others (*e.g.*, diuron, monuron, atrazine, simazine, chlorpropham, alachlor)	irritation to the skin, nose, and throat
Fungicides	
Pentachlorophenol	quite irritating to eyes, nose, and throat; anorexia; weakness, shortness of breath; chest pain; carcinogenic in animals
Creosote	extremely irritating to skin, eyes, nose, and throat
Ferbam, thiram	moderate irritation to eyes, nose, and throat; mild skin irritation; allergic contact dermatitis
Fumigant nematocides	
1,2-dibromo-3-chloropropane (DBCP)	mildly irritating to skin, eyes, and nose; testicular damage; carcinogenic in animals
Ethylene dibromide	severe irritation to skin, eyes, and throat; headache; anorexia; CNS depression; carcinogenic in animals
Methyl bromide	headache, nausea, vomiting, drowsiness, emotional disturbances, tremors, convulsion, coma, lung irritation, bronchial inflammation
Rodenticides	
Warfarin	internal bleeding
Strychnine	restlessness, increased audio and visual sensitivities, muscular stiffness in face and legs followed by convulsion
Thallium	hair loss; skin eruptions; intestinal bleeding; anorexia; nausea; vomiting; injuries of peripheral nerves, liver, and kidney
Plant growth regulators	
Daminozide (Alar)	carcinogenic in animals

bons, and insecticides derived from plants (botanical). Organophosphate and carbamate insecticides act by inhibiting acetylcholinesterase, the enzyme that degrades acetylcholine (the messenger of the parasympathetic nervous system). As a result, acetylcholine levels remain high, exaggerating the normal functions of the parasympathetic system (Table 1). Effects such as salivation, lacrimation, urination, defecation, twitching of the skeletal muscles, and in severe poisoning, death from respiratory depression occur.

DDT

Chlorinated hydrocarbons used as insecticides, such as chlorophenothane (DDT), are larger molecules than the chlorinated hydrocarbons used as organic solvents, such as chloroform. The former stimulate the central nervous system; the latter depress it. The major toxic effect produced by these insecticides is convulsions (Table 1). The use of DDT is banned in many countries because of its environmental effects and because it may cause cancer in humans. DDT is a highly fat-soluble chemical that accumulates in fish, and, when birds eat such fish, the chemical also accumulates in their fat tissues. The DDT in the birds results in fragile eggs, which are prone to breakage. This will ultimately decrease the population of fish-eating birds.

In general, insecticides derived from plants are low in toxicity. Pyrethrins are widely used insecticides in the home. They have a rapid "knockdown" for insects and have a low potential for producing toxicity in humans. The major toxicity of pyrethrins is allergy. Rotenone is a mild irritant and animal carcinogen (Table 1).

Herbicides. Herbicides are chemicals used to kill plants. Their potential to produce toxicity in humans is rather low. High doses of 2,4-D, however, can produce muscular and neurological symptoms (Table 1). The systemic toxicity of 2,4,5-T is lower than that of 2,4-D, but 2,4,5-T is more irritating.

During the Vietnam War, Agent Orange, a mixture of 2,4-D and 2,4,5-T, was used as a defoliant. The 2,4,5-T used in the Agent Orange was contaminated with tetrachlorodibenzodioxin (TCDD), or dioxin. Although TCDD is extremely toxic to some animals, it is less so to others, but it does cause birth defects and cancer in laboratory animals. The major toxicity of TCDD in humans is in the production of chloracne, a condition characterized by acne that appears between the eyes and the ears. In more severe form, acne may be found on the face, trunk, and buttocks. (Significant adverse health effects in the soldiers exposed to low amounts of TCDD in Vietnam have not been clearly established.) Polychlorinated biphenyls (PCBs) also produce chloracne by damaging the sebaceous glands in skin.

Warfarin

Rodenticides. Warfarin was originally developed as a drug to treat thromboembolism, a disease caused by blood clots, since it inhibits the synthesis of a factor essential for the clotting of blood. The inhibition of blood clotting by warfarin can lead to internal bleeding (Table 1), however. Because of its ability to induce internal bleeding, warfarin is also used as a rodenticide.

Plant growth regulator. Daminozide, also known as Alar, is a plant growth regulator used to improve the appearance and shelf life of apples. Because of its carcinogenicity in animals (Table 1), concerns have been raised that daminozide may produce tumours in children who consume apples. As a result, the use of daminozide has greatly decreased.

Industrial chemicals. The term industrial chemicals is used to refer to chemicals used neither in agriculture nor as drugs. Therefore, it includes chemicals used in industry, as well as chemicals found in or near households. Poisoning with industrial chemicals occurs most often by either percutaneous or inhalation routes.

Depression of the central nervous system is a common effect of most hydrocarbons (Table 2). Examples of common hydrocarbons include gasoline, toluene, and heptanes; *n*-hexane; and benzene. The hydrocarbons are lipid-soluble and dissolve in the membrane of nerve cells in the brain, perturbing their function. Depression, such as drowsiness, occurs as a result. In addition, many of the hydrocarbons sensitize the heart to fibrillation by epinephrine. The hydrocarbon *n*-hexane also causes damage to peripheral nerves. Benzene is toxic to organs like the bone marrow that form blood cells and can lead to the production of leukemia.

Alcohols

Most alcohols produce depression of the central nervous system, but some alcohols cause certain unique toxicities. Examples of common alcohols include methanol, ethanol, isopropanol, ethylene glycol, and phenol. Methanol can produce blindness after being metabolized to formic acid, which also leads to acidosis, characterized by an acidic pH in the body (lower than the normal pH of 7.4). Ethanol produces birth defects in both laboratory animals and humans. It also produces fetal alcohol syndrome, a major cause of mental retardation, in children of mothers who drink excessively while pregnant. Ethanol is toxic to the liver in chronic alcoholism and is a major cause of cirrhosis, a condition characterized by hardening of the liver. Phenol differs from other alcohols in causing damage to multiple organs. Finally, ethylene glycol, which is widely used as an antifreeze agent in automobiles, causes renal damage when it is biotransformed to oxalic acid, which crystallizes in the renal tubule (Table 2).

The major toxicity produced by aldehydes, such as formaldehyde, is irritation (Table 2). Formaldehyde can also cause allergic reactions in people who have been sensitized to it. Examples of other common aldehydes include acetaldehyde, glutaraldehyde, and acrolein. The toxicities of ketones and esters are similar to those of aldehydes in causing mainly irritation of the respiratory tract if inhaled and the gastrointestinal tract if ingested. (Table 2).

Aromatic amines and nitro compounds, for example, aniline, toluidine, and nitrobenzene, produce depression of the central nervous system and methemoglobinemia (Table 2). Methemoglobinemia is a condition in which the ferrous ion in hemoglobin, which is responsible for carrying oxygen, is oxidized to the ferric form. Oxidized hemoglobin, called methemoglobin, can still carry oxygen, but it does not readily release oxygen to tissues, so that the body, in effect, has a lack of oxygen. Some aromatic amines and nitro groups are known to cause bladder cancer.

Anhydrides and isocyanates

Because both anhydrides and isocyanates are highly reactive, they are extremely irritating to the upper respiratory tract (Table 2). If the airborne concentration is sufficiently high, the upper respiratory tract cannot remove all of the isocyanate or anhydride molecules, and pulmonary injury (mainly edema) results. Such a situation occurred in Bhopal, India, in the mid-1980s, when methyl isocyanate from a chemical plant was inadvertently released into the air, killing as many as 2,500 people and injuring thousands of others. Because they are chemically reactive, anhydrides and isocyanates also tend to cause hypersensitivity responses, such as asthma and allergic contact dermatitis. Common examples of anhydrides include maleic anhydride and phthalic anhydride; examples of isocyanates include methyl isocyanate and toluene diisocyanate.

Miscellaneous organic chemicals include such compounds as phosgene, carbon disulfide, and the halogenated aromatic compounds. Phosgene gained notoriety when it was used in chemical warfare in World War I. Like anhydrides and isocyanates, phosgene is highly reactive. Instead of reacting with the mucosal linings of the upper respiratory tract, however, it tends to react with the lungs, causing edema. As a result, the lungs' defenses against bacteria are weakened, and pneumonia may occur. Halogenated aromatic compounds with more than one ring, such as polychlorinated biphenyls (PCBs), polybrominated biphenyls (PBBs), and 2,3,7,8-tetrachlorodibenzodioxin TCDD, can produce a number of toxic effects in laboratory animals, including cancer, birth defects, liver injury, porphyria, and immunotoxicity (Table 2). The PCBs have been extensively used as a cooling agent in electrical transformers. It appears that humans are more resistant to the toxicity of these compounds than are some species of laboratory animals, and the main toxic effect observed in humans is chloracne, similar to juvenile acne.

Metal compounds

Examples of metal compounds toxic to humans include manganese, lead, cadmium, nickel, and arsenic compounds, beryllium oxide, and the elemental vapours, inor-

Table 2: Industrial Chemicals

chemicals	toxicity, symptoms, and signs
Hydrocarbons	
Gasoline, toluene, xylene, hexanes, *n*-hexane, heptanes	CNS depression, headache, nausea, vomiting, irritation of skin and eyes
Chlorinated hydrocarbons	
Chloroform, carbon tetrachloride, methylene chloride, and others	CNS depression, sensitization of heart muscle; many cause liver and kidney injuries; some cause liver tumours in animals
Alcohols	
Methanol	headache; nausea; vomiting; diarrhea; abdominal pain; restlessness; cold, clammy limbs; shortness of breath; CNS depression; blurred vision; blindness
Ethanol	irritation of stomach, CNS depression, fetal alcohol syndrome; brain damage, amnesia, sleep disturbances, heart damage, fatty liver, liver cirrhosis
Aldehydes	
Formaldehyde	irritation of eyes, nose, and throat; headache; bronchitis; lung edema; asthma and allergic contact dermatitis; carcinogenic in animals
Ketones	irritation of eyes, nose, and throat
Esters	irritation of eyes, nose, and throat; pulmonary edema
Aromatic amines and nitro compounds	CNS depression, methemoglobinemia; some are carcinogenic
Anhydrides and isocyanates	irritation of skin, eyes, nose, and throat; asthma; allergic contact dermatitis
Miscellaneous organic compounds	
Polychlorinated biphenyls (PCB), polybrominated biphenyls (PBB), tetrachlorodibenzodioxin (TCDD)	chloracne, liver injury; carcinogenic and teratogenic in animals
Metals	
Lead compounds	Colic; abnormal red blood cells; injuries to kidney, peripheral nerves (weakness and palsy), and brain (irritability, restlessness, excitement, confusion, delirium, vomiting, visual disturbance); lead acetate is carcinogenic in rats
Arsenic compounds	Edema, heart damage, low blood pressure, vomiting of blood, bloody stool, skin lesions, injuries of nervous systems, liver and kidney damage, cancers of skin and lung
Corrosives (acids and alkalies)	corrosion of skin, mouth, throat, stomach, and intestine on contact; irritation of eyes, nose, and throat if inhaled
Miscellaneous inorganic compounds	
Hydrogen cyanide, potassium cyanide, sodium cyanide	drowsiness, dizziness, headache, rapid breathing, palpitations, weakness, muscle twitches, cyanosis, coma, convulsion
Hydrogen sulfide, chlorine	irritating to skin, eyes, nose, throat, and lung; chest pain; lung edema; shortness of breath; pneumonia; headache; dizziness; nausea; vomiting
Sodium fluoride, stannous fluoride	irritations of mouth, stomach and intestine; CNS depression; tooth mottling; increased bone density
Bleaches (sodium hypochlorite, calcium hypoclorite)	irritation or corrosion of esophagus, stomach, and intestine; irritation of eyes and skin; acidic condition in the body; rapid breathing; aspiration-induced lung inflammation
Silica dust, asbestos fibres	lung fibrosis; shortness of breath; cough; chest pain; cancers of the lung, linings of the lung and abdomen, and intestine (asbestos)
Air pollutants	
Sulfur dioxide	irritation of eyes, nose, throat, and lung; nausea and vomiting; shortness of breath; alterations in sense of smell and taste; unconsciousness
Nitrogen oxides, ozone	irritation of eyes, nose, throat, and lung (dry throat with ozone); shortness of breath; bluish pale appearance; rapid breathing and pulse; pneumonia; nitrogen oxides also cause the destruction of red blood cells and cause liver and kidney damage
Carbon monoxide	weakness, confusion, headache, nausea and vomiting, dizziness, drowsiness, jaw stiffness, shortness of breath, seizures, coma, lung edema, pneumonia

ganic salts, and organic compounds of mercury. Chronic manganese exposure can damage the brain, resulting in a condition with symptoms similar to Parkinson's disease, such as slurred speech, masklike face, and rigidity. Mercury can also damage the brain, leading to behavioral changes; however, mercury is also toxic to the peripheral nervous system, causing sensory and motor symptoms. In addition, mercury is toxic to the kidney. Methyl mercury is especially toxic to the developing brain of a fetus.

Lead is probably the most ubiquitous metal poison. Used for numerous purposes, before World War II it was a major constituent in paint, and it has been used in gasoline. Like mercury, lead is toxic to the nervous system and kidney (Table 2), but its toxicity is age-dependent. In children, the blood–brain barrier is not fully developed, and more lead enters the brain. The extent of damage depends on the exposure; at lower levels of exposure, small decreases in intelligence and behavioral changes may result, whereas high levels result in severe brain damage and death. In adults, lead tends to cause paralysis or weakness, indicative of peripheral nervous system damage.

In acute cadmium poisoning by ingestion, irritation of the gastrointestinal tract is the major toxicity, causing nausea, vomiting, diarrhea, and abdominal cramps. With chronic exposure by inhalation, however, kidneys and lungs are the target organs. Arsenic compounds damage many organs. They cause skin lesions, decrease in heart contractility, blood vessel damage, and injuries of the nervous system, kidney, and liver. Arsenic compounds also produce skin and lung tumours in humans. Certain nickel and hexavalent chromium compounds, as well as beryllium oxide, are toxic to the lungs and can cause lung cancer.

Acids, such as sulfuric and hydrochloric acids, and strongly alkaline compounds, such as sodium hydroxide, and potassium hydroxide are corrosive to tissues on contact and can cause severe tissue injuries (Table 2). Sulfuric acid, sodium hydroxide, and potassium hydroxide are active ingredients in drain cleaners, the ingestion of which can cause severe chemical burns of the mouth and esophagus.

Hypochlorites

Hypochlorites are often used as bleaching agents. In low concentrations, as in household bleaches, hypochlorites have little toxicity but may be irritating to tissues; they can, however, be corrosive at high concentrations (Table 2). Cyanide ions poison the oxidative metabolic machinery of cells so that insufficient energy is generated. The effect is as if there were a lack of oxygen for the cells, even though there is plenty of oxygen in the blood (Table 2). Hydrogen sulfide and chlorine are highly irritating to the respiratory tract, with pulmonary edema the major toxic effect (Table 2). Chronic fluoride poisoning is called fluorosis, which is characterized by tooth mottling and increased bone density. These changes, especially of the bone, are related to a change in body calcium caused by fluoride. Silica and asbestos remain in the lungs for long periods of time, and both produce lung fibrosis (Table 2). In addition, asbestos is a well-known human carcinogen.

General air pollutants. Sulfur dioxide, an acidic pollutant, irritates the respiratory tract. It causes violent coughing when it irritates the throat, and may result in shortness of breath, lung edema, and pneumonia when it reaches the lungs.

Both ozone and nitrogen oxides are oxidizing pollutants. Like sulfur dioxide, they cause respiratory irritation;

ozone and nitrogen oxides, however, tend to be more irritating to the lung than to the upper respiratory tract.

Carbon monoxide

Carbon monoxide, an asphyxiating pollutant, binds to hemoglobin more strongly than oxygen does. Such binding produces a hemoglobin molecule that cannot carry its normal load of four oxygen molecules. In addition, once carbon monoxide is bound, the hemoglobin molecule does not as readily release to the tissues the oxygen molecules already bound to it. Therefore, tissues lack oxygen, resulting in many toxic effects. Because the brain is especially sensitive to the lack of oxygen, most of the symptoms are neurological. Lack of oxygen is termed asphyxiation, and thus carbon monoxide is an asphyxiant.

DRUGS AND HEALTH CARE PRODUCTS

Poisoning with drugs predominantly involves oral exposures. With drugs, therefore, irritation of the respiratory tract is rare, but anorexia, nausea, and vomiting resulting from gastrointestinal irritation are common.

Painkillers. Painkillers (analgesics) are the most commonly used drugs and account for many poisoning cases. Examples include aspirin and acetaminophen. Aspirin interferes with the oxidative burning of fuel by cells. To get energy, the cells switch to a less efficient way of burning fuel that does not use oxygen but generates a lot of heat. Increased perspiration develops to counteract a rise in body temperature, leading to dehydration and thirst. Aspirin also alters the pH in the body, affecting the central nervous system (Table 3). The major toxicity of acetaminophen is liver damage.

The major toxicity from narcotic analgesics, like morphine, is depression of the central nervous system, especially the brain centre controlling respiration. The cause of death in morphine overdoses is usually respiratory failure. Nausea is caused by morphine's stimulation of the chemoreceptor trigger zone in the brain, and constipation is caused by morphine's depression of muscular activity in the intestine (Table 3).

Benzodiazepines

Tranquilizers and sleeping pills. Benzodiazepines, such as diazepam, clonazepam, and chloridazepoxide, have a wide margin of safety when used at prescribed doses. Their major toxic effect is depression of the central nervous system, which results in muscular incoordination and slurred speech (Table 3). For sleeping pills containing barbiturates, chloral hydrate, paraldehyde, and meprobamate, however, the margin of safety is much narrower, and the major toxicity is severe depression of the central nervous system, leading to respiratory and cardiovascular failure (Table 3).

Antipsychotic drugs. Like benzodiazepines, antipsychotic drugs such as chlorpromazine, perphenazine, and haloperidol have a relatively large therapeutic index, rarely causing fatalities. They occasionally may block the action of the parasympathetic and sympathetic nervous systems and thus produce such undesired effects as dry mouth and blurred vision from the former and a drop in blood pressure upon standing in the latter (Table 3).

Cold medications. Nasal decongestants, antihistamines, and cough medicine, which are found in over-the-counter preparations for treating the symptoms of colds, have a low potential to produce toxicity. Nasal decongestants, such as ephedrine, mimic the action of epinephrine by stimulating the sympathetic nervous system, and consequently, an overdose of ephedrine produces symptoms related to stimulation of the sympathetic and central nervous systems (Table 3). Depression of the central nervous system and parasympathetic blockade are two common toxicities of antihistamines such as diphenhydramine (Table 3). Depression of the central nervous system is also the major toxicity of dextromethorphan and codeine, both used to suppress coughing.

Antiseptics. Most antiseptics (*e.g.*, hydrogen peroxide, benzoyl peroxide, resorcinol, benzalkonium chloride, parabens, and cetylpyridinium chloride) produce gastrointestinal irritation if ingested (Table 3). Benzoyl peroxide and parabens applied to the skin may be toxic. Among the most toxic antiseptics are hexachlorophene, benzalkonium, and cetylpyridinium chloride, any of which can cause injuries to internal organs. Systemic toxicity (double vision, drowsiness, tremor, seizures, and death) with hexachlorophene is more likely to occur in babies because the relatively thin stratum corneum of their skin is highly permeable.

Vitamins and iron pills. Deficiencies as well as excesses of vitamins are harmful. Excessive vitamin A (retinol, or retinoic acid), known as hypervitaminosis A, can result in

Excessive vitamin A

Table 3: Drugs and Health Care Products

drugs	toxicity, symptoms, and signs
Painkillers	
Aspirin, sodium salicylate	increased perspiration, respiration increased initially, dehydration, acidity in the body, hypoglycemia, CNS depression, respiration decreased, nausea, vomiting, diarrhea, confusion, coma, convulsion, lung edema, death
Acetaminophen	skin rash, decreases in blood cells, liver and kidney injuries, hypoglycemia, coma
Morphine	nausea, vomiting, pinpoint pupil, depressed respiration, delusions, confusion, muscle flaccidity, constipation, coma, death
Tranquilizers and sleeping pills	
Benzodiazepines	increased salivation, muscular incoordination, slurred speech, weakness, seizures, irritability, loss of appetite
Barbiturates	slowed respiration, CNS depression, depressed heartbeat, low blood pressure, shock, kidney failure, lung edema, pneumonia, muscular incoordination, slurred speech, pinpoint pupil, coma, death
Antipsychotic drugs	sympathetic blockade reflex increase in heart rate, parasympathetic blockade, tremors, rigidity, restlessness, jaundice
Nasal decongestants	nervousness, dizziness, tremor, confusion, increased blood pressure and heart rate
Antihistamines	drowsiness, dizziness, ear ringing, blurred vision, lack of coordination, headache, nausea, vomiting, loss of appetite, heartburn, dry mouth and throat, cough, palpitations, decrease in blood pressure, chest tightness, tingling of the hands
Cough medicine	CNS depression
Antiseptics	irritation of esophagus and stomach when ingested
Vitamins and iron pills	
Vitamin A	fatigue, dizziness, severe headache, vomiting, edema, dry and peeling skin, enlarged liver and spleen, teratogenic, red skin eruptions, abnormal hair growth
Iron	nausea, upper abdominal pain, diarrhea, bloody or brown vomit, dehydration, intestinal bleeding, liver damage, drowsiness, acidic condition in the body, rapid breathing, shock
Antidepressants	
Tricyclic antidepressants	parasympathetic blockade, CNS damage, cardiovascular system damage
Lithium salts	thyroid enlargement, edema, increased urination, abnormal heart rhythm, vomiting, diarrhea, tremor, muscle flaccidity, seizures, coma
Drugs of abuse	primarily CNS effects
Cardiovascular drugs	
Digitalis (*e.g.*, digoxin, digitoxin)	gastrointestinal irritation, abdominal discomfort, salivation, fatigue, facial pain, visual disturbances, confusion, delirium, hallucinations
Beta blockers (*e.g.*, propanolol, metoprolol)	constriction of bronchi, nausea, vomiting, diarrhea, constipation, headache, insomnia, dizziness, abnormal heart rhythm
Verapamil	headache, dizziness, gastrointestinal symptoms, edema, rash, abnormal heart rhythm, lowered blood pressure
Procainamide, quinidine	anorexia, nausea, vomiting, confusion, delirium, psychotic behaviour, abnormal heart rhythm, lowered blood pressure
Therapeutics for asthma	CNS stimulation

skin lesions, edema, and liver damage (Table 3). Overconsumption by Alaskan natives of polar bear liver, a rich source of vitamin A, has produced acute toxicities, characterized by irresistible sleepiness and severe headaches. Chronic poisoning with vitamin A can cause neurological symptoms, including pain, anorexia, fatigue, and irritability (Table 3).

Excess vitamin C can lead to kidney stones. Apart from irritation of the skin and respiratory tract, the most severe toxicity of vitamin K excess is the increased destruction of red blood cells, which leads to anemia and the accumulation of bilirubin, one of the products of hemoglobin degradation (Table 3). Excess bilirubin can result in brain damage in newborns, a condition known as kernicterus. Because the blood–brain barrier is not well developed in newborns, bilirubin enters and damages the brain. Due to the blood–brain barrier, kernicterus is not seen in adults.

Iron, a metal that is necessary for normal health, can also cause poisoning. The toxicity of iron is a result of its corrosive action on the stomach and intestine when present in high concentrations. As a result, intestinal bleeding occurs, which can lead to the development of shock (Table 3).

Antidepressants. Among tricyclic antidepressants, amitriptyline and imipramine account for most of the fatal cases of poisoning. These drugs have a number of effects, including blockage of the parasympathetic system and damage to the central nervous system, the latter producing symptoms such as fatigue, weakness, lowered body temperature, seizures, and respiratory depression (Table 3). Death is usually caused by damage to the heart. Lithium salts, used to treat manic depression, have a relatively low therapeutic index.

Drugs of abuse. Mind-altering drugs commonly abused include amphetamines, cocaine, phencyclidine, heroin, and methaqualone. These drugs are primarily toxic to the central nervous system; amphetamine and cocaine cause stimulation of the system (hallucinations and delirium), and heroin causes the depression of the system (depressed respiration and coma). In contrast, phencyclidine and methaqualone are biphasic, producing first depression (drowsiness) and then stimulation of the central nervous system (delirium and seizures). Amphetamines also affect the gastrointestinal tract (anorexia, nausea, vomiting, diarrhea) and stimulate the cardiovascular system (increased blood pressure and heart rate, palpitations, and abnormal heart rhythm). In addition to hallucinations and delirium, cocaine causes euphoria, sexual arousal, confusion, and sympathetic stimulation. Phencyclidine is also known to cause aggression and psychotic behaviour, while methaqualone produces excessive dreaming and amnesia.

Digitalis

Cardiovascular drugs. Digitalis (*e.g.,* digoxin) is a class of drugs used for congestive heart failure, with a very narrow margin of safety. Digitalis overdose usually begins with gastrointestinal symptoms, such as anorexia, nausea, and vomiting, followed by sensory symptoms, such as pain and visual disturbances (Table 3). There are also effects on the central nervous system, characterized by delirium and hallucinations.

The major toxicities of beta blockers (*e.g.* propranolol and metoprolol) result from the blockage of sympathetic effects on the tracheobronchial tree (lung) and heart. Sympathetic stimulation relaxes smooth muscles in the tracheobronchial wall and makes the heart beat faster and more forcefully. Blockage produced by propranolol or metoprolol can cause bronchoconstriction and heart failure (Table 3).

Antiasthmatics. Drugs for treating asthma, such as theophylline and aminophylline, are structurally similar to caffeine. Like caffeine, which is a stimulant, theophylline and aminophylline also stimulate the central nervous system. Therefore, excitement, delirium, rapid breathing, increased heart rate, and seizures occur with an overdose. With excessive stimulation of the heart, palpitations and irregular heart rhythm (arrhythmia) can result, leading to sudden death. (C.D.Kl./K.L.W.)

POISONS OF BIOLOGICAL ORIGIN

Biotoxins can be conveniently grouped into three major categories: (1) microbial toxins, poisons produced by bacteria, blue-green algae, dinoflagellates, golden-brown algae, etc., (2) phytotoxins, poisons produced by plants, and (3) zootoxins, poisons produced by animals. The geographic distribution of poisonous organisms varies greatly; poison-producing microorganisms tend to be ubiquitous in their distribution. Poisonous plants and animals are found in greatest abundance and varieties in warm-temperate and tropical regions. Relatively few toxic organisms of any kind are found in polar latitudes.

Evolutionary aspects of biotoxins

Knowledge of the evolutionary significance and development of most biotoxins is largely speculative and poorly understood. In some instances they may have developed during the evolution of certain animal species as part of the food procurement mechanism (*e.g.,* in snakes; cnidarians, jellyfishes, and their relatives; mollusks, octopuses, and others; and spiders). Biotoxins may also function as defensive mechanisms, as in some snakes, fishes, arthropods (*e.g.,* insects, millipedes), and others. The defense may be quite complex—as in the protection of territorial rights for reproductive purposes—and inhibitory or antibiotic substances may be produced that result in the exclusion of competitive animal or plant species. Certain marine organisms and terrestrial plants may release into the water, air, or soil inhibitory substances that discourage the growth of other organisms; well-known examples include the production of antibiotic substances by microorganisms. Similar chemical-warfare mechanisms are used in battles for territorial rights among the inhabitants of a coral reef, a field, or a forest. Thus biotoxins play important roles in the regulation of natural populations. Of increasing interest has been the discovery that certain substances, which may be toxic to one group of organisms, may serve a vital function in the life processes of the source organism.

Entry methods for biological poisons

Importance to humans. Venom-producing animals and stinging and dermatogenic (*i.e.,* skin-poisoning) plants capable of inflicting pain and sometimes death by means of parenteral contact (*i.e.,* by bringing poisons into the body other than through the digestive tract) constitute environmental hazards. Biotoxic agents may produce their injurious effects by becoming involved in the food supply; ingestion of a poisonous microbial organism, plant, or marine animal or one of their toxic by-products may cause intoxication. An example is that of the shore fishes of many tropical islands; otherwise valuable food fishes are frequently contaminated by a poison called ciguatoxin. The poison, a potent neurotoxin (nerve poison), is accidentally ingested by the fishes in their food; such fish can no longer be used for either human or animal consumption.

Some of the effects produced by biotoxins on humans are of an acute nature, and the injuries they cause are readily discernible. The effects of some of the mycotoxins (poisons produced by fungi) and poisons produced by plants, however, are long-term and chronic; they result in the development of cancerous growths and other chronic degenerative changes that are sometimes difficult to detect.

Microbial toxins. Microbial poisons are produced by the Monera (bacteria and blue-green algae) and Protista (algae, protozoa, and others), and the Fungi. Various classifications have been proposed for the microbial poisons, but none is entirely satisfactory. The problems encountered when dealing with these organisms result from a lack of precise knowledge concerning their biological nature and their phylogenetic relationships; in addition, their poisons show great diversity and chemical complexity. The following outline, however, is useful in dealing with this subject.

Moneran toxins. The prefixes "exo-" and "endo-" are retained in classifying the bacterial toxins mainly for historical reasons rather than because they are found either outside or inside the bacterial cell. The main differences in these toxins lie in their chemical structure.

Bacterial exotoxins

Poisonous proteins from bacteria are sometimes referred to as bacterial exotoxins. The exotoxins are generally produced by gram-positive organisms (*i.e.,* bacteria that react in certain ways to the staining procedure known as Gram staining); at least two bacteria, *Shigella dysenteriae* and *Vibrio cholerae,* that produce exotoxins are gram-nega-

Table 4: Representative Toxic Microfungi

	toxin	comments
Claviceps purpurea (ergot)	"ergotaxine," a complex of toxic alkaloids, ergocryptine, ergocornine, ergocristine, and others	causes poisoning in animals and humans; produces vomiting, abdominal pain, numbness, nervous disorders, convulsions, gangrene, and abortion
Stachybotrys alternans	stachybotryotoxin	causes a toxicosis in animals and humans; produces stomatitis (inflammatory disease of the mouth), rhinitis (inflammation of the mucous membranes of the nose), conjunctivitis (inflammation of the inner surface of the eyelid), failure of blood to clot, blood abnormalities, neurological disturbances, and death
Aspergillus flavus and other species, *Penicillium* species	aflatoxin complex (16 or more known toxins)	causes toxicosis in animals and possibly man; toxins damage liver and kidneys; aflatoxin is one of the most potent liver-cancer-producing agents known
Fusarium sporotrichioides and other *Fusarium* species	fusariogenin, epicladosporic acid, fagicladosporic acid	causes alimentary toxic aleukia in animals and man; produces burning sensation of the mouth, tongue, throat, and stomach; causes nausea, vomiting, headache, cold extremities, hemorrhagic spots, convulsions, anemia, gangrene, death
Cladosporium epiphyllum and other species of *Cladosporium*	same as *Fusarium* species	
Pithomyces chartarum (*Sporidesmium bakeri*)	sporidesmin	causes facial eczema (an eruptive severe rash) in cattle; produces sensitization of the skin to sunlight, resulting in scab formation and sores; there may also be severe liver damage
Fusarium species, *Rhizopus* species, *Aspergillus* species, *Penicillium islandicum*, and others	luteoskyrin, "islanditoxin," citrinin, citreoviridin, and others; a large complex of poisons is involved	causes moldy or yellowed rice, which is toxic to animals and man; the effects in man have not been well defined; may cause nausea, vomiting, diarrhea, prostration, liver damage, and death; the effects vary greatly because of the various poisons involved

tive, however. The exotoxins usually do not contain any nonprotein substances, and most are antigenic; *i.e.,* they stimulate the formation of antibodies. The exotoxins may appear in the culture medium in which the bacteria are growing during the declining phases of growth; in some cases they are released at the time of normal destruction of the cells after death (autolysis). The exotoxins are less stable to heat than are the endotoxins, and they may be detoxified by agents that do not affect endotoxins. They are more toxic than endotoxins, and each exotoxin exerts specific effects which are collectively known as pharmacological properties. Exotoxins are neutralized by homologous antibodies—*i.e.,* the active agents in blood serum produced by a process involving the bacteria against which the serum is to be used.

Endotoxins are antigens composed of complexes of proteins, polysaccharides (large molecules built up of numerous sugars), and lipids (fats). The protein part determines the antigenicity, or quality of being reacted against as a foreign substance in a living organism. The polysaccharide part determines the immunological specificity, or limitations on the types of antibodies that can react with the endotoxin molecule and neutralize it (the immunological reaction). Some of the lipids possibly determine the toxicity. Endotoxins are derived from the bacterial cell wall and, when cells are grown in culture, are released only on autolysis. Endotoxins are not neutralized by homologous antibodies and are relatively stable to heat; all of them have the same pharmacological properties.

The Cyanobacteria, or blue-green algae, are among the most primitive and widely distributed of all organisms. They have extreme temperature tolerances. Some strains of a species are toxic; other strains of the same species are not. Water blooms of blue-green algae have been responsible for the death of fishes, waterfowl, cattle, horses, swine, and other animals. Blue-green algae have also been implicated as causes of human intoxications.

Mycotoxins. Fungi are plantlike members of the kingdom Fungi (Mycota) that do not contain chlorophyll. A significant number are known to produce poisons of various types. Toxic fungi can be roughly divided into two main categories on the basis of their size: the smaller microfungi and the larger mushrooms. The toxic microfungi are members of one of two classes: Ascomycetes, or the sac fungi, and the Deuteromycetes, or the imperfect fungi (*i.e.,* fungi in which no sexual reproductive stages are known). The large toxic mushrooms, or toadstools, are mostly members of the class Basidiomycetes, although some Ascomycetes, such as the poisonous false morel (*Gyromitra esculenta*), may attain a size as large as some of the mushrooms.

The ability of certain fungi, such as ergot (*Claviceps purpurea*) and some mushrooms, to produce intoxication has long been known. During the 19th century it was recognized that molds are responsible for such diseases as yellow-rice toxicoses in Japan and alimentary toxic aleukia in Russia. The eruption of so-called turkey X disease in England in 1960 and the resulting discovery of the substance known as aflatoxin (see Table 4) stimulated study of the subject of mycotoxicology. Because mycotoxins have now been recognized as potential cancer-producing agents (carcinogens) that can become involved in man's food supply, they have become important in the study of environmental carcinogenesis.

Poisonous mushrooms

Poisonous mushrooms, or toadstools as they are commonly called, are the widely distributed members of the class Basidiomycetes, although only a few are known to be poisonous when eaten (see Table 5); some of the poisons, however, are deadly. Most deaths attributed to mushroom poisoning result from eating members of the genus

Table 5: Representative Poisonous Mushrooms

	toxin	type of poisoning
Lorchel, or false morel (*Gyromitra esculenta*)	gyromitrin	toxicity to people is variable; causes severe liver damage accompanied by nausea, vomiting, abdominal pain, jaundice, enlarged and tender liver, coma, convulsions; fatality rate about 15 percent
Fly mushroom, or fly agaric (*Amanita muscaria*)	muscarine	symptoms develop rapidly and are severe, consisting of severe gastrointestinal disturbances, delirium, hallucinations, convulsions; rarely causes death
Death cap (*Amanita phalloides*)	amanitine, phalloidine	symptoms develop slowly, about 6–15 hours after eating: extreme abdominal pain, nausea, vomiting, excessive thirst, anuria (absence or defective excretion of urine), prostration, weakness, jaundice, cyanosis, convulsion, death; fatality rate about 50 percent; no known antidote but some treatment available
Boletus miniato-olivaceus	muscarine(?)	causes gastrointestinal and visual disturbances
Jack-o'-lantern fungus (*Clitocybe illudens*)	muscarine	causes gastrointestinal upset, not fatal
Inky cap (*Coprinus atramentarius*)	unknown	some people experience a peculiar type of intoxication after eating this mushroom and then drinking an alcoholic beverage: giddiness, gastrointestinal upset, prostration, and tachycardia (rapid heart action); the alcohol is believed to increase the solubility and absorption of the poison
Entoloma lividum	unknown	causes gastrointestinal upset, usually not fatal
Inocybe patouillardi	muscarine	symptoms are similar to *A. muscaria* poisoning
Lepiota morgani	unknown	causes gastrointestinal upset; fatalities have been reported
Mexican hallucinogenic mushroom (*Psilocybe mexicana*)	psilocybin, psilocin	causes euphoria, loss of sense of distance and size, and hallucinations

Table 6: Representative Poisonous Plants

name and distribution	toxic principle	toxic effects and comments
Plants poisonous to eat		
Rosary pea, or jequirity bean (*Abrus precatorius*); tropical regions	abrin (*N*-methyltryptophan) and abric acid	onset of symptoms may be delayed several hours to two days; vomiting, diarrhea, acute gastroenteritis, chills, convulsions, death from heart failure; one seed chewed may be fatal to a child
Aconite, or monkshood (*Aconitum napellus*); North America, Europe	aconite and a complex of other alkaloids	tingling, burning sensation in tongue, throat, skin; restlessness, respiratory distress, muscular uncoordination, vomiting, diarrhea, convulsions, possible death; an extremely poisonous plant
Corn cockle (*Agrostemma githago*); North America, Europe	githagin, agrostemmic acid (saponins)	dizziness, diarrhea, respiratory distress, vomiting, headache, sharp pains in spine, coma, death: frequent ingestion of small amounts results in chronic githagism (a disease, similar to lathyrism, that results in pain, burning and prickling sensations in lower extremities, and increasing paralysis); milled seeds may be found in wheat flour
Locoweed (*Astragalus* species); Northern Hemisphere	locoine	dullness, weakness, irregular behaviour, impaired vision, edema of eyelids, loss of muscular control, loss of appetite, emaciation, starvation, death in sheep, horses, and cattle
Belladonna (*Atropa belladonna*); United States, Europe, Asia	hyoscyamine, atropine, hyoscine, and a complex of other alkaloids	dryness of the skin, mouth, throat; difficulty in swallowing, flushing of the face, cyanosis (a bluish discoloration of skin due to insufficient oxygen), nausea, vomiting, slurred speech, coma, death; children and animals frequently poisoned by eating fruit
Akee (*Blighia sapida*)	hypoglycin A, B	sudden vomiting, drowsiness, muscular and nervous exhaustion, prostration, coma, death
Rape (*Brassica napus*)	glycosides (isothiocyanates)	pulmonary emphysema, respiratory distress, anemia, constipation, irritability, blindness in cattle
Marijuana (*Cannabis sativa*); United States, Mexico, tropical America	cannabinol, canabidiol, and related compounds	exaltation, inebriety, confusion, followed by central nervous system depression; prolonged, frequent use may produce dullness or mania; ingestion in large quantities or injection of the purified extract may produce death by cardiac depression
Water hemlock (*Cicuta maculata*); northern temperate regions	cicutoxin	abdominal pain, nausea, vomiting, diarrhea, respiratory distress, hypersalivation, convulsions, death; among the most poisonous plants
Poison hemlock (*Conium maculatum*); temperate United States, South America, northern Africa, Asia	coniine, conhydrine, *N*-methyleoniine, coniceine, and other alkaloids	muscular weakness, paralysis of extremities, blindness, respiratory paralysis, death; responsible for many human fatalities; leaves most toxic when plant is flowering
Purging croton (*Croton tiglium*); Asia, Pacific Islands, Africa	croton, croton resin, ricinine	vomiting, violent purging, collapse, death; croton oil is also a skin irritant, causing reddening, swelling, and pustules
Daphne (*Daphne mezereum*); temperate regions	glycoside involving aglycone dihydroxycoumarin	vomiting, burning sensation of the mouth, ulceration of the oral mucosa, diarrhea, stupor, weakness, convulsions, and death
Jimsonweed or thornapple (*Datura stramonium*); temperate and tropical regions	hyoscine, hyoscyamine, atropine	headache, nausea, vomiting, dizziness, thirst, dry and burning sensation in skin, mental confusion, mania, loss of memory, convulsions, death; children are often poisoned by eating seeds or sucking flowers
Larkspur (*Delphinium* species); northern temperate regions	delphinine, delphinoidine, delphisine, and other alkaloids	burning and inflammation of mouth, nausea, vomiting, respiratory distress, itching, cyanosis; one of the greatest causes of death in grazing livestock
Dumbcane (*Dieffenbachia seguine*); widely cultivated in temperate regions, tropical regions	protoanemonine, calcium oxalate	irritation and burning of the mouth, tongue, and lips; hypersalivation, swelling of the tongue, difficulty in swallowing and breathing
Foxglove (*Digitalis purpurea*); Europe, North America	glycosides, digitoxigenin, and others	loss of appetite, nausea, vomiting, slow pulse and irregular heartbeat, diarrhea, abdominal pain, headache, fatigue, drowsiness, convulsions, death
Wild yam (*Dioscorea hispida*); southern Asia, Pacific Islands	dioscorine	discomfort, then burning of the throat, giddiness, vomiting of blood, respiratory distress, drowsiness, exhaustion, paralysis of the nervous system, death; raw tubers are a frequent cause of death in the Philippines
Huanuco cocaine (*Erythroxylon coca*); tropics of both hemispheres	cocaine and other alkaloids	central nervous system stimulation followed by depression, numbness of tongue, paralysis of respiratory centres, cyanosis, respiratory distress, death; leaves are commonly chewed by Indians of Peru and Bolivia as a stimulant
Manchineel (*Hippomane mancinella*); Florida, Central America, South America, West Indies	physostigmine or a similar alkaloid plus a sapogenin	fruit causes gastroenteritis, which may be fatal, and causes ulceration of intestinal tract; sap causes burning of skin, swelling and hemorrhage of the eyes; sap is used as an arrow poison
Black henbane (*Hyoscyamus niger*); North America, Europe, Asia, Oceania	hyoscyamine, hyoscine, atropine, and other alkaloids	similar to belladonna poisoning caused by *Atropa belladonna;* children are poisoned by eating seeds and pods

Amanita. Wild mushrooms should be eaten only if they have been accurately identified by an experienced person; the safest procedure is to eat only cultivated species. The problem of toxicity in mushrooms is complex; no single rule or test method exists by which the toxicity of a mushroom can be determined. The most poisonous species closely resemble some of the most prized edible species; in addition, toxicity within a given wild species may vary from one set of ecological conditions or from one geographical locality to the next. Moreover, although some mushrooms that are poisonous when fresh are edible when cooked, dried, salted, or preserved in some other way, others remain poisonous in spite of all preparation procedures. It has also been observed that some people may become poisoned by eating mushrooms that apparently do not affect others. As with microfungi, the mushroom poisons vary in their chemical and biological properties from species to species.

Protistan poisons. The dinoflagellates, important producers of the primary food supply of the sea, are microscopic one-celled organisms that are dependent upon various inorganic nutrients in the water and upon radiant energy for photosynthesis, the process by which they produce their own food supplies. Although dinoflagellates inhabit both marine waters and freshwaters, most species are marine. Dinoflagellates are most often found in cool or temperate waters. During periods of planktonic blooms (times of high concentrations of microscopic organisms in the water) dinoflagellates multiply in large numbers. These planktonic blooms, sometimes referred to as red tide because they discolour the water, are often associated with weather disturbances that may bring about changes in water masses or upwellings. During periods of bloom large numbers of toxic dinoflagellates may be ingested by shellfish; the poisons accumulate in their digestive glands. Animals and humans may in turn be poisoned by eating poisoned shellfish. Certain species of dinoflagellates are capable of producing some of the most toxic substances known. The two species of dinoflagellates most commonly involved in human intoxications have been *Gonyaulax catenella* along the Pacific coast of North America and *G. tamarensis* along the eastern coast of North America. Intoxications from these organisms are known as paralytic shellfish poisoning. The symptoms, which begin with a tingling or burning sensation, then numbness of the lips, gums, tongue, and face, gradually spread. Gastrointestinal upset may be present. Other symptoms include weakness, joint aches, and muscular paralysis; death may result. There is no specific treatment or antidote. The poison, variously called paralytic shellfish poison, mussel poison, and saxitoxin, is a complex nonprotein nitrogen-containing compound. Paralytic shellfish poisoning is best avoided by following local public-health quarantine regulations.

Symptoms of paralytic shellfish poisoning

Respiratory irritation may result from the inhalation of toxic products in the windblown spray from red-tide areas containing the toxic dinoflagellate *Gymnodinium breve,* which is found in the Gulf of Mexico and Florida; the nature of the poison is unknown. Deaths of large numbers of brackish-water pond fishes because of *Prymnesium parvum* have been reported in Israel; the poison is known as prymnesin.

Plant poisons (phytotoxins). The study of plant poisons

Table 6: Representative Poisonous Plants (continued)

name and distribution	toxic principle	toxic effects and comments
Barbados nut (*Jatropha cucas*); tropics	curcin	burning of the throat, bloating, dizziness, vomiting, diarrhea, drowsiness, dysuria, leg cramps, violent purgative action; may be fatal to children
Mountain laurel (*Kalmia latifolia*); North America	andromedotoxin	hypersalivation, tears, impaired vision, tingling of skin, dizziness, vomiting, muscular paralysis, convulsions, coma, death; children are poisoned by eating leaves
Grass peavine (*Lathyrus sativus*); North America, Europe, northern Africa, Asia	β-aminopropionitrile	back pain, weakness in legs, paralysis; has caused death in children
Cassava (*Manihot esculenta*); tropics	cyanophoric glycosides	nausea, respiratory distress, twitching, staggering, convulsions, coma, death
Chinaberry (*Melia azedarach*); North America, southern Africa, Asia	azadarin	stomatitis with violent and bloody vomiting, paralysis
Opium poppy (*Papaver somniferum*); Europe, Asia, tropics	morphine, codeine, thebaine, papavarine, narcotine	central nervous system depression, pinpoint pupils, depressed respiration, cyanosis, coma, death
Pokeberry (*Phytolacca americana*); North America, Europe, southern Africa	phytolaccine	burning, bitterness in mouth, vomiting, purging, spasms, convulsions, death
Castor bean (*Ricinus communis*); United States, tropics	ricin, a toxalbumin	burning of mouth, throat, and stomach, vomiting, diarrhea, abdominal cramps, dulled vision, convulsions, respiratory distress, paralysis, death; one to three seeds may be fatal to children
Black nightshade (*Solanum nigrum*); North America, Europe	solanine, a glycoalkaloid	nausea, vomiting, abdominal pain, diarrhea, trembling, paralysis, coma, death
Plants poisonous by contact		
Euphorbia, spurge (*Euphorbia* species); worldwide	a complex of substances including alkaloids, glycosides, and others	eye irritation, blindness, blistering of the skin, swelling around the mouth, burning of the mouth, unconsciousness, death; milky sap is used as an arrow poison
Spurge nettle (*Jatropha stimulosus*); North America, Europe, Asia	toxin unknown	contact produces instant, intense stinging and itching due to an irritating substance injected into the skin by the stinging hairs; results in a skin eruption of minute red papular (small conical elevations of the skin) rash, which lasts about 30 minutes; a dull purplish discoloration of the skin may remain for several weeks
Shiney-leaf stinging tree, tree nettle (*Laportea photiniphylla*); Australia	5-hydroxytryptamine (and other toxic substances?)	contact with the stinging hairs of this plant produces intense, rapidly spreading pain, reddened rash, and later a severe skin eruption; severe stings may result in intense, unbearable pain; fatalities have been reported; dried leaves may cause intense sneezing
Poisonwood (*Metopium toxiferum*); West Indies, Florida	similar to poison ivy	contact with any part of the tree, especially sap, turns the skin black, causes a rash, blisters, etc.; smoke from a burning tree is very irritating, causing illness and temporary blindness
Strophanthus (*Strophanthus* species); Florida, tropical America, Africa	an alkaloid, trigonelline, and a large number of cardiac glycosides and aglycones	vomiting, slow and irregular pulse, blurred vision, delirium, circulatory failure, death; used as an arrow poison
Curare (*Strychnos toxifera*); Central America and northern South America	toxiferines, caracurines, and other alkaloids	haziness of vision, relaxation of facial muscles, inability to raise head, loss of muscle control of arms, legs, and respiratory muscles, death; used as a poison for arrows and for blowgun darts
Poison ivy (*Toxicodendron radicans*, also called *Rhus toxicodendron*); North America	urushiol	skin irritation, swelling, blistering, itching; may be fatal to young children; smoke from burning plant is toxic
Plants that produce photosensitization		
Buckwheat (*Fagopyrum sagittatum*); North America, Europe	fagopyrin, a naphthodianthrone derivative	ingestion of the leaves by animals causes liver dysfunction, thereby resulting in deposition of a photosensitizing pigment in the skin; sunlight then causes redness of the skin, nervousness, swelling of the eyelids, convulsions, and prostration in farm animals
St. Johnswort (*Hypericum perforatum*); North America, Europe	hypericin, a naphthodianthrone derivative	similar to buckwheat
Plants that produce airborne allergies		
Box elder (*Acer negundo*); Northern Hemisphere	oleoresin and a water-soluble antigen	hay fever (respiratory allergy), may also cause an eczematous dermatitis of the exposed parts of the body

is known as phytotoxicology. Most of the poisonous higher plants are angiosperms, or flowering plants, but only a small percentage are recognized as poisonous. Several systems have been devised for the classification of poisonous plants, none of which is completely satisfactory. Poisonous plants may be classified according to the chemical nature of their toxic constituents, their phylogenetic relationship, or their botanical characteristics. The following classification, which is based on their toxic effects, has been found to be useful: (1) plants that are poisonous to eat, (2) plants that are poisonous upon contact, (3) plants that produce photosensitization, and (4) plants that produce airborne allergies (see Table 6).

Chemical variability of plant poisons

Plant poisons, or phytotoxins, comprise a vast range of biologically active chemical substances, such as alkaloids, polypeptides, amines, glycosides, oxalates, resins, toxalbumins, and a large group of miscellaneous compounds whose chemical structure has not yet been determined. Alkaloids, most of which are found in plants, are characterized by the presence of nitrogen and their ability to combine with acids to form salts. They are usually bitter in taste. It has been estimated that about 10 percent of the plant species contain some type of alkaloid. Only a few of the 5,000 alkaloids characterized thus far do not produce any biological activity; most cause a strong physiological reaction when administered to an animal. Amines are organic compounds containing nitrogen. A polypeptide is a string of three or more amino acids. A few polypeptides and amines are toxic to animals. Some glycosides, which are compounds that yield one or more sugars and one or more other compounds—aglycones (nonsugars)—when hydrolyzed (chemically degraded by the introduction of water molecules between adjacent subunits), are extremely toxic to animals. Toxicity resides in the aglycone component or a part of it. Oxalates are salts of oxalic acid, which under natural conditions is not toxic but becomes so because of the oxalate ion. Resins, a heterogeneous assemblage of complex compounds, differ widely in chemical properties but have certain similar physical properties. Some resins are physiologically very active, causing irritation to nervous and muscle tissue. Toxalbumins are highly toxic protein molecules that are produced by only a small number of plants. Ricin, a toxalbumin from the castor bean (*Ricinus communis*), is one of the most toxic substances known.

Under certain ecological conditions plants may become poisonous as a result of the accumulation of toxic inorganic minerals such as copper, lead, cadmium, fluorine, manganese, nitrates, or selenium. Photosensitization, an unusual toxic reaction resulting from the ingestion of certain plants, may be of two types. The toxic substance may be obtained directly from the plant, which thereupon acts on the skin (primary photosensitivity), or the toxicity may result from liver damage caused by the metabolism of a toxic plant and failure of the breakdown products to be eliminated by the liver (hepatic photosensitivity). In either case the animal reacts by becoming restless; in addition, the skin reddens, and a severe sloughing of the skin develops. Death seldom occurs.

A large number of poisonous plants occur throughout the world; a few representative species and their poisons are listed in Table 6.

Animal poisons (zootoxins). Poisonous animals are widely distributed throughout the animal kingdom; the

Table 7: Representative Animals Poisonous When Eaten

name and distribution	toxic principle	toxic effects and comments
Protozoans—one-celled animals		
Dinoflagellate (*Gymnodinium breve*); Gulf of Mexico, Florida	unknown	irritation of mucous membranes of nose and throat; causes sneezing, coughing, respiratory distress due to inhalation of windblown spray from red tide areas
Dinoflagellate (*Gonyaulax catenella*); Pacific coast of North America	paralytic shellfish poison, saxitoxin ($C_{10}H_{17}N_7O_3 \cdot 2$ HCl)	tingling, burning sensation and numbness of lips, tongue, face, spreading elsewhere in the body; weakness, dizziness, joint aches, hypersalivation, intense thirst, difficulty in swallowing, muscular paralysis, and death; extremely toxic; usually involved with the eating of shellfish that have been feeding on toxic dinoflagellates
Mollusks—octopus, squid, shellfish, and others		
California mussel (*Mytilus californianus*); Pacific coast of North America	paralytic shellfish poison, saxitoxin	these mollusks become poisonous to eat because of feeding on toxic dinoflagellates; symptoms same as for dinoflagellates (*Gonyaulax* species)
Butter clam (*Saxidomus giganteus*); Alaska to California	same as California mussel	
Whelk (*Neptunea* species); Europe, Pacific region	tetramine	nausea, vomiting, diarrhea, weakness, fatigue; dizziness, photophobia (intolerance to light), impaired vision, and dryness of the mouth; poison is believed to be restricted to the salivary glands of the whelk
Turban shell (*Turbo argyrostoma*); tropical Pacific Ocean	poison believed to be related to ciguatoxin	diarrhea, weakness of the legs, fatigue; cold water produces a painful stinging sensation, itching; the illness closely resembles ciguatera fish poisoning
Callista shellfish (*Callista brevisiphonata*); Japan	a histamine-like substance, choline	flushing of the face, itching, urticaria (stinging sensation of skin), sensation of constriction of the chest, abdominal pain, nausea, respiratory distress, asthmatic attacks, paralysis, hypersalivation, numbness of the tongue, throat; recovery usually within 10 days
Arthropods—joint-legged animals: crabs, spiders, and others		
Shore crab (*Demania toxica*); Philippines	unknown	nausea, vomiting, diarrhea, muscular weakness, respiratory distress, difficulty in speaking, hypersalivation, muscular paralysis, convulsions, death
Crab (*Zozymus aeneus*); Indo-Pacific	similar to tetrodotoxin; toxicity of these crabs variable	tingling about the mouth, nausea, vomiting, muscular paralysis, coma, convulsions, death
Asiatic horseshoe crab (*Tachypleus tridentatus*); Southeast Asia	unknown	dizziness, headache, nausea, vomiting, abdominal pain, cardiac palpitation, numbness of the lips, weakness, muscular paralysis, hypersalivation, loss of consciousness, death
Sharks, eels, and other fish		
Greenland shark (*Somniosus microcephalus*); Arctic	unknown; flesh toxic; liver of tropical sharks very toxic and may also cause death	nausea, vomiting, diarrhea, abdominal pain, tingling and burning sensation of the tongue, throat, and esophagus, muscular cramps, respiratory distress, coma, death
Moray eel (*Gymnothorax javanicus*); Indo-Pacific†	ciguatoxin*	symptoms may develop rapidly or slowly; tingling about the lips, tongue, and throat, followed by numbness, nausea, vomiting, abdominal cramps, muscular weakness, paralysis, convulsions, teeth feeling loose, visual impairment, skin rash, hot objects feeling cold and vice versa ("Dry Ice" or "electric shock" sensation); loss of muscular coordination, coma, death in about 12 percent of the cases; known as ciguatera fish poisoning, this is one of the most common forms of fish poisoning
Red snapper (*Lutjanus bohar*); Indo-Pacific†	same as Moray eel	

*Fish poisoning is categorically referred to as "ichthyosarcotoxism," but there are several different forms of fish poisoning, such as ciguatera fish poisoning, clupeotoxism, scombrotoxism, and others. †More than 400 species of tropical reef fishes have been involved in ciguatera fish poisoning. These fish are normally edible but under certain conditions may become toxic.

only major group that seems to be exempt is the birds.

Categories of zootoxins

Zootoxins can be divided into several categories: (1) oral poisons—those that are poisonous when eaten; (2) parenteral poisons, or venoms—those that are produced by a specialized poison gland and administered by means of a venom apparatus; and (3) crinotoxins—those that are produced by a specialized poison gland but are merely released into the environment, usually by means of a pore.

Oral zootoxins (see Table 7) are generally thought to be small molecules; most venoms (Table 8) are believed to be large molecules, usually a protein or a substance in close association with one. Venoms, which are produced by specialized poison glands, are injected by means of a mechanical device that is able to penetrate the flesh of the victim. Little is known about the biological or chemical properties of most crinotoxins (Table 9). The term poisonous may be used in the generic sense to refer to all three categories of zootoxins.

Some of the most complex relationships in biotoxicology are found in the marine environment. Certain marine biotoxins, such as ciguatera fish poison, apparently originate in marine plants, are ingested by herbivores and then passed on to carnivores and eventually to humans. The extremely complex mechanism by which this is accomplished is not clear. With the buildup of toxic industrial chemical pollutants in the marine environment, the problems of toxicity in marine organisms are becoming increasingly more serious. There is evidence that under certain conditions chemical pollutants may trigger biotoxicity cycles in marine organisms. The outbreaks in Japan of Minamata disease were the result of such a cycle: microorganisms, algae, shellfishes, and fishes ingested or absorbed industrial wastes with highly toxic organic compounds containing mercury and were in turn consumed by humans, causing a number of deaths among the population.

The relationships of representative poisonous animals and their position in the total framework of the animal kingdom can best be appreciated by categorizing them according to the group in which they belong. They are further grouped as to whether they are poisonous to eat, venomous, or crinotoxic in Tables 7, 8, and 9.

(B.W.H./Ed.)

RADIATION

Radiation, radioactivity, and radioisotopes. Radiation is a flow of energy through space or matter. It takes the form of particles (*e.g.*, alpha and beta particles) or electromagnetic waves (*e.g.*, X rays, gamma rays, and visible and ultraviolet [UV] light). Radiation can be classified as either ionizing or nonionizing depending on its ability to produce ions in the matter it interacts with. Ionizing radiation is more toxic than nonionizing radiation.

Radioactivity is the emission of radiation caused by the disintegration of unstable nuclei of radioisotopes. After disintegration, a radioisotope may become a radioisotope of another element, which will further disintegrate. The disintegration series continues until a stable isotope is formed.

Ionizing radiation. Ionizing radiation is radiation that produces ions in matter during interaction with atoms in the matter. The toxic effect of ionizing radiation is related to the ionization. It is believed that ionization of tissues, composed mainly of water, generates H_2O^+ and H_2O^- ions, which in turn form H and OH radicals. Because radicals are very reactive chemically, biological damage, such as attacks on DNA and proteins, results.

There are two classes of ionizing radiation: particulate and eletromagnetic. Alpha particles, beta particles, neutrons, and positrons are examples of particulate ionizing radiation. Gamma rays and X rays are electromagnetic ionizing radiation.

Among particulate ionizing radiation, alpha and beta particles are the forms most commonly encountered in the environment and are biologically the most significant.

Table 7: Representative Animals Poisonous When Eaten (continued)

name and distribution	toxic principle	toxic effects and comments
Thread herring (*Clupanodon thrissa*); Indo-Pacific	ciupeotoxin, chemical nature unknown	metallic taste, nausea, vomiting, abdominal pain, vascular collapse, hypersalivation, numbness, muscular paralysis, convulsions, coma, death; this form of poisoning develops rapidly and violently; mortality rate is about 50 percent and death may come within a few minutes; this form of fish poisoning is known as clupeotoxism; no known antidote
Castor-oil fish (*Ruvettus pretiosus*); tropical Atlantic, Indo-Pacific	oleic acid	produces a painless diarrhea; poisoning known as gempylotoxism; no treatment needed
Skipjack tuna (*Euthynnus pelamis*); tropical seas Bluefin tuna (*Thunnus thynnus*); subtropical and temperate seas	saurine, a histamine-like substance; scombroid fishes (mackerels, tunas, swordfishes, and allies) contain a chemical constituent in their flesh called histidine; when histidine is acted upon, it forms a histamine-like substance called saurine; this occurs when the fishes are permitted to stand at room temperature for several hours; scombroid fishes are more susceptible to the development of saurine poison than most other kinds of fishes	the symptoms of acute scombroid poisoning resemble those of severe allergy: sharp, peppery taste, headache, throbbing of the large blood vessels of the neck, nausea, vomiting, massive red welts, and intense itching; recovery after 8-12 hours; this is probably the most common and cosmopolitan form of fish poisoning; antihistamines are used for treatment
Deadly death puffer (*Arothron hispidus*); tropical Pacific, Indian Ocean, Red Sea	tetrodotoxin ($C_{11}H_{17}O_8N_3 \pm$ y H_2O)	tingling of lips and tongue, loss of motor coordination, floating sensation, hypersalivation, numbness of the entire body, muscular paralysis, difficulty in swallowing, weakness, nausea, vomiting, convulsions, about 60 percent fatality in humans; no known antidote
Amphibians		
California newt (*Taricha torosa*); California	tarichatoxin; poison said to be identical to tetrodotoxin; the eggs of this newt are extremely toxic	effects in humans are unknown; no known antidote
Reptiles		
Hawksbill turtle (*Eretmochelys imbricata*); tropical seas Leatherback turtle (*Dermochelys coriacea*); temperate and tropical seas	chelonitoxin, chemistry unknown; the flesh of some species of marine turtles is extremely poisonous	nausea, vomiting, diarrhea; burning sensation of lips, tongue, mouth; tightness of the chest, difficulty in swallowing, hypersalivation, foul breath, skin rash, sloughing of the skin, enlargement of the liver, coma, death; fatality rate is high; no known antidote
Mammals		
Sei whale (*Balaenoptera borealis*); North Pacific and North Atlantic oceans	unknown; livers of many marine mammals are toxic	intense headache, neck pain, photophobia, desquamation (peeling in scales) around the mouth and face, flushing of the face; antihistamines are used in treatment
White whale (*Delphinapterus leucas*); Arctic seas	unknown	flesh is poisonous and has caused fatalities in humans; no known antidote
Polar bear (*Thalarctos maritimus*); Arctic	vitamin A and possibly other toxic substances	intense throbbing headaches, nausea, vomiting, diarrhea, abdominal pain, dizziness, drowsiness, irritability, muscle cramps, visual disturbances, collapse, coma, rarely death

Composed of two neutrons and two protons and thus containing a 2^+ charge, alpha particles are the heaviest ionizing particles. Although they do not penetrate tissue very well, alpha particles turn many atoms in their short paths into ions, producing intense tissue ionization.

In contrast to alpha particles, beta particles are electrons of little mass carrying only one negative charge. They penetrate up to several millimetres in soft tissues. Their low mass and low charge mean that only moderate ionization is produced in tissues when beta particles collide with atoms in its path.

Gamma rays and X rays are electromagnetic radiation of similar properties, with gamma rays having higher energy than X rays. Gamma rays usually accompany the formation of alpha or beta particles. Neither gamma rays nor X rays carry a charge, and neither have mass; consequently, they can penetrate tissues easily, creating moderate ionization along their paths.

Biological damage is related to the degree of tissue ionization produced by radiation. Thus, a physical dose of alpha particles does not produce the same amount of damage as that produced by the same dose of beta particles, gamma rays, or X rays.

Radiation sources. Radiation is either natural or man-made. Natural radiation includes cosmic radiation, terrestrial radiation, radioisotopes inside human bodies, and radon gas. Cosmic radiation consists of charged particles from outer space, and terrestrial radiation of gamma rays from radionuclides in the Earth. Radioisotopes in human bodies come from the food, water, and air consumed. Cosmic and terrestrial radiation, together with radioisotopes inside human bodies, contribute only one-third of the total natural radiation dose. The remaining two-thirds can be attributed to radon, a radioactive gas released from soil that may reach a high level inside buildings with poor ventilation. Man-made radiation consists of radiation from medical and dental diagnostic procedures, atmospheric tests of atomic bombs, emissions from nuclear plants, certain occupational activities, and some consumer products. The largest nonoccupational radiation sources are tobacco smoke for smokers and indoor radon gas for the nonsmoking population.

Emissions from nuclear power plants contribute only a very small portion of the total yearly radiation received. The low dose reflects the negligible amount of radionuclides released during normal operation, although the amount released can be much higher after a nuclear reactor accident. Not every reactor accident is a disaster, however. The 1979 accident at the Three Mile Island nuclear power station, near Harrisburg, Pa., released only a small amount of radiation (0.8 and 0.015 mSv within a 16- and 80-km radius, respectively), less than the background annual radiation dose. The nuclear reactor accident at Chernobyl in the Soviet Union, in 1986, however, was much more devastating, leading to more than 30 deaths and the evacuation of thousands of nearby residents.

Adverse effects of ionizing radiation. Ionizing radiation quickly kills rapidly dividing cells. In general, immature blood cells in bone marrow, cells lining the mucosa of the gastrointestinal tract, and cells in the lower layers of the epidermis and in hair follicles are the most rapidly dividing cells in the body. As a result, radiation leads to the decreased production of blood cells, nausea, vomiting, diarrhea, malabsorption by the intestine, skin burns, and hair loss. Because of its relatively selective lethal effect on rapidly dividing cells, however, ionizing radiation is used in the treatment of certain cancers. Some cells in the embryo and fetus also divide rapidly, and thus ionizing radiation can cause malformations and even fetal death. Ionizing radiation can also produce mutations by altering the DNA, and it can result in cancer.

Toxicities of whole-body ionizing radiation. X rays and gamma rays irradiate the body uniformly and acutely affect all of the tissues discussed above. At sufficiently high doses, this type of radiation can lead to a condition known as acute radiation syndrome. The most sensitive tissue is the bone marrow, where blood cells are generated. The next tissue affected is the gastrointestinal tract. If the dose is high, the central nervous system is affected and the person becomes uncoordinated and disoriented

Table 8: Representative Venomous Animals That Inflict a Sting

name and distribution	toxic principle	toxic effects and comments
Cnidarians		
Portuguese man-of-war (*Physalia* species); tropical seas	tetramine, 5-hydroxytryptamine	immediate, intense stinging, throbbing, or burning sensation, shooting sensation, inflammatory rash, blistering of the skin, shock, collapse, in very rare cases death
Sea wasp (*Chironex fleckeri*); northern and northeast Australia	cardiotoxin	immediate, extremely painful stinging sensation, seared reddened lines wherever the tentacles touch the skin, large indurated wheallike lesions, prostration, dizziness, circulatory failure, respiratory distress, rapid death in a high percentage of cases
Sea anemone (*Actinia equina*); Mediterranean, Black Sea, etc.	nature of venom unknown	burning, stinging sensation, itching, swelling, redness, ulceration, nausea, vomiting, prostration; no specific antidote available
Mollusks		
Cone shell (*Conus* species); tropical Indo-Pacific region	quaternary ammonium compounds and others	blanching at the site of injection, cyanosis of the surrounding area, numbness, stinging or burning sensation, blurring of vision, loss of speech, difficulty in swallowing, nausea, extreme weakness, coma, and death in some cases; no specific antidote
Spotted octopus (*Octopus maculosus*); Indo-Pacific, Indian Ocean	cephalotoxin, a neuromuscular poison	sharp burning pain, similar to a bee sting, numbness of the mouth and tongue, blurring of vision, loss of tactile sensation, difficulty in speech and swallowing, paralysis of legs, nausea, prostration, coma, and death in a high percentage of cases
Arthropods		
Kissing bug (*Triatoma* species); Latin America, United States	unknown	bite usually painless; later itching, edema about the bite, nausea, palpitation, redness; the bite is of relatively minor importance but spreads Chagas disease caused by a trypanosome (protozoan)
Puss caterpillar (*Megalopyge* species); United States, Latin America	unknown	stinging hairs of the caterpillar associated with poison-secreting glands; contact with the hairs produces an intense burning pain, itching, pustules, redness, nausea, fever, numbness, swelling, and paralysis; recovery usually within about six days
Honeybee (*Apis* species); worldwide	neurotoxin, hemolytic, melittin, hyaluronidase, phospholipase A, histamine, and others	sting produces acute local pain or burning sensation, blanching at the site of the sting surrounded by a zone of redness, and itching; local symptoms usually disappear after 24 hours; severe cases may develop massive swelling, redness, shock, prostration, vomiting, rapid heartbeat, respiratory distress, trembling, coma, and death; it is estimated that 500 stings delivered in a short period of time can provide a lethal dose to a human; bees kill more people in the United States than do venomous reptiles
Bumblebee (*Bombus* species); temperate regions	similar to honeybee (*Apis*) venom	stings are similar to honeybee (*Apis*) stings; bumblebees are not as vicious as honeybees
Yellow jacket, hornet (*Vespula* species); temperate regions	similar to bee venom; also acetylcholine	they can both bite and sting; the sting is similar to that of the honeybee's but more painful; yellow jackets are quite aggressive; stings may be fatal
Wasp (*Polistes* species and *Vespa* species); worldwide	similar to bee venom; also acetylcholine	wasps are less aggressive than hornets, and their sting is similar to the honeybee's but generally less painful than the hornet's; stings may be fatal
Harvester ant (*Pogonomyrmex* species); United States	bradykinin, formic acid, hyaluronidase, hemolytic, phospholipase A, and others	ant stings cause immediate intense burning, pain, blanched area at site of sting surrounded by redness, ulceration, fever, blistering, itching, hemorrhaging into the skin, eczematoid dermatitis, pustules, respiratory distress, prostration, coma, and death in some instances
Fire ant (*Solenopsis* species); United States, Latin America	similar to harvester-ant venom	similar to above; stings are very painful, burning sensation, etc.
Millipede (*Apheloria* species and others); temperate areas	hydrogen cyanide and benzaldehyde	toxic liquid or gas from lateral glands causes inflammation, swelling, and blindness in contact with eyes, and brown stain, redness, swelling, and vesicle formation in contact with skin
Centipede (*Scolopendra* species); temperate and tropical regions	hemolytic phospholipase and serotonin	local pain, swelling, and redness at bite site
Brown spider (*Loxosceles* species); United States, South America, Europe, Asia	cytotoxic, hyaluronidase, hemolytic, and others	bite causes stinging sensation or burning pain, blanching at site of bite surrounded by redness, blistering, hemorrhages into the skin and internal organs, ulceration, vomiting, fever, cardiovascular collapse, convulsions, sometimes death
Black widow (*Latrodectus* species); tropical and temperate regions	neurotoxic	bite may or may not be painful, two tiny red dots at site of bite, localized swelling after a few minutes; intense cramping pain of abdomen, legs, chest, back; rigidity of muscles lasting 12-48 hours, nausea, sweating, respiratory distress, priapism (abnormal, painful erection of the penis) in males, chills, skin rash, restlessness, fever, numbness, tingling; about 4 percent are fatal; antiserum is available
Tarantula (*Dugesiella* and *Lycosa* species); temperate and tropical regions	venom varies, usually mild	most of the large tarantulas found in the United States, Mexico, and Central America are harmless to humans; some of the large tropical species may be more poisonous, but their effects are largely localized
Scorpion (species of *Centruroides, Tityus,* and *Leiurus*); warm temperate and tropical regions	neurotoxin, cardiotoxin, hemolytic, lecithinase, hyaluronidase, and others	symptoms vary depending upon the species of scorpion; sting from the tail stinger causes a sharp burning sensation, swelling, sweating, restlessness, salivation, confusion, vomiting, abdominal pain, chest pain, numbness, muscular twitching, respiratory distress, convulsions, death; the mortality rate from stings from certain species of scorpions is very high; antiserum is available
Echinoderms		
Crown-of-thorns starfish (*Acanthaster planci*); Indo-Pacific	nature of poison unknown	penetration of the spines produces a painful wound, redness, swelling, vomiting, numbness, and paralysis
Long-spined sea urchin (*Diadema setosum*); Indo-Pacific	nature of poison unknown	penetration of the spines produces an immediate and intense burning sensation, redness, swelling, numbness, muscular paralysis
Sea urchin (*Toxopneustes pileolus*); Indo-Pacific	nature of poison unknown	bites from the stinging jaws or pedicellariae (small pincerlike organs) produce an immediate, intense, radiating pain, faintness, numbness, muscular paralysis, respiratory distress, and occasionally death
Sharks and rays		
Stingray (*Dasyatis* species); warm temperate and tropical seas	stingray venom, cardiotoxin, chemistry unknown	penetration of the tail spines inflicts jagged painful wounds that produce sharp, shooting, throbbing pain, fall in blood pressure, nausea, vomiting, cardiac failure, muscular paralysis, rarely death; no known antidote; stingrays are among the most common causes of envenomations in the marine environment
Bony fishes		
Weever fish (*Trachinus draco*); Mediterranean Sea	weever fish venom, chemistry unknown	opercular and dorsal fin spines can produce instant pain, burning, stabbing or crushing sensation; pain spreads and becomes progressively more intense, causing the victim to scream with anguish and suffer loss of consciousness, numbness about the wound, swelling, redness, nausea, delirium, difficulty in breathing, convulsions, and death; no known antidote
Scorpion fish (*Scorpaena* species); temperate and tropical seas	scorpion fish venom, chemistry unknown	fin spines can inflict painful stings and intense, immediate pain that may cause victim to scream followed by redness, swelling, loss of consciousness, ulceration of the wound, paralysis, cardiac failure, delirium, convulsions, nausea, prostration, and respiratory distress, but rarely death; no known antidote
Stonefish (*Synanceja* species); Indo-Pacific region	stonefish venom, chemistry unknown	produces an extremely painful sting by means of the dorsal fin spines; symptoms similar to other scorpion fish stings but more serious
Reptiles		
Gila monster (*Heloderma suspectum*); southwestern United States	heloderma venom, primarily a neurotoxin	all of the teeth are venomous; bite causes local pain, swelling, weakness, ringing of the ears, nausea, respiratory distress, cardiac failure; may cause death; no antiserum available

Table 8: Representative Venomous Animals That Inflict a Sting (continued)

name and distribution	toxic principle	toxic effects and comments
Rear-fanged snakes		
Boomslang (*Dispholidus typus*); Africa	coagulates the blood; proteolytic and other enzymes are present	local pain, swelling, hemorrhages; bleeding from nose, mouth, and skin; headache, vomiting, collapse, death; antiserum is available
Front-fixed-fanged snakes		
Death adder (*Acanthophis antarcticus*); Australia, New Guinea	coagulates blood; proteolytic and other enzymes present	effects of envenomation develop slowly (15–60 minutes); nausea, vomiting, faintness, drowsiness, staggering, slurred speech, respiratory distress, hemorrhages, death; antiserum available
Indian krait (*Bungarus candidus caeruleus*); India, Myanmar, Malaya, Indonesia	cholinesterase, protease, and other enzymes present	little pain or local reaction, latent period up to 12 hours; abdominal pain, staggering gait, difficulty in swallowing, stiffness of jaws, coma, respiratory distress, cardiac failure, death; antiserum available
Brown snake (*Demansia textilis*); Australia	cholinesterase, ophio-adenosinetri-phosphatase, phospholipase, and other enzymes	latent period up to 12 hours; abdominal pain, vomiting, headache, dizziness, blood in the urine, weakness, respiratory and circulatory collapse, death; antiserum available; this snake is responsible for most of the deaths due to snake bite in Australia; very dangerous
Black mamba (*Dendroaspis polylepis*); Ethiopia, Somalia, Natal, southwest Africa	cholinesterase, L-amino-acid oxidase, ophio-adenosine, triphosphatase, and other substances	local pain, swelling, salivation, paralysis of vocal cords, sweating, vomiting, restlessness, drowsiness, collapse, coma, respiratory distress, death; one of the most dangerous of all living snakes; antiserum available; a large black mamba secretes enough venom to kill 10 men
Coral snake (*Micrurus corallinus*); subtropical South America	cholinesterase, hyaluronidase, L-amino-acid oxidase, and other enzymes	numbness without pain at bite, headache, swelling of face and lips, hyperesthesia (unusual sensitivity of skin), sore throat, drooping of eyelids, photophobia, vomiting, rapid heart rate, backache, irritability, salivation, death; there are about 40 species in the genus *Micrurus;* all are dangerous; antiserum available for some species but not others
India cobra (*Naja naja*); southern Asia, Indonesia, Taiwan, Philippines	carboxypolypeptidase, cholinesterase, diastase, dipeptidase, polypeptidase, protease, and other enzymes; one of the most complex snake venoms	pain radiating from site of bite, edema, numbness, drooping of the eyelids and head, salivation, difficulty in speech, muscular incoordination, weakness, respiratory distress, blindness, incontinence, convulsions, death; antiserum available
Black-necked cobra (*Naja nigricollis*); Africa	cholinesterase, L-amino-acid oxidase, phospholipase, protease, and other enzymes	similar to poisonings by the Indian cobra; venom sometimes sprayed at eyes; produces an intense irritation, but permanent damage to eyes is rare; antiserum available
Tiger snake (*Notechis scutatus*); Australia	coagulates blood; cholinesterase, hyaluronidase, ophio-adenosine-triphosphatase, phosphatase, and other enzymes	latent period of 15–60 minutes; nausea, vomiting, faintness, sweating, drowsiness, staggering, slurred speech, difficulty in swallowing, drooping of eyelids, respiratory distress, death; antiserum available
King cobra (*Ophiophagus hannah*); India, Myanmar, Philippines, Thailand	cholinesterase, L-amino-acid oxidase, ophio-adenosinetriphosphotase	symptoms similar to those produced by other cobras; symptoms develop rapidly, death often occurs in 30–60 minutes; one of the largest of venomous snakes—obtains a length of 16 feet; antiserum available
Taipan (*Oxyuranus scutellatus*); Australia, New Guinea	lecithinase, phospholipase	symptoms similar to those produced by the tiger snake; its long fangs and potent venom make this one of the world's most dangerous snakes; produces flaccid paralysis, respiratory paralysis, often rapidly fatal; antiserum available
Australian black snake (*Pseudechis porphyriacus*); Australia	L-amino-acid oxidase, lecithinase, phospholipase	local pain and swelling, vomiting, hemorrhages from nose and mouth, prostration, hematuria (presence of blood in the urine), and sometimes death; usually the bites are nonfatal; antiserum available
Sea snakes		
Beaked sea snake (*Enhydrina schistosa*); Indo-Pacific	lecithinase	no local reaction; latent period minutes to hours; giddiness, muscle ache, muscular weakness, drooping of the eyelids, lockjaw, respiratory failure, acute renal (kidney) failure, death; all sea snakes are venomous, but only a few species have caused death in man
Erectile-fanged snakes		
Copperhead (*Ancistrodon contortrix*); eastern and southern United States	cholinesterase, L-amino-acid oxidase, phosphodiesterase, and other enzymes	local pain, swelling, necrosis, nausea, vomiting, shock, petechiae (minute bleeding spots), bloody stools, death; antiserum available
Eastern cottonmouth (*Ancistrodon piscivorus*); southeastern United States to central Texas	cephalinase, cholinesterase, hyaluronidase, lacithinase, and other enzymes	similar to poisoning by the copperhead but more severe; local necrosis more marked; death; antiserum available
Fer-de-lance (*Bothrops atrox*); central Mexico, south into South America	cholinesterase, deoxyribonuclease, L-amino-acid oxidase, protease, and other enzymes	local pain, bleeding from bite, gums, nose, mouth, rectum; blood fails to clot; shock, respiratory distress, hemorrhages into muscles and nervous system, death; antiserum available
Eastern diamondhead rattlesnake (*Crotalus adamanteus*); southeastern United States	cephalinase, cholinesterase, hyaluronidase, and other enzymes	local pain, edema, hemorrhaging into the tissues, dryness of the mouth, vomiting, shock, anemia, tingling sensation, blood in feces, difficulty in speech, yellow vision, loss of consciousness, death; antiserum available
Western diamondback rattlesnake (*Crotalus atrox*); southwestern United States	bradykininogen, cholinesterase, hyaluronidase, L-amino-acid oxidase, and other substances	symptoms similar to poisoning by *C. adamanteus,* but neurological symptoms less marked; antiserum available
Cascabel (*Crotalus durissus terrificus*); southern Mexico south to Argentina	cholinesterase, deoxyribonuclease, hyaluronidase, lecithinase, and other enzymes	symptoms similar to cobra bites, except that no edema occurs; antiserum available
Bushmaster (*Lachesis muta*); Costa Rica to northern South America	protease	very few bites recorded and little is known concerning the effects of the venom; severe prostration, rapid death; antiserum available
Taiwan habu (*Trimereurus mucrosquamatus*); southeastern China, Taiwan	neurotoxin, amino-acid oxidase, hemorrhagin, and other substances	local pain, blistering, bleeding into the tissues, shock, death; antiserum available
True vipers		
Puff adder (*Bitis arietans*); Africa, Saudi Arabia	fibrinolytic, bradykininogen, cholinesterase, L-amino-acid oxidase, protease, and other substances	severe local edema, necrosis and sloughing of the tissues, restlessness, respiratory distress, gastrointestinal hemorrhages, death; antiserum available
Saw-scaled viper (*Echis carinatus*); India, Iraq, Saudi Arabia, Africa	carboxypolypeptidase, cholinesterase, dipeptidase, endopeptidase, and other enzymes	local pain and edema; hemorrhages into the skin, mucous membranes, and gastrointestinal tract; liver damage, shock, death; antiserum available
European viper (*Vipera berus*); Europe, Asia, Japan	bradykininogen, L-amino-acid oxidase	local pain and edema, hemorrhages along lymphatics, very little general reaction, vomiting, abdominal pain, shock, death; antiserum available
Russell's viper (*Vipera russelli*); Asia, Indonesia	carboxypolypeptidase, cholinesterase, dipeptidase, lipase, lecithinase, protease, and other enzymes	rapidly spreading edema with hemorrhages into the tissues, abdominal pain, vomiting; blood does not clot; loss of consciousness, circulatory failure, shock, death; antiserum available
Mammals		
Duck-billed platypus (*Ornithorhynchus anatinus*); Australia	protease and other enzymes	may inflict envenomation by means of the spurs on the hind legs of the male platypus; immediate intense pain, swelling, redness in the area of the sting, faintness, numbness about the wound, restlessness, prostration; no death recorded; recovery is uneventful after a few days
Short-tailed shrew (*Blarina brevicauda*); North America	neurotoxic, proteolytic	the lower incisor teeth are associated with submaxillary poison glands; bites produce localized pain, redness in the area of the bite, shooting pains, and general discomfort of the affected part; gradual recovery after several days; no fatalities recorded

Table 9: Representative Crinotoxic Animals*

name and distribution	toxic principle	toxic effects and comments
Sponges		
Red moss (*Microciona prolifera*); eastern United States coastal waters	unknown	contact with the sponge produces a chemical irritation of the skin, redness, stiffness of the finger joints, swelling, blisters, and pustules
Flatworms		
Flatworm (*Leptoplana tremellaris*); European coastal waters	unknown	poison is produced by epidermal skin glands; no human intoxications recorded; extracts from the skin of these worms injected into laboratory animals produces cardiac arrest
Arthropods—joint-legged animals		
Blister beetles (*Cantharis vesicatorea*); United States	cantharidin	toxic substance does not seem to be produced by special glands but is found throughout the body of the beetle; no discomfort at time of initial contact; after about 8–10 hours large blisters develop on the skin accompanied by slight burning or tingling sensation; swallowing of the beetles may cause kidney damage; cantharidin is used as an aphrodisiac known as Spanish fly—a very dangerous substance to use; ingestion can cause severe gastroenteritis, kidney damage, blood in the urine, priapism, profound collapse, and death
Millipedes (species of *Orthoporus, Rhinocrichus, Julus,* and *Spirobolus*); temperate and tropical regions	unknown	repugnatorial (distasteful to enemies) fluid may be exuded or forcefully squirted from body pores a distance up to 30 inches or more; contact with the skin causes mild to moderately intense burning pain, redness, and pigmentation of the skin; toxic fluid squirted in the eyes may cause temporary blindness, an inflammatory reaction, and pain
Venomous ticks (species of *Ixodes* and *Ornithodoros*); temperate and tropical regions	unknown	tick bites result in swelling, redness, intense pain, headache, muscle cramps, loss of memory, etc.
Fishes		
Sea lamprey (*Petromyzon marinus*); Atlantic Ocean	unknown	slime of the lamprey is toxic; ingestion may cause diarrhea
Soapfish (*Rypticus saponaceus*); tropical and subtropical Atlantic	neurotoxic	slime of fish is toxic; produces an irritation of the mucous membrane
Amphibians†		
European earth salamander (*Salamandra maculosa*); Europe	skin glands of the salamander are poisonous; contains the alkaloids samandarine, samandenone, samandine, samanine, samandarone, samandaridine, and others	effects on humans not known; affects the heart and nervous system; causes in animals convulsions, cardiac irregularity, paralysis, and death
Toads (*Bufo* species); temperate and tropical regions	bufotoxin, bufogenins, and 5-hydroxytryptaline; poison includes a complex of many substances	produces a poisonous secretion in the parotid glands and skin; handling of some toads may cause a skin irritation; ingestion causes nausea, vomiting, numbness of the mouth and tongue, and tightness of the chest; the poison has a digitalis-like action
Frogs (some species of *Dendrobates, Physalaemus,* and *Rana*); northern South America and Central America	skin secretions are poisonous; histamine, bufotenine, physalaemin, serotonin, and other substances; composition varies with the species	skin secretions produce a burning sensation when handled; used by Indians as an arrow poison
Tree frogs (some species of *Hyla* and *Phyllobates*); northern South America, Central America	skin secretions are poisonous; batrachotoxin, steroidal alkaloids, serotonin, histamine, and other substances; bufotenine varies with the species	some frog species produce a burning sensation and a skin rash when handled; skin secretions in the eye may produce a severe inflammatory reaction; if ingested, poison causes vomiting and abdominal pain; batrachotoxin is extremely toxic if injected; used by Indians as an arrow poison

*Animals in which poison glands are present and poison is released into the environment through a pore. †Poisonous amphibians are sometimes referred to as "venomous," but they do not possess a true venom apparatus. They possess only poison glands.

and experiences tremors, convulsions, and coma. At even higher doses, the skin, eyes, and ovaries and testes are affected. Death may follow from 2 to 35 days after exposure. Exposure to radiation can also result in cancers of the bone marrow (leading to leukemia), lungs, kidneys, bladder, esophagus, stomach, colon, thyroid, or breasts.

Radioisotopes that are absorbed and distributed evenly throughout the body also can result in whole-body irradiation. Examples are tritium and cesium-137, both of which release beta particles that can lead to bone marrow toxicities and even, in the case of cesium-137, to death. The toxicity of tritium is less severe than that of cesium-137 because the beta particles generated by tritium are less energetic and because cesium-137 also releases gamma rays.

Local toxicities of common beta-particle emitters. Unlike tritium and cesium-137, the isotopes strontium-90, iodine-131, and cerium-144 emit beta particles that are not distributed evenly in the body. Strontium-90 releases only beta particles, while iodine-131 and cerium-144 release both beta particles and gamma rays, but their toxicities are primarily caused by the beta particles. These radioisotopes produce toxicities in the tissues where they are stored or concentrated. Strontium-90 and cerium-144 chemically resemble calcium and as a result are stored in bone. Therefore, these two radioisotopes produce bone cancer and leukemia, which is a result of the irradiation of bone marrow. Iodine-131 is concentrated in the thyroid and produces thyroid damage and tumours.

Local toxicities of common alpha-particle emitters. There are radioisotopes that emit primarily alpha particles, together with some gamma rays. Because the destructive effect on tissues of alpha particles is far greater than that of gamma rays, the toxicities of these radioisotopes are contributed mainly by the alpha particles. Because of the limited penetrability of alpha particles, only tissues in the near vicinity of the isotopic molecules are affected. These radioisotopes typically produce tumours at the storage site.

Most of the common alpha-particle emitters belong to the uranium series, which consists of radioisotopes that form one after another, via a nuclear decay reaction, and release mainly alpha particles. The series starts with uranium-238. The nuclear disintegration of uranium-238 forms radium-226 which disintegrates to form radon gas (radon-222). Radon decays to form a series of daughter nuclides, most of which are alpha-particle-releasing isotopes, such as polonium-210. The radioisotopes in the uranium series are important because uranium is the starting fuel for many nuclear reactors and because daughter nuclides in this series are commonly found in the environment.

The toxicity of uranium-238 depends on the water solubility of the uranium compound. Water-soluble forms mainly cause kidney injury, while the insoluble forms produce fibrosis and cancer of the lung. Because of its similarity to calcium, radium-226 is stored mainly in the bone, and it produces abnormal changes in the bone marrow, including anemia and leukemia, cancers of the bone, and paranasal sinuses. The next radioisotope in the uranium series is radon, radon-222. Although radon is radioactive, its toxicity is not due to retention of the gas by the lungs but rather to the series of radioactive daughter nuclides in particulate form. These particulate daughter nuclides are deposited on the respiratory tract when inhaled, the respiratory tract is irradiated by the alpha particles released, and lung cancer can result.

Other radioisotopes do not belong to the uranium series. For example, radium-224, which is deposited mainly on

bone surfaces, has been used in Europe to treat ankylosing spondylitis. Because of its short half-life (3.6 days), it affects only the bone surface and not the bone marrow. Its major toxicity is the production of bone cancer. Like uranium-238, plutonium-239, which is used in some nuclear reactors and in nuclear bombs, primarily releases alpha particles. Although there are no human data, animal studies indicate that the toxicity of plutonium-239 is similar to that of insoluble uranium-238 in causing fibrosis and cancer of the lung.

Nonionizing radiation. Nonionizing radiation includes ultraviolet light, infrared radiation, microwaves, and radio frequencies, all of which are electromagnetic waves. The toxicity of radio frequencies is rather low. On the whole, nonionizing radiation is not as toxic as ionizing radiation, and the various forms of nonionization radiation share common target organs; particularly the skin and eyes.

Ultraviolet radiation. The toxicity of ultraviolet light depends on its wavelength. Ultraviolet-A (near UV) has a wavelength of 315–400 nanometres, ultraviolet-B (mid UV) has one of 280–315 nanometres, and ultraviolet-C (far UV) has one of 200–280 nanometres. Ultraviolet-A affects primarily the skin and causes burns at high energy levels. The toxicities of ultraviolet-B and ultraviolet-C are similar, but ultraviolet-C is less toxic because it does not penetrate tissues as deeply. Both ultraviolet-B and ultraviolet-C cause injuries to the eyes and skin. Ultraviolet-B is the major component of sunlight and accelerates the aging of skin by damaging the collagen fibres under it. Ultraviolet-B also is the cause of an occupational disease known as "welder's flash," or "arc eye," which is characterized by photophobia, tears in the eyes, spasm of the eyelids, and eye inflammation. Finally, ultraviolet-B can cause skin cancer, which may be a result of the linking of thymidines, a base in DNA, produced by ultraviolet-B radiation.

Infrared radiation and microwaves. The major mechanism of toxicity of infrared radiation and microwaves is the production of heat in tissues. Infrared-A (wavelength 0.8–1.4 micrometres) penetrates the skin, causing burns and pigmentation. Infrared-A also penetrates the liquid content of the eye to reach the retina and can therefore produce damage to all parts of the eye. In contrast, infrared-B and infrared-C (wavelength 1.4–3,000 micrometres) are almost completely absorbed by the superficial layers of the skin and eyes, and the damage is thus confined to the surface. Microwaves (wavelength 1 millimetre to 1 metre) produce heat in tissues. Because testes and eyes do not dissipate heat well, due to low blood flow through these organs, temporary sterility and cataracts can be produced by microwaves.

Lasers. Lasers are high-energy light beams, visible and nonvisible, generated by atoms at an excited state and further amplified by optics. Like most other nonionizing radiation, lasers can produce skin burns. Visible lasers, with a wavelength from 0.4 to 1.4 micrometres, will cause retinal damage if they enter the eyes and are focused by the lens onto the retina.

BIBLIOGRAPHY

Toxic substances: CURTIS D. KLAASSEN, MARY O. AMDUR, and JOHN DOULL (eds.), *Casarett and Doull's Toxicology: The Basic Science of Poisons,* 3rd ed. (1986), contains in-depth discussions of basic toxicology principles and information on toxins classified according to use and target organs. Analysis of the chemical structure of toxins influencing their effect is provided in STANLEY E. MANAHAN, *Toxicological Chemistry: A Guide to Toxic Substances in Chemistry* (1989). MICHAEL A. KAMRIN, *Toxicology: A Primer on Toxicology Principles and Applications* (1988), is a concise, nontechnical general introduction. ERNEST HODGSON, RICHARD B. MAILMAN, and JANICE E. CHAMBERS, *Dictionary of Toxicology* (1988), explains concepts and terminology and covers organizations and authorities in the field.

Exposure and response to poisons: ROBERT H. DREISBACH and WILLIAM O. ROBERTSON, *Handbook of Poisoning: Prevention, Diagnosis, & Treatment,* 12th ed. (1987), contains concise but essential information on the toxicity and treatment of poisoning by biological toxins and drugs. AVRAM GOLDSTEIN, LEWIS ARONOW, and SUMNER M. KALMAN, *Principles of Drug Action: The Basis of Pharmacology,* 2nd ed. (1973), describes the principles governing chemical absorption, distribution, and excretion of the substances. MATTHEW J. ELLENHORN and DONALD G. BARCELOUX, *Medical Toxicology: Diagnosis and Treatments of Human Poisoning* (1988), focuses on the poisons derived from biological sources. SIDNEY KAYE, *Handbook of Emergency Toxicology: A Guide for the Identification, Diagnosis, and Treatment of Poisoning,* 5th ed. (1988), surveys almost 200 toxic substances.

Sources of poisoning: GEORGE D. CLAYTON and FLORENCE E. CLAYTON (eds.), *Patty's Industrial Hygiene and Toxicology,* 3rd. rev. ed., vol. 2, *Toxicology,* parts A, B, and C (1981–82), is an extensive compendium of information on the toxicology of industrial chemicals. For concise reference on the subject, see ALICE HAMILTON, *Hamilton and Hardy's Industrial Toxicology,* 4th ed., rev. by ASHER J. FINKEL (1983). The toxicology of chemicals found in commercial products is examined in ROBERT E. GOSSELIN, ROGER P. SMITH, and HAROLD C. HODGE, *Clinical Toxicology of Commercial Products,* 5th ed. (1984), where information on chemical ingredients of many commercial products is detailed. ALFRED GOODMAN GILMAN *et al.* (eds.), *Goodman and Gilman's The Pharmacological Basis of Therapeutics,* 7th ed. (1985), is an authoritative source of information on the toxicology of drugs and includes a section on the toxicity of industrial chemicals.

The toxicology of food is the subject of JOINT FAO/WHO COMMITTEE ON FOOD ADDITIVES, *Toxicological Evaluation of Certain Food Additives* (1988), including lists of acceptable consumption of flavourings, preservatives, and colours; JOSE M. CONCON, *Food Toxicology,* 2 vol. (1988), a comprehensive survey of food contamination and poisoning; PALLE KROGH (ed.), *Mycotoxins in Food* (1987); and W.F.O. MARASAS and PAUL E. NELSON, *Mycotoxicology: Introduction to the Mycology, Plant Pathology, Chemistry, Toxicology, and Pathology of Naturally Occuring Mycotoxicoses in Animals and Man* (1987). BRUCE W. HALSTEAD, *Poisonous and Venomous Marine Animals of the World,* 2nd rev. ed. (1988), is an exhaustive and profusely illustrated compendium on the toxic marine animals of the world, from protozoans to polar bears, including their historical background, names, geographic distribution, biology, mechanism of intoxication, and medical, toxicological, pharmacological, and chemical aspects.

(C.D.Kl./K.L.W.)

Poland

The Republic of Poland (Polish: Polska Rzeczpospolita) lies at the physical centre of the European continent, approximately between latitudes 49° and 55° N and longitudes 14° and 24° E. Except for its southern mountainous regions, the country consists almost entirely of lowlands within the North European Plain. The total area of Poland is 120,728 square miles (312,685 square kilometres). Its capital is Warsaw (Warsawa).

Over the past millennium, the name Poland has been applied to a shifting territorial base. At one time, in the mid-1500s, Poland was the largest state in Europe. At other times there was no Polish state at all. Its current frontiers, stretching for 2,198 miles (3,538 kilometres), were drawn in 1945. Poland is bordered to the north by the Baltic Sea, to the northeast by Russia and Lithuania, and to the east by Belarus and Ukraine. To the south the border follows the watershed of the Beskid, Carpathian, and Sudeten (Sudety) mountains, which separate Poland from Slovakia and the Czech Republic, while to the west the border with Germany is defined by the Neisse (Nysa Łużycka) and Oder (Odra) rivers.

This article is divided into the following sections:

Physical and human geography 930
- The land 930
  - Relief
  - Physiographic regions
  - Drainage and soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 934
  - Ethnic composition
  - Language
  - Religion
  - Demographic trends
  - Social structures
- The economy 935
  - General considerations
  - Resources
  - Agriculture
  - Industry
  - Finance
  - Trade
  - Transportation
- Administration and social conditions 937
  - Government
  - Education
  - Health and welfare
- Cultural life 938
  - Folk culture
  - Literature
  - Music
  - Visual arts
  - Theatre and motion pictures
  - Museums
  - Sports and recreation
  - Media
  - Censorship

History 940
- The Piast monarchy 940
  - The early state
  - Collapse and restoration
  - The period of divisions
  - Revival of the kingdom
- The states of the Jagiellonians 942
  - The waning of the Middle Ages
  - The golden age of the Sigismunds
- The Commonwealth 943
  - Báthory and the Vasas
  - The 17th-century crisis
  - Decline and attempts at reform
  - The Saxons
  - Reforms, agony, and partitions
- Partitioned Poland 948
  - The Legions and the Duchy of Warsaw
  - From the Congress of Vienna to 1848
  - The January 1863 uprising and its aftermath
  - Accommodation with the ruling governments
- Poland in the 20th century 951
  - The rebirth of Poland
  - From the Treaty of Versailles to the Treaty of Riga
  - The Second Republic
  - World War II
  - Communist Poland
  - Poland since 1989

Bibliography 956

Physical and human geography

THE LAND

Relief. The natural landscape of Poland can be divided broadly into three relief groups: the lowlands, the highlands, and the mountains. The eastern extremes of Poland display characteristics common to eastern Europe, but the rest of the country is linked to western Europe by structure, climate, and the character of its vegetation. The lowland characteristics predominate: the average elevation of the whole country is only 568 feet (173 metres) above sea level, while more than three-fourths of the land lies below 650 feet.

Poland's relief was formed by the actions of Ice Age glaciers, which advanced and receded over the northern part of the country several times during the Pleistocene Epoch (1,600,000 to 10,000 years ago). The great expanses of the Polish lowlands, part of the North European Plain, are composed of geologically recent deposits that lie over a vast structural basin. In the southern part of the country older and more diverse geologic formations are exposed, and the topography is dominated by the mountainous arc of the Carpathians, dating from the Tertiary mountain-building period of about 25 million years ago. Around the northern rim of the Carpathians lies a series of structural basins, separating the mountain belt proper from a much older structural mass, or foreland, that appears in the relief patterns of the region as the Bohemian Massif, the Sudeten, and the Little Poland (Małopolska) Uplands.

Carpathian Mountains

Physiographic regions. The relief structure can be divided more specifically into a series of east-west–trending zones. To the north lie the swamps and dunes of the Baltic coast; south of these is a belt of morainic terrain with thousands of lakes, the southern boundary of which marks the limit of the last ice sheet. The third zone consists of the central lowlands, whose minimal relief was created by streams issuing from the retreating glaciers. This zone is the Polish heartland, the site of agriculture in places where loess has been deposited over the relatively infertile fluvioglacial deposits. The fourth zone is made up of the older mountains and highlands to the south; though limited in extent, it offers spectacular scenery. Along the southern border of the country are the Sudeten and Carpathian ranges and their foothills.

The coastal plain. The Baltic Coastal Plain stretches across northern Poland from Germany to Russia, forming a low-lying region built of various sediments. It is largely occupied by the ancient province of Pomerania, which means "Along the Sea." The scarcely indented Baltic coastline was formed by wave action after the retreat of the ice sheet and the raising of sea levels. The Pomeranian (Pomorska) Bay in the west and the Gulf of Gdańsk in the


BALTIC SEA
Gulf of Gdańsk
Pomeranian Bay
Vistula Lagoon
Curonian Lagoon
RUSSIA
LITHUANIA
BELARUS
GERMANY
CZECH REPUBLIC
SLOVAKIA
UKRAINE
HUNGARY
Kaliningrad
Kaunas
Gdańsk
Elbląg
Olsztyn
Białystok
Bydgoszcz
Toruń
Włocławek
Gorzów Wielkopolski
Poznań
Konin
Warsaw (Warszawa)
Brest
Zielona Góra
Leszno
Kalisz
Łódź
Pabianice
Tomaszów Mazowiecki
Radom
Siedlce
Lublin
Chełm
Wrocław
Częstochowa
Kielce
Zamość
Opole
Katowice
Sosnowiec
Dąbrowa Górnicza
Bytom
Zabrze
Gliwice
Rybnik
Tychy
Kraków
Tarnów
Rzeszów
Przemyśl
Lviv
Nowy Sącz
Bielsko-Biała
Ostrawa
Cottbus
Dresden
Prague
Berlin
Frankfurt
Szczecin
Koszalin
Słupsk
Gdynia
Sopot
Hel
Darłowo
Kołobrzeg
Świnoujście
Stargard
Gryfice
Trzebiatów
Kamień Pomorski
Police
Szczecinek
Chojnice
Tczew
Malbork
Nowy Dwór Gdański
Braniewo
Olsztynek
Ostróda
Iława
Ełk
Suwałki
Augustów
Hrodna
Siemiatycze
Sokółka
Łapy
Zambrów
Ostrołęka
Mława
Pasłęk
Ostróda
Grudziądz
Chełmno
Świecie
Tuchola
Chojnice
Starogard Gdański
Tczew
Kwidzyn
Sztum
Nidzica
Szczytno
Pisz
Mrągowo
Giżycko
Węgorzewo
Gołdap
Ełk
Olecko
Kętrzyn
Lidzbark Warmiński
Bartoszyce
Orneta
Braniewo
Elbląg
Malbork
Nowy Dwór Gdański
Sztum
Kwidzyn
Grudziądz
Chełmno
Świecie
Tuchola
Chojnice
Człuchów
Czarne
Czersk
Skarszewy
Kościerzyna
Starogard Gdański
Tczew
Gdańsk
Wejherowo
Puck
Władysławowo
Leba
Ustka
Słupsk
Darłowo
Kołobrzeg
Świnoujście
Stargard
Gryfice
Trzebiatów
Kamień Pomorski
Police
Szczecin
Goleniów
Stargard Szczeciński
Pyrzyce
Choszczno
Pełczyce
Drawsko Pomorskie
Wałcz
Piła
Nakło nad Notecią
Chodzież
Czarnków
Wągrowiec
Wronki
Międzyrzecz
Skwierzyna
Szamotuły
Kostrzyn
Słubice
Świebodzin
Lubsko
Żary
Żagań
Nowa Sól
Głogów
Legnica
Jawor
Złotoryja
Lubin
Polkowice
Głogów
Rawicz
Leszno
Gostyń
Krotoszyn
Ostrów Wielkopolski
Jarocin
Środa
Śrem
Nowy Tomyśl
Wolsztyn
Oborniki
Turek
Konin
Kutno
Sochaczew
Płock
Gostynin
Włocławek
Inowrocław
Aleksandrów Kujawski
Lipno
Rypin
Brodnica
Nidzica
Mława
Pasłęk
Ostróda
Iława
Ełk
Suwałki
Augustów
Hrodna
Siemiatycze
Sokółka
Łapy
Zambrów
Ostrołęka
Pułtusk
Wyszków
Sokołów Podlaski
Węgrów
Mińsk Mazowiecki
Otwock
Garwolin
Puławy
Lubartów
Opole Lubelskie
Kraśnik
Janów Lubelski
Biłgoraj
Zamość
Hrubieszów
Tomaszów Lubelski
Stalowa Wola
Nisko
Leżajsk
Łańcut
Jarosław
Przemyśl
Lubaczów
Sanok
Krosno
Jasło
Gorlice
Nowy Targ
Zakopane
Tatra National Park
Pieniny National Park
Dunajec Gorge National Park
Beskid Mountains
Bukowina Tatrzańska
Nowy Targ
Krynica
Bielsko-Biała
Żywiec
Rabka
Wadowice
Oświęcim
Auschwitz Concentration Camp
Birkenau Concentration Camp
Jaworzno
Chrzanów
Olkusz
Zawiercie
Myszków
Kłobuck
Lubliniec
Kluczbork
Namysłów
Oleśnica
Trzebnica
Oława
Brzeg
Strzelin
Świdnica
Jelenia Góra
Karkonosze National Park
Nowa Ruda
Kłodzko
Ząbkowice Śląskie
Wałbrzych
Świdnica
Legnica
Głogów
Polkowice
Lubin
Żagań
Żary
Lubsko
Cottbus
Dresden
Prague
Berlin
Frankfurt
Szczecin
Gorzów Wielkopolski
Międzyrzecz
Skwierzyna
Szamotuły
Poznań
Swiebodzin
Lubuskie
Wielkopolskie
Great Poland Lowland
Great Poland National Park
Great Poland Lakeland
Lake District
Pomeranian Lakeland
Lake District
Kashubian Lakeland
Lake District
Vistula Lagoon
Curonian Lagoon
Pomeranian Bay
Gulf of Gdańsk
Kashubian Lakeland
Pomeranian Lakeland
Lake District
Great Poland Lowland
Silesian Lowland
Sudeten
Karkonosze National Park
Table Mountains
Kłodzko Valley
Silesian Upland
Opolskie
Śląskie
Małopolskie
Podkarpackie
Świętokrzyskie
Świętokrzyskie Mts.
Lysogory National Park
Roztocze National Park
Lublin Upland
Lublin Polesie
Polesie National Park
Bug Gorge National Park
Podlaskie
Białystok
Biebrza National Park
Narew National Park
Knyszyńska Forest
Augustów Forest
Hajnówka
Białowieża Forest
Białowieża National Park
Mazurian Lakeland
Lake District
Warmińsko-Mazurskie
Olsztyn

Jarocin . . . . . 51 58 N 17 31 E
Jarosław . . . . . 50 01 N 22 41 E
Jasło . . . . . 49 45 N 21 28 E
Jawor . . . . . 51 03 N 16 11 E
Jędrzejów . . . . . 50 38 N 20 18 E
Jelenia Góra . . . . . 50 54 N 15 44 E
Kalisz . . . . . 51 45 N 18 05 E
Kamień Pomorski . . . . . 53 58 N 14 47 E
Katowice . . . . . 50 16 N 19 01 E
Kazimierza Wielka . . . . . 50 16 N 20 29 E
Kędzierzyn-Koźle . . . . . 50 21 N 18 12 E
Kępno . . . . . 51 17 N 17 59 E
Kętrzyn . . . . . 54 05 N 21 23 E
Kielce . . . . . 50 50 N 20 40 E
Kłodzko . . . . . 50 26 N 16 39 E
Kluczbork . . . . . 50 59 N 18 13 E
Kolno . . . . . 53 25 N 21 56 E
Kołobrzeg . . . . . 54 11 N 15 35 E
Konin . . . . . 52 13 N 18 16 E
Kościerzyna . . . . . 54 07 N 17 59 E
Kostrzyn . . . . . 52 35 N 14 39 E
Koszalin . . . . . 54 12 N 16 11 E
Kraków . . . . . 50 05 N 19 55 E
Krasnik . . . . . 50 55 N 22 14 E
Krasnystaw . . . . . 50 59 N 23 11 E
Krosno . . . . . 49 41 N 21 47 E
Krynica . . . . . 49 26 N 20 58 E
Kutno . . . . . 52 14 N 19 22 E
Kwidzyn . . . . . 53 44 N 18 55 E
Łapy . . . . . 52 59 N 22 52 E
Łeba . . . . . 54 45 N 17 33 E
Łęczyca . . . . . 52 04 N 19 13 E
Legnica . . . . . 51 12 N 16 12 E
Leszno . . . . . 51 51 N 16 35 E
Lezajsk . . . . . 50 16 N 22 25 E
Lidzbark Warmiński . . . . . 54 08 N 20 35 E
Lipno . . . . . 52 50 N 19 12 E
Łódź . . . . . 51 45 N 19 28 E
Łomża . . . . . 53 11 N 22 05 E
Łowicz . . . . . 52 07 N 19 56 E
Lubaczów . . . . . 50 10 N 23 08 E
Lubań . . . . . 51 07 N 15 17 E
Lubartów . . . . . 51 28 N 22 38 E
Lubawa . . . . . 53 30 N 19 45 E
Lubin . . . . . 51 24 N 16 12 E
Lublin . . . . . 51 15 N 22 34 E
Lubsko . . . . . 51 48 N 14 58 E
Malbork . . . . . 54 02 N 19 03 E
Miechów . . . . . 50 22 N 20 02 E
Międzyrzecz . . . . . 52 26 N 15 35 E
Mielec . . . . . 50 17 N 21 25 E
Mińsk Mazowiecki . . . . . 52 11 N 21 34 E
Mława . . . . . 53 07 N 20 23 E
Mogilno . . . . . 52 40 N 17 58 E
Morąg . . . . . 53 55 N 19 56 E
Mrągowo . . . . . 53 52 N 21 18 E
Nakło nad Notecią . . . . . 53 09 N 17 36 E
Namysłów . . . . . 51 05 N 17 43 E
Nidzica . . . . . 53 22 N 20 26 E
Nowa Ruda . . . . . 50 35 N 16 30 E
Nowa Sól . . . . . 51 48 N 15 43 E
Nowogard . . . . . 53 39 N 15 07 E
Nowy Sącz . . . . . 49 38 N 20 43 E
Nowy Targ . . . . . 49 28 N 20 02 E
Nowy Tomyśl . . . . . 52 19 N 16 09 E
Nysa . . . . . 50 30 N 17 20 E
Olecko . . . . . 54 02 N 22 31 E
Oleśnica . . . . . 51 12 N 17 23 E
Olkusz . . . . . 50 17 N 19 34 E
Olsztyn . . . . . 53 47 N 20 29 E
Opatów . . . . . 50 48 N 21 26 E
Opoczno . . . . . 51 22 N 20 17 E
Opole . . . . . 50 40 N 17 57 E
Opole Lubelskie . . . . . 51 09 N 21 58 E
Orneta . . . . . 54 07 N 20 08 E
Ostróda . . . . . 53 42 N 19 59 E
Ostrołęka . . . . . 53 05 N 21 34 E
Ostrów Mazowiecka . . . . . 52 48 N 21 54 E
Ostrów Wielkopolski . . . . . 51 39 N 17 49 E
Ostrowiec Świętokrzyski . . . . . 50 56 N 21 24 E
Ostrzeszów . . . . . 51 25 N 17 57 E
Oświęcim . . . . . 50 02 N 19 14 E
Pabianice . . . . . 51 40 N 19 22 E
Parczew . . . . . 51 38 N 22 54 E
Piła . . . . . 53 09 N 16 45 E
Pińczów . . . . . 50 32 N 20 32 E
Piotrków Trybunalski . . . . . 51 24 N 19 41 E
Pisz . . . . . 53 38 N 21 48 E
Płock . . . . . 52 33 N 19 42 E
Płońsk . . . . . 52 38 N 20 23 E
Poznań . . . . . 52 25 N 16 58 E
Prudnik . . . . . 50 19 N 17 35 E
Prudnik . . . . . 52 10 N 20 50 E
Przasnysz . . . . . 53 02 N 20 53 E
Przemyśl . . . . . 49 47 N 22 47 E
Puck . . . . . 54 42 N 18 25 E
Puławy . . . . . 51 25 N 21 58 E
Pułtusk . . . . . 52 43 N 21 06 E
Pyrzyce . . . . . 53 08 N 14 53 E
Ratibórz . . . . . 49 37 N 19 57 E
Racibórz . . . . . 50 05 N 18 12 E
Radom . . . . . 51 25 N 21 09 E
Radomsko . . . . . 51 04 N 19 27 E
Radzyń Podlaski . . . . . 51 47 N 22 37 E
Rawicz . . . . . 51 37 N 16 52 E
Rybnik . . . . . 50 07 N 18 32 E
Ryjin . . . . . 53 04 N 19 27 E
Rzeszów . . . . . 50 03 N 22 00 E
Sandomierz . . . . . 50 41 N 21 45 E
Sanok . . . . . 49 34 N 22 12 E
Siedlce . . . . . 52 10 N 22 18 E
Siemiatycze . . . . . 52 27 N 22 53 E
Sieradz . . . . . 51 36 N 18 45 E
Sierpc . . . . . 52 53 N 19 40 E
Skarzysko-Kamienna . . . . . 51 07 N 20 54 E
Skierniewice . . . . . 51 58 N 20 09 E
Skwierzyna . . . . . 52 36 N 15 30 E
Słupsk . . . . . 54 27 N 17 02 E
Sochaczew . . . . . 52 14 N 20 15 E
Sokółka . . . . . 53 25 N 23 30 E
Sokołów Podlaski . . . . . 52 24 N 22 15 E
Sopot . . . . . 54 27 N 18 34 E
Sosnowiec . . . . . 50 18 N 19 10 E
Śrem . . . . . 52 08 N 17 01 E
Środa . . . . . 52 14 N 17 17 E
Stalowa Wola . . . . . 50 34 N 22 03 E
Starachowice . . . . . 51 04 N 21 04 E
Stargard Szczeciński . . . . . 53 20 N 15 03 E
Staszów . . . . . 50 34 N 21 40 E
Stettin, see Szczecin
Strzegom . . . . . 50 58 N 16 21 E
Strzelin . . . . . 50 47 N 17 04 E
Suwałki . . . . . 54 06 N 22 56 E
Świdnica . . . . . 50 51 N 16 30 E
Świdwin . . . . . 53 46 N 15 47 E
Świebodzin . . . . . 52 15 N 15 32 E
Świecie . . . . . 53 25 N 18 27 E
Świnoujście . . . . . 53 55 N 14 15 E
Szamotuły . . . . . 52 37 N 16 35 E
Szczebrzeszyn . . . . . 50 42 N 22 58 E
Szczecin (Stettin) . . . . . 53 25 N 14 35 E
Szczecinek . . . . . 53 43 N 16 42 E
Szczytno . . . . . 53 34 N 21 00 E
Tarnobrzeg . . . . . 50 35 N 21 41 E
Tarnów . . . . . 50 01 N 20 59 E
Tarnowskie Góry . . . . . 50 27 N 18 52 E
Tczew . . . . . 54 06 N 18 48 E
Tomaszów Lubelski . . . . . 50 27 N 23 25 E
Tomaszów Mazowiecki . . . . . 51 32 N 20 01 E
Toruń . . . . . 53 02 N 18 36 E
Trzcianka . . . . . 53 02 N 16 27 E
Trzebiatów . . . . . 54 04 N 15 16 E
Trzebnica . . . . . 51 19 N 17 03 E
Tuchola . . . . . 53 35 N 17 51 E
Turek . . . . . 52 02 N 18 30 E
Tychy . . . . . 50 08 N 18 59 E
Ustka . . . . . 54 35 N 16 51 E
Ustroń . . . . . 49 43 N 18 48 E
Wąbrzeźno . . . . . 53 17 N 18 57 E
Wągrowiec . . . . . 52 48 N 17 12 E
Wałbrzych . . . . . 50 46 N 16 17 E
Wałcz . . . . . 53 16 N 16 28 E
Warsaw (Warszawa) . . . . . 52 15 N 21 00 E
Węgorzewo . . . . . 54 13 N 21 44 E
Węgrów . . . . . 52 24 N 22 01 E
Wejherowo . . . . . 54 36 N 18 14 E
Wieluń . . . . . 51 13 N 18 33 E
Władysławowo . . . . . 54 48 N 18 25 E
Włocławek . . . . . 52 39 N 19 05 E
Włoszczowa . . . . . 50 52 N 19 58 E
Wodzisław Śląski . . . . . 50 00 N 18 28 E
Wołomin . . . . . 52 21 N 21 15 E
Wolsztyn . . . . . 52 07 N 16 08 E
Wrocław (Breslau) . . . . . 51 06 N 17 02 E
Wronki . . . . . 52 43 N 16 24 E
Ząbkowice Śląskie . . . . . 50 35 N 16 48 E
Zabrze . . . . . 50 19 N 18 47 E
Zakopane . . . . . 49 18 N 19 58 E
Zambrów . . . . . 52 59 N 22 15 E
Zamość . . . . . 50 43 N 23 15 E
Żary . . . . . 51 38 N 15 09 E
Zawiercie . . . . . 50 30 N 19 26 E
Zgierz . . . . . 51 51 N 19 25 E
Zielona Góra . . . . . 51 56 N 15 30 E
Złocieniec . . . . . 53 32 N 16 00 E
Złotoryja . . . . . 51 07 N 15 55 E
Żory . . . . . 50 03 N 18 42 E
Żywoń . . . . . 51 21 N 21 36 E
Żyrardów . . . . . 52 04 N 20 26 E
Żywiec . . . . . 49 41 N 19 13 E

Physical features and points of interest

Auschwitz Concentration Camp . . . . . 50 02 N 19 14 E
Babia, Mount . . . . . 49 34 N 19 32 E
Baltic Sea . . . . . 55 00 N 17 00 E
Barycz, *river* . . . . . 51 41 N 16 16 E
Beskid Mountains . . . . . 49 25 N 20 15 E
Białowieża Forest National Park . . . . . 52 40 N 23 50 E
Bieszczady National Park . . . . . 49 05 N 22 44 E
Bóbr, *river* . . . . . 52 03 N 15 05 E
Brda, *river* . . . . . 53 08 N 18 08 E
Bug, *river* . . . . . 52 31 N 21 05 E
Bydgoszcz Canal . . . . . 53 08 N 17 37 E
Bzura, *river* . . . . . 52 25 N 20 15 E
Carpathian Mountains . . . . . 49 20 N 21 00 E
Dąbie, Lake . . . . . 53 29 N 14 40 E
Drawa, *river* . . . . . 52 52 N 16 00 E
Drwęca, *river* . . . . . 53 00 N 18 42 E
Dukla (Dukajec) Pass . . . . . 49 25 N 21 42 E
Dunajec, *river* . . . . . 50 14 N 20 44 E
Gdańsk,Gulf of . . . . . 54 40 N 19 15 E
Gliwicki Canal . . . . . 50 22 N 18 09 E
Gopło, Lake . . . . . 52 36 N 18 21 E
Gorce National Park . . . . . 49 35 N 20 12 E
Great Poland (Wielkopolska) Lakeland, *physical region* . . 52 30 N 16 30 E
Great Poland (Wielkopolska) Lowland, *physical region* . . 52 10 N 16 30 E
Great Poland (Wielkopolska) National Park . . . . 52 16 N 16 47 E
Gwda, *river* . . . . . 53 03 N 16 44 E
Jeziorak, Jezioro, *lake* . . . . . 53 44 N 19 38 E
Kampinos National Park . . . . . 52 20 N 20 35 E
Karkonosze, *mountains* . . . . . 50 45 N 15 40 E
Karkonoskie National Park . . . . . 50 45 N 15 40 E
Kraków, Historic Centre of . . . . . 50 05 N 20 00 E
Krzna, *river* . . . . . 52 09 N 23 32 E
Łebsko, Jezioro, *lake* . . . . . 54 43 N 17 25 E
Little Poland (Małopolska) Uplands, *physical region* . . 50 45 N 20 00 E
Liwiec, *river* . . . . . 52 36 N 21 34 E
Lublin (Lubelska) Uplands, *physical region* . . 50 50 N 22 40 E
Łyna, *river* . . . . . 54 37 N 21 14 E
Łysica, *mountain* . . . . . 50 54 N 20 54 E
Mamry, Jezioro, *lake* . . . . . 54 10 N 21 42 E
Masurian (Mazurskie) Lakeland, *physical region* . . 53 40 N 21 00 E
Mazovian (Mazowiecka) Lowlands, *physical region* . . 52 10 N 21 15 E
Miedwie, Lake . . . . 53 17 N 14 54 E
Narew, *river* . . . . . 52 26 N 20 41 E
Neisse (Nysa Łużycka), *river* . . . 52 04 N 14 46 E
Ner, *river* . . . . . 52 09 N 18 41 E
Nida, *river* . . . . . 50 18 N 20 55 E
Noteć, *river* . . . . . 52 44 N 15 25 E
Nysa Łużycka, see Neisse
Oder (Odra), *river* . . . 53 32 N 14 38 E
Ojców National Park . . . . . 50 13 N 19 50 E
Parsęta, *river* . . . . . 54 11 N 15 34 E
Pieniny National Park . . . . . 49 25 N 20 24 E
Pilica, *river* . . . . . 51 52 N 21 17 E
Pisa, *river* . . . . . 53 14 N 21 52 E
Podhale, *valley* . . . . 49 25 N 20 01 E
Podlasian (Podlaska) Lowland, *physical region* . . 53 00 N 22 40 E
Południowy Kanał Obry, *canal* . . . . 52 01 N 15 59 E
Pomeranian (Pomorskie) Lakeland, *physical region* . . 54 00 N 16 30 E
Pomeranian (Pomorska) Bay . . . . . 54 00 N 14 15 E
Prosna, *river* . . . . . 52 09 N 17 40 E
Rega, *river* . . . . . 54 09 N 15 17 E
Roztocze National Park . . . . . 50 34 N 23 01 E
Rysy, *mountain* . . . 49 11 N 20 05 E
San, *river* . . . . . 50 44 N 21 50 E
Sandomierz (Sandomierska) Basin, *physical region* . . 50 15 N 21 45 E
Silesian (Śląska) Lowland, *physical region* . . 51 05 N 16 30 E
Silesian (Śląska) Upland, *physical region* . . 50 20 N 18 45 E
Słowiński National Park . . . . . 54 43 N 17 25 E
Śniardwy, Lake . . . 53 46 N 21 45 E
Śnieżnik, *mountain* . 50 13 N 16 51 E
Sudeten (Sudety), *mountains* . . . . . 50 30 N 16 00 E
Świętokrzyski National Park . . . . . 50 53 N 20 57 E
Świętokrzyskie (Holy Cross) Mountains . . . . . 50 48 N 20 50 E
Szczecin (Szczeciński) Lagoon . . . . . 53 46 N 14 14 E
Tarnica, *mountain* . 49 04 N 22 44 E
Tatra National Park . . . . . 49 15 N 20 00 E
Tatry, *mountains* . . 49 10 N 20 00 E
Vistula (Wisła), *river* . . . . . 54 21 N 18 57 E
Vistula (Wisłany) Lagoon . . . . . 54 20 N 19 30 E
Warsaw, Historic Centre of . . . . . 52 15 N 21 00 E
Warta, *river* . . . . . 52 36 N 14 37 E
Wełna, *river* . . . . . 52 39 N 16 49 E
Wieliczka Salt Mines, *historic site* . . . . . 49 59 N 20 04 E
Wieprz, *river* . . . . . 51 33 N 21 50 E
Wisłok, *river* . . . . . 50 12 N 22 33 E
Wisłoka, *river* . . . . . 50 27 N 21 23 E
Wkra, *river* . . . . . 52 27 N 20 45 E
Wolin National Park . . . . . 53 55 N 14 30 E

Coastal settlements

east are the two major inlets. In the southern portion of the former, two islands block off the Szczecin (Szczeciński) Lagoon, into which the Oder discharges its waters. In the Gulf of Gdańsk, the Vistula (Wisła) River forms a large delta. Sandbars, on which the winds have created large dunes, line much of the coast, separating the coastal lakes and lagoons from the sea. The main urban centres are the ports of Szczecin (German: Stettin) on the lower Oder and Gdańsk (Danzig) and Gdynia in the east. The central portion of the Baltic Coastal Plain is scantily populated—there are only small fishing ports—and the landscape has a desolate beauty.

The lake region and central lowlands. The belt immediately to the south of the coastal plain presents, with its many lakes, a varied, hilly landscape of glacial origin. Wide river valleys divide the region into three parts: the Pomeranian (Pomorskie) Lakeland; the Masurian (Mazurskie) Lakeland, east of the lower Vistula; and the Great Poland (Wielkopolskie) Lakeland. The larger settlements and the main communications routes of this zone lie in and along the river valleys; the remainder of the area is mostly wooded and thinly populated. Only the eastern portion of the Great Poland Lakeland has a developed agriculture.

The extensive central lowlands contain isolated relief features shaped by the oldest glaciations, but their character is generally flat and monotonous. The postglacial lakes have long since been filled in, and glacial outwash masks the weakly developed meltwater valley channels. The basins of the main rivers divide the area into the Silesian (Śląska) Lowland, which lies in the upper Oder; the southern Great Poland Lowland, which lies in the middle Warta basin; and the Mazovian (Mazowsze) and Podlasian lowlands, which lie in the middle Vistula basin. Lower Silesia and Great Poland are important agricultural areas, but large industrial centres are found in many parts of the central lowlands. Warsaw, the capital, situated on the middle Vistula, is most prominent.

The Little Poland Uplands. These uplands, south of the central lowlands, extend from east to west but are folded transversely. In the west is the Silesian-Kraków upthrust, with rich deposits of coal. A second upthrust is formed by the ancient rocks of the Świętokrzyskie ("Holy Cross") Mountains, which reach a maximum elevation of 2,008 feet. Between these two regions lies the Nida River basin, with an average height of 650 to 1,000 feet. East of the Świętokrzyskie Mountains, the uplands are cut by the valley of the Vistula, beyond which lie the Lublin (Lubelska) Uplands. In the south occur patches of loess on which fertile brown- and black-earth soils have developed. The older geologic regions contain valuable minerals: in the Silesian-Kraków uplands there are coal, iron, zinc, and lead deposits. These mineral resources have made possible the rise of Poland's most important industrial region, and the landscape of Upper Silesia is highly urbanized. Katowice is the largest centre, and the region is closely linked with that around Kraków. The Little Poland Uplands protect the Little Poland Lowlands, in which Kraków lies, from the colder air of the north. To the north the Staropolski ("Old Polish") Basin, situated in the foothills of the Świętokrzyskie Mountains, has a long history of industrial production. Kielce is the area's urban centre.

The industrial region

The Sudeten. The Sudeten and its foreland, part of the larger Bohemian Massif, have a long and complex geologic history. They owe their present rugged form, however, to earth movements that accompanied the Carpathian uplift, and the highest portion, the Karkonosze ("Giant Mountains"), reaches 5,256 feet above sea level. The region contains rich mineral deposits, notably coking coal, which has occasioned the growth of an industrial centre around Wałbrzych. The region has many small towns. Resorts and spas are found in more secluded areas. The foreland of the Sudeten, separated by a large fault from the larger mass, contains many granite quarries.

The Carpathians. The southernmost and also most scenic portion of Poland embraces the Carpathian Mountains and their associated chains and basins, created in the geologically recent mountain-building Tertiary Period (about 66.4 to 1.6 million years ago). Within the Polish frontiers lie the Oświęcim and Sandomierz basins, a portion of the Beskid Mountains, the Orawka-Podhale Basin, and the Tatra (Tatry) Mountains. The sub-Carpathian basins contain deposits of salt, sulfur, and natural gas and some petroleum. There are both a large rural population and many towns of medium size. The highest peak of the Beskid Mountains, Mount Babia, reaches 5,659 feet; the Tatras, with a maximum elevation of 8,199 feet (2,499 metres), are the highest portion of the Polish Carpathians. Zakopane, the largest tourist and resort centre in Poland, lies at their feet. The Bieszczady Mountains—rolling, carpeted in beech woods, and sparsely inhabited—lie in the extreme southeast.

Drainage and soils. Virtually the entire area of Poland drains to the Baltic, about half via the Vistula River and a third via the Oder. Polish rivers experience two periods of high water each year. In spring the lowland rivers are swollen by melted snow; the effect intensified both by ice dams (which block the rivers for one to three months) and by the fact that the thaw first strikes the upper reaches of the northward-flowing rivers. The summer rains bring a second maximum around the beginning of July.

There are some 9,300 Polish lakes with areas of more than 2½ acres (1 hectare), and their total area is about 1,200 square miles, or 1 percent of the national territory. The majority, however, are found in the northern glaciated belt, where they occupy more than 10 percent of the surface area.

Polish soils are varied and without clearly marked regional types. The greatest area is covered by podzol and pseudopodzol types, followed by the less widely distributed brown-earth soils, which are richer in nutrients. In the south are extensive areas of fertile loess-based soils. The rendzinas, formed on limestone rocks, are a unique type. Alluvial soils filling the river valleys and peaty swamp soils found in the lake area and in poorly drained valleys are also of importance.

Climate. Varying types of air masses collide over Poland, influencing the character of both weather and climate. The major elements involved are oceanic air masses from the west, cold polar air from Scandinavia or Russia, and warmer, subtropical air from the south. A series of barometric depressions moves eastward along the polar front year-round, dividing the subtropical from the colder air, bringing to Poland, cloudy, wet days. In winter, polar-

© Bernard and Catherine Desjeux—Corbis



The picturesque town of Kazimierz Dolny along the Vistula River.

continental air often becomes dominant, bringing crisp, frosty weather, with still colder Arctic air following in its wake. Warm, dry, subtropical-continental air often brings pleasant days in late summer and autumn.

Six seasons

The overall climate of Poland has a transitional—and highly variable—character between maritime and continental types. Six seasons may be clearly distinguished: a snowy winter of one to three months; an early spring of one or two months, with alternating wintry and springlike conditions; a predominantly sunny spring; a warm summer with plenty of rain and sunshine; a sunny, warm autumn; and a foggy, humid period signifying the approach of winter. Sunshine reaches its maximum over the Baltic in summer and the Carpathians in winter, and mean annual temperatures range from 46° F (8° C) in the southwestern lowlands to 44° F (7° C) in the colder northeast. The climate of the mountains is determined by altitude.

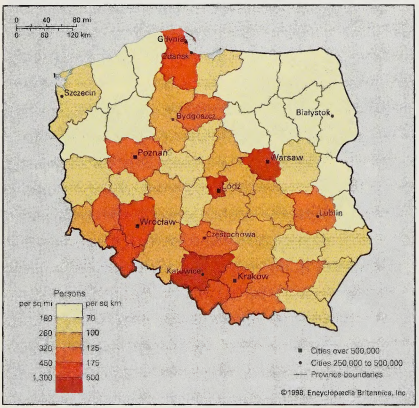
The annual average precipitation is about 24 inches (600 millimetres), but in the mountains the figure approaches 31 to 47 inches, dropping to about 18 inches in the central lowlands. In winter, snow makes up about half the total in the plains and almost all of it in the mountains.

Plant and animal life. *Vegetation.* The vegetation of Poland consists of some 2,250 species of seed plants, 630 mosses, 200 liverworts, 1,200 lichens, and 1,500 fungi. Holarctic elements (*i.e.,* those appertaining to the temperate belt of the Northern Hemisphere) are dominant among the seed plants.

Trees

The northeastern limits of certain trees—notably beech, fir, and the variety of oak known as pedunculate—run through Polish territory. There are few endemic species: the Polish larch (*Larix polonica*) and the Ojców birch (*Betula oycoviensis*) are two examples. Some relics of tundra vegetation have been preserved in the peat bogs and mountains. More than a fourth of the country is wooded. Poland lies in the zone of mixed forests, but in the southeast a fragment of the forest-steppe vegetation zone intrudes. In the northeast there are portions of the eastern European subtaiga, with spruce as a characteristic component. In the mountains the vegetation, like the climate, is determined by elevation. Fir and beech woods give way to the spruce of the upper woods, which in turn fade into subalpine, alpine, and snow-line vegetation.

Fauna. Poland's animal life belongs to the European–West Siberian zoogeographic province of the Palearctic subregion and is closely linked with the vegetation cover. The vertebrate fauna includes nearly 400 species, including many types of mammals and more than 200 native birds. Deer and wild pigs roam the woods, elk are found in the coniferous forests of the northeast, and steppe rodents, such as the brindled gopher, are found in the south. Brown bear and wildcat live in the mountain woods, and the chamois and marmot are found at the highest levels. The great Białowieża Forest, a national park contiguous with the Byelavyezhskaya Forest in Belarus, provides shelter for a small herd of wisent, or European bison, all that remain of the vast numbers that once roamed the continent.



Population density of Poland.

The environment. Rapid industrialization since World War II in not only Poland but also the neighbouring areas of the Czech Republic and eastern Germany has severely polluted many areas of the country. Upper Silesia and Kraków, in particular, have suffered some of the highest levels of atmospheric and groundwater pollution in Europe, but several areas of central Poland, where cement is produced and brown coal (lignite) is burned, also suffer from air pollution. The country's major rivers are all badly polluted by industrial and urban effluents, and almost all of Poland's cities and larger towns are major sources of pollution. Tree growth has been adversely affected in many of the forests in the Sudeten and western Carpathians. The problem was not officially recognized until the early 1970s, and little action was taken until the Solidarity movement began agitating in the early 1980s. Much greater reductions in the emission of pollutants occurred, however, as a consequence of the rapid fall in industrial production in the early 1990s, following the abandonment of communism and the introduction of economic reforms.

Settlement patterns. *Rural settlement.* Until the mid-20th century the pattern of rural settlement differed widely from one part of Poland to another. In the centre and east of the country, many villages were small and irregular in shape, reflecting their origin as self-sufficient clusters of cultivators and pastoralists in forest clearings; in the mountains they stretched along the valleys; and in Lower Silesia they were larger and more orderly, associated with the planned settlement of the area by Teutonic people in medieval times. In the north, rural settlement was dominated by large, landed estates, which had belonged to the Prussian Junkers. Many houses in the centre, east, and south were wooden. Since the 1950s, however, there have been marked changes. Some attempt has been made to retain traditional building styles in the mountains, but many older, single-story houses in all parts of the country have been replaced with ones of two or three stories built of cinder block; and many villages have expanded, especially those close to larger cities and in regions popular with tourists.

Warsaw

Urban settlement. Warsaw is the largest city in Poland, with a population twice that of the next largest city. It consists of a small, historic core, on the west bank of the Vistula, comprising a medieval town, Old Town (Stare Miasto), and its 18th-century suburbs—New Town (Nowe Miasto) to the north and Krakowskie Przedmieście to the south. About 85 percent of the city's buildings were destroyed during World War II; much of the city therefore dates from the period since 1950. Its skyline is dominated by the Palace of Culture and Science, a skyscraper built in the Soviet style in the 1950s. Many of Warsaw's inhabitants live in large, unattractive blocks of flats that were built around the edge of the city in the 1960s and '70s.

Kraków (the original capital), Gdańsk, Poznań, and Wrocław (German: Breslau) share many characteristics with Warsaw, having medieval and early modern cores surrounded by 19th- and, especially, 20th-century suburbs containing a mixture of manufacturing industry and poor-quality apartment-style housing. In contrast, Łódź, Poland's second largest city, dates from the 19th century, when it grew rapidly to become one of the most important centres of the textile industry in the Russian Empire. The other major urban area is that of southern Upper Silesia, a conurbation of mining and industrial settlements stretching some 30 miles from Dąbrowa Górnicza to Gliwice.

THE PEOPLE

Ethnic composition. Before World War II the Polish lands were noted for the richness and variety of their ethnic communities. In the provinces of Silesia, Pomerania, and Masuria (then in Germany) there was a significant

minority of Germans. Ukrainian settlements predominated in the regions east of Chełm and in the Carpathians east of Nowy Sącz. In all the towns and cities there were large concentrations of Yiddish-speaking Jews. The Polish ethnographic area stretched eastward: in Lithuania, Belarus, and western Ukraine, all of which had a mixed population, Poles predominated. There were significant Polish minorities in Daugavpils (in Latvia), Minsk (in Belarus), and Kiev (in Ukraine).

The war, however, killed vast numbers of people, precipitated massive migrations, and radically altered borders. As a consequence the population of Poland became one of the most ethnically homogeneous in the world. Virtually all of Poland's people claim Polish nationality, with Polish as their native tongue. Ukrainians, the largest minority group, are scattered in various northern districts. Lesser numbers of Belarusians and Lithuanians live in areas adjoining Belarus and Lithuania. The Jewish community, almost entirely Polonized, has been greatly reduced. In Silesia a significant segment of the population, of mixed Polish and German ancestry, tends to declare itself as Polish or German according to political circumstances.

West Slavic language group

Language. The Polish language belongs to the West Slavic branch of Slavic languages. It has several dialects that correspond in the main to the old tribal divisions; the most significant of these (in terms of numbers of speakers) are Great Polish (spoken in the northwest), Little Polish (spoken in the southeast), Mazovian, and Silesian. Mazovian shares some features with Kashubian, whose remaining speakers live west of Gdańsk near the Baltic Sea. Elsewhere, Polish has been influenced by contact with foreign languages. In Silesia the inimitable regional patois contains a mixture of Polish and German elements. Since 1945, as the result of mass education and mass migrations, standard Polish has become far more homogeneous, although regional dialects persist. In the western and northern territories, resettled in large measure by Poles from the Soviet Union, the older generation came to speak a language characteristic of the former eastern provinces. Small numbers of people also speak Belarusian, Ukrainian, and German as well as several varieties of Romany.

Literary Polish developed from the medieval period onward, on the basis of the dialects of Great Poland and Little Poland. By the 19th century Polish was well established both as a literary vehicle and as the dominant language of common speech in Poland, despite attempts of the partitioning powers to Germanize or Russify the population. Indeed, quite the opposite happened, and the Polish language became the main touchstone of national identity.

Religion. The overwhelming majority of the Polish population is Roman Catholic, and a large number are practicing Catholics. Poland is among the most uniformly Catholic nations in the world, and the Roman Catholic church in Poland enjoys immense social prestige and political influence.

After World War II all religious institutions became subject to the control of the state office for religious denominations. In practice the Roman Catholic church wielded a full measure of independence, partly through the sheer force of the faithful and partly because in all important matters it answered to the pope in Rome and not to the government in Warsaw. Those opposed to communism within Poland were greatly encouraged by the election in 1978 of the archbishop of Kraków, Karol Cardinal Wojtyła, as Pope John Paul II, the first non-Italian pope since the 16th century. The religious minorities—comprising primarily Lutherans and Evangelicals, along with Eastern Orthodox, Uniate, Jewish, Seventh-day Adventist, and Jehovah's Witness denominations—though encouraged by the anti-Roman Catholic policies of the communist state, were barely visible except in local areas. The influence of the Catholic church became very great after the fall of communism in 1989, leading to its greater involvement in state schools and the replacement of the country's liberal abortion law with much more restrictive legislation. However, subsequently church attendance and income fell sharply, and the abortion law was relaxed again in the mid-1990s.

Demographic trends. The population of Poland was transformed during and immediately after World War II. Of the nearly 35 million people who had been living within the Polish frontiers in 1939, only about 24 million were to be found within the new frontiers by 1946. The decrease of some 11 million can be accounted for mainly by war losses but also in part by changes in frontiers and by the migration of former Polish citizens whose homelands were incorporated into the Soviet Union.

War losses and postwar migrations

Polish war losses are the subject of some controversy. The official figure, issued in 1947, was 6,028,000, although this referred exclusively to losses within the postwar frontiers. As a result of the changes in frontiers, millions of Germans were forcibly expelled in 1946–47, and millions of Poles were transferred from the Soviet Union in the same period. An estimated 500,000 Ukrainians and Belarusians also were transferred into the Soviet Union. At the same time, there were vast internal movements into the new northern and western territories annexed from Germany.

Population losses and movements on this scale introduced long-term distortions into demographic structures and trends, especially in such categories as males, urban dwellers, and the educated classes as a whole. However, the immediate postwar generation had an unprecedented birth rate, and the population grew rapidly again, especially in the northern and western territories, returning to its prewar level in 1977. The birth rate fell sharply after the early 1980s, and population growth slowed.

Emigration has been a permanent feature of Polish life for most of the past two centuries, and roughly one Pole in three lives abroad. Wave after wave of political émigrés has left Poland since the mid-18th century. By far the greatest numbers of people have left, however, for economic reasons. Starting in the mid-19th century, Polish emigrants moved into the new industrial areas of Europe and later to America.

Social structures. Polish society since World War II has been transformed by two great movements: by the growth of a dominant, urban, industrialized working class and by the continuing drift of peasants from the rural areas into towns and cities. Whereas in 1946 there were nearly twice as many people in the countryside as in towns, the two numbered equally in the late 1960s, and by the mid-1990s the 1946 position had been reversed. So-called peasant workers, who tended to live on the fringes of industrial regions, contrived to benefit from both movements: while one part of the family maintained the farm, other family members earned wages in local factories.

Development under the communist government stressed the classless and proletarian nature of society. However, peasants resisted the collectivization of their holdings, and about 2.5 million small farms remained in private ownership throughout the communist period. Some development of a privately owned, small-scale service industry was also allowed under communism. The party elite formed a third group, enjoying a range of privileges unavailable to ordinary workers.

Since 1989, in contrast, Polish society has become much more differentiated. Many private businesses have been established, and a small number of people have become wealthy. Many elderly people on fixed incomes, however, have suffered sharp falls in their standard of living.

THE ECONOMY

General considerations. Before World War II Poland was a free-market economy based largely upon agriculture but with a few important centres of manufacturing and mining. After the initiation of communist rule (1948), the country developed an increasingly industrial, state-run command economy based on the Soviet model. It operated within the rigid framework of the Council on Mutual Economic Assistance (Comecon), an organization of Eastern-bloc countries dominated by the Soviet Union.

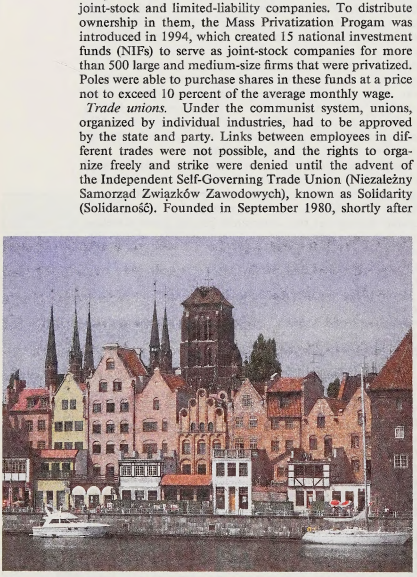
From the mid-1970s the Polish economy experienced limited growth, largely as a result of an antiquated industrial infrastructure, government subsidies that masked inefficient production, and wages that were artificially high relative to the standard of living. In the late 1980s a swelling government deficit and hyperinflation brought

about economic crisis. With the fall of communism and the demise of Comecon, the Polish economy became increasingly involved with the market-oriented global economy, for which it was ill-suited. To try to achieve economic stability, the postcommunist government introduced "shock therapy," which sought both to control inflation and to expedite Poland's transition to a market economy. As part of that plan, wages were frozen, price controls were removed, subsidies to state-owned enterprises were phased out, and large-scale private enterprise was again permitted. As a result, in the early 1990s, output and gross domestic product (GDP) dropped significantly. Unemployment grew, affecting as many as one in seven Poles. Privatization of some of Poland's large industries proved to be a slow process. Inflation, however, began to drop, and by the mid-1990s production and GDP recorded dramatic turnarounds and unemployment decreased. Poland's balance of payments improved (partly as the result of debt forgiveness), and the country became one of the leading economies of the former Eastern bloc. In 1996 it became a member of the Organisation of Economic Co-operation and Development and awaited full membership in the European Union (EU).

"Shock therapy"

Privatization. Under communism the principal branches of industry, services, and trade were directly owned by the state. There was, however, a surprisingly large sector of legal self-employment, and small-scale private businesses—including workshops, services, and restaurants—proliferated. Moreover, some three-fourths of Poland's farmland was privately owned. A government collectivization campaign begun in 1949 was abandoned in 1956. After the fall of communism, both industry and agriculture became increasingly privatized. By the early 1990s more than half the economy was in private ownership, while more than four-fifths of shops were privately owned. The privatization of larger enterprises was more complicated. A number of these were transformed into joint-stock and limited-liability companies. To distribute ownership in them, the Mass Privatization Progam was introduced in 1994, which created 15 national investment funds (NIFs) to serve as joint-stock companies for more than 500 large and medium-size firms that were privatized. Poles were able to purchase shares in these funds at a price not to exceed 10 percent of the average monthly wage.

Trade unions. Under the communist system, unions, organized by individual industries, had to be approved by the state and party. Links between employees in different trades were not possible, and the rights to organize freely and strike were denied until the advent of the Independent Self-Governing Trade Union (Niezależny Samorząd Związków Zawodowych), known as Solidarity (Solidarność). Founded in September 1980, shortly after widespread strikes organized by the Interfactory Strike Committee, Solidarity broke the monopoly of the official party unions, quickly gained mass support, even among party members, and extended its activities far beyond narrow syndicalist concerns. After December 1981 Solidarity was suppressed and its leaders declared criminals, but in 1988 renewed labour unrest and nationwide strikes forced negotiations between the government and Solidarity (held in early 1989) that resulted in the legalization of Solidarity and in Poland's first free elections since World War II.



Waterfront of Gdańsk, on the Motława River.
Leo de Wys Inc./Friedmar Damm

Resources. *Mineral resources.* Poland is relatively well endowed with natural resources. Its principal mineral asset is bituminous coal. Most of the output is derived from the rich Upper Silesian coalfield. Brown coal is mined as well. Other fuel resources include small amounts of petroleum and moderately large deposits of natural gas.

Sulfur is Poland's second most important mineral, and the republic ranks among the world leaders in both reserves and production. Other important nonmetallic minerals include barite, salt, kaolin, limestone, chalk, gypsum, and marble.

Among metallic minerals, copper, silver, and zinc are the most important. Zinc production in Poland declined in importance relative to world output in the 1990s. Poland is a major world producer of copper and silver.

Other resources. Soil quality varies, but among the large fertile areas are Lower Silesia and the Little Poland Lowlands, the Kujawy, the Vistula delta, and the Lublin area. The soil is somewhat poorer in large parts of central and northern Poland. Some four-fifths of the country's wooded land is occupied by coniferous trees, with pine, larch, and spruce the most economically important. The bulk of the country's hydroelectricity comes from the Carpathians, the Sudeten region, and the Brda and Vistula rivers.

Agriculture. Polish agriculture was unique in the Soviet bloc in that private farms accounted for the greater part of total output. Most of those private farms continue to be smaller than 12 acres (5 hectares). In postcommunist Poland farm incomes declined rapidly in real terms as the prices of industrial products rose, and imported processed foods from western Europe competed strongly with lower-quality Polish products. Many state farms collapsed after 1989, as did the system of state purchase upon which much of the private sector had relied. In the early 1990s the number of people employed in agriculture declined by one-fifth, partly because of the liquidation of state farms but also through retirement among the aging agricultural population. In the mid-1990s there was a reversal of this trend, but, even as Polish agriculture recovered from the drought of the early 1990s, less than one-fourth of the working population was employed in agriculture. In that same time period agriculture's portion of GDP fell from nearly one-fifth to less than one-tenth.

Nevertheless, Poland remains one of the world's leading producers of rye and potatoes. Other principal crops include wheat and sugar beets. Most farming is mixed, and beef cattle, dairy cows, and pigs are raised throughout the country. As Poland became increasingly integrated into the global economy in the mid-1990s, about half its agricultural exports went to the EU.

Principal crops

Industry. During the period of communist rule, remarkable quantitative advances in industrial production were overshadowed to some extent by shortcomings in quality and by problems of organization. Moreover, industrial production in Poland—governed almost solely by quantitative requirements and dependent on inexpensive raw materials provided through Comecon—was largely inefficient and poorly prepared to compete in the global marketplace. Industrial output fell dramatically after the demise of communism, especially during the first years of shock therapy, with declines of a third or more in almost all of manufacturing and mining following the freeing of prices and the collapse of Comecon. As Polish industry began to downsize, however, production improved, and by the mid-1990s industry accounted for about two-fifths of GDP. The principal branches of the manufacturing sector are machinery and transport equipment, food products, metals and metal products, chemicals, beverages, tobacco, and textiles and clothing.

All but a small portion of the total electrical power output is derived from thermal plants using bituminous coal and lignite. Natural gas has largely replaced manufactured gas. Almost all petroleum and petroleum products are imported.

Finance. All financial institutions were owned by the state after 1944–45 and formed an integral part of centralized economic planning after 1949. The National Bank of Poland (Narodowy Bank Polski) acted as the main agent of the government's financial policy, managing everything from the currency and money supply to wages and prices, credit, investment, and the detailed business of all state enterprises. Other important banks included the Commercial Bank of Warsaw (Bank Handlowy w Warszawie) for foreign trade and the General Savings Bank (Bank Polska Kasa Opieki). In the late 1980s and early '90s, the banking industry was reorganized. The National Bank became an independent central bank, with responsibility for regulating the banking sector and the currency. A number of private banks also were established.

Currency reform

Until 1990, internal monetary operations were conducted in inconvertible local currency, while external operations were conducted either in foreign currency, especially U.S. dollars, or, for the Soviet bloc, in special units of account such as convertible rubles. Exchange rates against foreign hard currency were flexible according to the needs of the state bank. In 1990, as part of a government program to move the Polish economy toward a free-market system, the exchange rate of the złoty, Poland's currency, was allowed to be set freely on the international currency markets. In 1991 Poland established a stock exchange.

Trade. Poland's trade was greatly affected by the fall of communism. Hitherto, much was conducted within Comecon, including the export of coal and machinery to the Soviet Union and other countries in eastern Europe. Since 1990, however, two-thirds has been with western Europe, nearly half of which has been with Germany. Less than one-sixth has been with other former Comecon countries. Machinery, metals, textiles and clothing, coal, and food account for the bulk of exports, and machinery, chemicals, and fuels are the major imports. Germany is the largest market for almost all categories of exports, while Russia remains by far the most important source of energy imports, and Germany and Italy the chief sources of foreign machinery and chemicals. (E.I.U./A.H.D./N.D.)

Transportation. *Railways.* The communications system in Poland developed in the 19th and early 20th centuries, when the country was divided between Russia, Germany, and Austria. The three areas thus developed in different economic and political conditions, and the main railway lines were centred on the capitals of the three empires. The density of the railway networks in the three sectors was uneven. In 1918 independent Poland took over the system and redesigned and rebuilt it according to the standard European gauge. Among the most important railway lines built after that date were those linking Warsaw with Poznań and Kraków and a coal trunk line linking Upper Silesia with the newly built seaport of Gdynia.

After the devastation of World War II, the railway system was reconstructed once again, and the most heavily used lines were converted to electric power. Because of the location of the country, Polish railway lines were important in the carriage of transit freight among the socialist countries of eastern Europe. Demand for rail transport fell sharply, however, after the end of communism, both for freight and passengers. The railways are administered by the Polish State Railways (Polskie Koleje Państwowe).

Highways. The highway system originally showed disproportions similar to those of the railways; that is, the densest network was on lands belonging to Germany and the least dense on lands belonging to Russia. An attempt to remedy this situation was made between 1918 and 1938 and again, but more intensively, after 1945. Modern multilane highways designed for high traffic volumes have been built in Warsaw, and projects have been undertaken to link Warsaw to provincial centres, but the road system in general is of low quality.

Inland waterways and merchant shipping. The middle course of the main Polish river, the Vistula, contains many navigational hazards, and thus the Vistula is a less important waterway than the smaller Oder. The Oder is linked by the modern Gliwice Canal to the Upper Silesian industrial region and carries coal to the port of Szczecin. The Oder basin is also linked to the lower Vistula by the Bydgoszcz Canal. Inland navigation is of little importance in Poland, however. Shipping is well developed, and there are three large seaports—Szczecin (the largest), Gdynia, and Gdańsk—as well as smaller fishing and coastal navigation ports.

Air transport. Domestic and international air transport is provided by LOT (from Polskie Linie Lotnicze), a state-owned enterprise scheduled for partial privatization. There are numerous international routes, centred on the airport at Warsaw.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. *Political parties.* Beginning in 1948, Poland was governed by the Polish United Workers' Party (PUWP; Polska Zjednoczona Partia Robotnicza), which was modeled on the Communist Party of the Soviet Union. The postwar government was run as a dual system in which state organs were controlled by parallel organs of the PUWP. The executive branch of government, therefore, was in effect the PUWP, with the party's first secretary acting as the de facto head of state and most powerful authority. The party's Political Bureau, or Politburo, operated as the central administration, and the party ensured its control over all offices and appointments by use of the *nomenklatura*, a list of politically reliable people.

Polish United Workers' Party

Two other parties, the United Peasants' Party (Zjednoczone Stronnictwo Ludowe; ZSL) and the Democratic Party (Stronnictwo Demokratyczne; SD), were permitted to exist, but only as entirely subservient allies of the PUWP. However, in 1989 economic and political problems obliged the government to recognize the independent trade union Solidarity (which had been banned not long after it came into being in 1980) and allow it to contest at least some seats in a general election. The PUWP was guaranteed 65 percent of the seats in the lower house of the Sejm (the state legislature), but Solidarity won all the rest and all but one of those in the Senate, going on to form Poland's first postcommunist government with the support of the SD and ZSL, which broke their alliance with the PUWP. In 1990 the Polish communist party voted to replace the PUWP with two social democratic parties, the Social Democracy of the Republic of Poland and the Social Democratic Union. In the same year, Lech Wałęsa, the leader of Solidarity, was elected president.

Thereafter, however, as the costs of economic reform were felt by Poles, support for Solidarity waned and the party split into several smaller groups. In the first completely free elections, in 1991, no party obtained more than an eighth of the vote, leading to a succession of short-lived coalition governments. In 1993 the postcommunist and Peasant parties won a majority of seats, and in the presidential election of 1995 Wałęsa was defeated by a former communist, Aleksander Kwasniewski. There has been no fundamental change in economic and political policy: all postcommunist governments have given high priority to the integration of Poland into the EU and the North Atlantic Treaty Organization (NATO). However, the government of former communists has reduced the role of the Roman Catholic church in schools and eased the highly restrictive antiabortion legislation of the early 1990s.

State constitution and institutions. The constitution of Poland's postwar socialist state, the Polish People's Republic, took effect in 1952 but was amended numerous times, most significantly in early 1989, when constitutional reforms worked out between the government and Solidarity were passed by the Sejm. Among the changes were the replacement of the Council of State by the office of president and the reinstatement of the Senate, which had been abolished in 1946. The existing Sejm, with 460 members, became the lower house of the new National Assembly, and the Senate, or the upper house, was assigned 100 members. Additional reforms passed later in 1989 by the National Assembly included the guarantee of

free formation of political parties and the return of the state's official name to the Republic of Poland.

(J.A.K./N.D./A.H.D.)

Adopted in 1992, the provisional Small Constitution was replaced by a new document in 1997. This new constitution confirmed the mixed presidential-parliamentary form of government that had been established during the period 1989–92. The president is directly elected to not more than two five-year terms, serves as commander in chief of the armed forces, has the power to declare martial law or a state of emergency, and can veto an act of the Sejm (which, in turn, can override that veto with a three-fifths vote). The president nominates the prime minister and, on the prime minister's recommendation, the cabinet, subject to the legislature's approval, but the president cannot dismiss the government. All members of the legislature—460 deputies in the Sejm and 100 senators—are popularly elected to four-year terms. Laws must be adopted by both houses, but, if there is disagreement between the two, the Sejm may override the Senate by a majority vote.

Regional administration. The regional administration is on three levels. The largest units are the 49 *województwa* (provinces), followed by 267 regional authorities and some 2,500 *gminy* (towns and rural communes). These last are the fundamental territorial units within Poland.

Judicial system. The constitution guarantees the judiciary's independence. Poland has a Supreme Court and other special judicial bodies (including the Supreme Administrative Court, military courts, and industrial tribunals) as well as general courts, comprising provincial and district courts. General courts deal with criminal, civil, and family matters; commercial courts deal with civil law disputes between businesses. There is also a Court of Appeals.

Armed forces. Poland's armed forces consist of three services—the army, the air force, and the navy. Under the communist government, the armed forces were highly politicized. The country was a founding member of the Warsaw Pact, a mutual-defense organization dominated by the Soviet Union, and supplied the second largest contingent to its forces. When the pact dissolved in 1991, Poland indicated that it wished to become part of NATO, which it did in 1999. Its forces have been depoliticized.

Security services. The regular defense of Poland's frontiers is provided by the border guard. The Office of State Protection (UOP), established in 1990, is charged with the country's intelligence services. Normal civilian police services are under the authority of the Ministry of Internal Affairs. Under the communist government, police services were undertaken by the Citizens' Militia—of which the Motorized Detachments of the Citizens' Militia (ZOMO) acted as a mobile, paramilitary riot squad—and the Security Service (SB), a secret political police force. The paramilitary nature of the Policja ("Police"), as they became known after 1990, has diminished.

Education. Schools of all types are free, the system of schooling is standard, and attendance from age 7 to 14 is compulsory. The system contains nursery, primary, and secondary schools. There are several types of secondary schools offering basic vocational training, vocational and technical training, and general college-preparatory education. Basically, all schools are subject to the Ministry of National Education, but medical schools and colleges are subject to the Ministry of Health and Social Welfare, army colleges to the Ministry of National Defense, and higher schools of art to the Ministry of Culture and Arts.

Secondary schools

Prominent universities include the University of Warsaw (founded 1818), the Jagiellonian University (1364) in Kraków, and the Catholic University of Lublin (1918). The highest academic institution is the Polish Academy of Sciences, which has numerous research institutes.

Health and welfare. Health care in Poland has been handled largely by the Ministry of Health and Social Welfare, which oversees the regional health departments. Facilities include clinics; hospitals; sanatoriums, rest homes, and spas; and ambulance services. Private medical and dental practices proliferated after the fall of communism, and the pharmaceutical industry also was privatized. In general the health care system was in transition in the 1990s. Medical services were seriously strained during periods of general economic crisis. In 1999, the government launched a major reform of the universal health care system.

Under communism social insurance for the health service covered free treatment for all workers and the members of their families, as well as for pensioners, invalids, students, and certain others. In addition, there was a social service whose purpose was to ensure a suitable means of support for the elderly and invalids. Services for the unemployed were established as part of the 1989–90 economic reforms. During the mid-1990s, however, a number of laws were enacted that reduced the formerly comprehensive coverage of the unemployment program.

As a result of the program of urbanization that began in the 1940s, Polish cities became overwhelmed by migrant workers from the countryside, and the demand for housing vastly exceeded supply. In urban areas, various cooperative housing schemes were put into operation by the local government authorities, but the standard family apartment was inadequate for many families. As a result of the low priority placed on the creation of housing during communist rule, housing shortages were extreme in the 1980s and '90s. In postcommunist Poland private ownership of housing increased significantly.

(J.A.K./N.D./A.H.D./Ed.)

CULTURAL LIFE

The culture of Poland has been nurtured by a great variety of traditions. Until World War II, because Poland was an area of multinational settlement, it drew not only on the dominant Polish culture but also on that of the minorities, especially the Germans, Jews, and Ruthenians. The common Slavic element traditionally has been weak, but aspects of it can be identified in language, literature, and folklore. The formative experience, undoubtedly, was the adoption of Roman Catholicism in the 10th century and the resultant millennium of involvement in "Western" civilization. German, French, and Italian influences have been particularly strong. Unlike Russia, Poland was deeply immersed in all the great movements of Western culture, and its cultural identity was already strong before the series of partitions of Polish territory began in 1772. Because of the loss of political independence, the 19th and 20th centuries were characterized by an unrelenting struggle to preserve the national culture and its values from foreign impositions and government policy.

The Roman Catholic church in Poland has played a social and cultural role far beyond the religious sphere. After World War II, catechism lessons exposed children to a nonofficial view of the world. Church-sponsored societies, such as the Catholic Intellectual Clubs, provided a unique forum for free public discussion. Parish halls provided shelter for a wide variety of uncensored exhibitions, plays, films, and meetings.

Folk culture. Because of rapid industrialization and urbanization, Poland's traditional folk culture has been seriously undermined since World War II. Regional dress, regional forms of speech, peasant arts and crafts, and religious and folk festivals have all been swamped by mass culture from the new cities and the media. In an effort to compensate, the Roman Catholic church has tried to preserve the religious elements of folk culture, notably in the large annual pilgrimages to shrines such as Częstochowa, Kalwaria, Zebrzydowska, Lanckorona, or Piekary Śląskie. The communist authorities supported folk music and folk dancing. The colourful and stylized repertoire of the State Folk Ensemble, Mazowsze, for example, won international acclaim. Several regional communities, including the Górale ("Highlanders") of Podhale, the Kurpie in the northeast, and the inhabitants of Łowicz, near Warsaw, have created an authentic blend of the old and the new.

Religious elements

Literature. Polish literature developed long ago into the main vehicle of national expression. For many Poles, their literature stands with their religion as the twin pillars of their heritage. It provides one of their most cherished links with Western civilization and one of the main safeguards of their national identity. The close relationship between local political events and literary trends, however, together with a necessary resort to elaborate allegories, allusions,

and symbols during the communist period, rendered many excellent Polish works inaccessible to the foreign public. (For detailed discussion, see POLISH LITERATURE.)

The first half of the 19th century produced the three greatest Polish poets; Adam Mickiewicz, Juliusz Słowacki, and Zygmunt Krasiński. In the second half of the 19th century and the beginning of the 20th century the great Polish prose writers—including the Nobel Prize winners Henryk Sienkiewicz and Władysław Reymont—were active. Among the postwar Polish poets are Zbigniew Herbert and the Nobel Prize winner Czesław Miłosz.

Music. Polish music, like Polish literature, has a continuous tradition reaching back into the Middle Ages. As the least political of the arts, it suffered less from official constraints. Founded on the inimitable rhythms and melodies of folk music—the krakowiak, mazurka, and polonaise—its native characteristics developed early, and a distinctive school of Polish church music already flourished during the Renaissance. The first major Polish opera, *Cud mniemany czyli Krakowiacy i Górale* ("The Pretended Miracle; or, The Krakovians and the Highlanders"), by Jan Stefani and Wojciech Bogusławski, was staged in 1794. In the 19th century Stanisław Moniuszko wrote a series of popular operas, including *Halka, Straszny dwór* ("The Haunted Manor"), and *Hrabina* ("The Countess"). Frédéric Chopin is usually considered to have created the quintessence of Polishness in music. In addition to his renown as one of the supreme master composers, he was the first of a constant stream of instrumentalists from Polish lands, among them many prominent Jewish performers, who have won international acclaim. Pianists such as Ignacy Paderewski and Artur Rubinstein and violinists such as Henryk Szeryng attest to the vitality of Polish musical life. Contemporary Polish composition has been dominated by the names of Karol Szymanowski, Witold Lutosławski, and Krzysztof Penderecki. All branches of classical music—opera, symphony, chamber, and choral—are well represented in Poland, and several orchestras and choirs appear regularly on the international circuit. Popular music in Poland is derived largely from Western styles, although Polish jazz has earned a reputation for experiment and excellence.

Architecture

Visual arts. From the Middle Ages many fine examples of Romanesque and Gothic architecture, both secular and religious, have been preserved, together with outstanding sculptures, among which the wooden altar of Wit Stwosz (Veit Stoss), in the Church of St. Mary (Mariacki) in Kraków, is the most famous. The vast red-brick castle of Malbork (Marienburg), once the headquarters of the Teutonic Knights, is among the most impressive in Europe. The architecture and sculpture of the Renaissance and Baroque periods were formed under Italian influence but nevertheless developed individual Polish forms, as in the town hall of Poznań or the decorated granaries at Kazimierz Dolny. Zamość, a model Renaissance city built in the 1580s, has survived virtually intact. The best-preserved urban architecture of the late Middle Ages and Renaissance is that of the Old Town and the Wawel castle in Kraków. The classicism of the end of the 18th century and the beginning of the 19th left its most valuable monuments in some of the great palaces, like that of the Radziwiłł at Nieborów or at Łazienki in Warsaw. There are many examples of imperial German and Russian architecture from the 19th century, among them Lublin Castle.

Painting attained its greatest development in the second half of the 19th century, encompassing styles developing at that time in western Europe, but again with specific national characteristics. The greatest fame was achieved by Jan Matejko, creator of monumental, romantic historical canvases, and by Henryk Siemiradzki. The greatest sensitivity was shown in portraits by Stanisław Wyspiański, an artist active also in drama and design.

Theatre and motion pictures. The Polish national theatre, as distinct from earlier religious, court, and foreign plays that had circulated since the Middle Ages, dates from the end of the 18th century. The great pioneer was Wojciech Bogusławski, an actor, director, and playwright. Political conditions during the period of partition (1772–1914) inhibited theatrical development, however, and most of the Romantic masterworks of Mickiewicz or Słowacki were never staged during their lifetimes. The comedian and satirist Aleksander Fredro earned a less exalted but no less lasting reputation. Kraków, in Austrian Galicia, became a centre of lively theatre at the turn of the century. Between World Wars I and II, Juliusz Osterwa in Warsaw and Leon Schiller in Łódź launched the experimental tradition. After 1956, once the era of Socialist Realism had passed, the avant-garde came into its own. The Theatre of the Absurd was explored alongside the revival of the classical repertoire. During the 1960s the Laboratory Theatre of Jerzy Grotowski (whose theories and methods emphasizing the nonverbal aspects of theatre had a broad impact, especially in the United States) gained international acclaim. Henryk Tomaszewski's Pantomime Theatre experienced parallel success. Tadeusz Kantor, a painter and designer, has been an important influence.

Cinema

The origins of Polish cinema date to 1909, but it was only in the late 1950s that its products began to attract worldwide attention. Just as the State Film School at Łódź earned high standing in the filmmaking profession, so the work of individual directors who broke free of official preferences achieved great success both at home and abroad. Undoubtedly, the leading name is that of Andrzej Wajda, whose films and theatre productions set precedents for independence and excellence in exploring the conflicts in Polish society. Among other distinguished directors are Andrzej Munk, Roman Polanski, Aleksander Ford, Tadeusz Konwicki, Krzysztof Zanussi, Agnieszka Holland, and Krzysztof Kieslowski. Historical epics have enjoyed great popularity.

Museums. In part because of colossal wartime destruction and in part also because historiography in Poland has long been subject to official controls, the need to preserve and cherish the records and artifacts of the past is felt with special urgency. Archives and museums of art, ethnography, archaeology, and natural history can be found in many Polish cities. The Czartoryski Museum in Kraków dates to the beginning of the 19th century, the Archaeology Museum in Poznań to 1857, and the National Museum in Warsaw to 1862. After World War II, official policy concentrated on the creation of new regional museums in cities recovered from German occupation, on museums connected with the history of the communist movement, on former private palaces and collections acquired by the state, and on sites connected with Nazi war crimes, such as Oświęcim (Auschwitz) or Majdanek. The Roman Catholic church is active in preserving and exhibiting the art treasures and records connected with Poland's religious heritage.

Sports and recreation. Team sports and spectator sports are encouraged in Poland. Professional association football (soccer) teams attract large crowds in the towns, while local authorities provide facilities for athletics and swimming. Skiing and mountaineering in the Tatras and sailing on the Baltic or the Masurian Lakes are well developed. Sports such as tennis or golf are less available.

The press

Media. With very few exceptions, during the period of communist rule, the mass media, including all television and radio stations, were either owned or controlled by the state or the party. The Polish press included the official organs of the party and state, such as *Trybuna ludu* ("People's Tribune"), the organ of the PUWP; a wide band of less closely controlled semi-party newspapers and journals, such as *Życie Warszawy* ("Warsaw Life"), *Polityka* ("Politics"; a lively weekly), and *Twórczość* ("Creativity"; an intellectual monthly); the independent sector, headed by the respected Kraków *Tygodnik powszechny* ("Universal Weekly") and by the Roman Catholic journals *Znak* ("The Sign") and *Więź* ("The Link"); and the underground "free sector," in which local newssheets circulated. Restrictions on the media were eased in 1989, and Solidarity supporters began publishing numerous journals and newspapers, including the daily newspaper *Gazeta Wyborcza* ("Voters Daily"; Eng. ed. *Gazeta International*).

Censorship. Under the communist government the Main Office for the Control of Presentations and Public Performances (GUKPIW), with its headquarters in Warsaw, controlled the media, publishing, films, theatres, exhibitions, advertising, and related activities. The bureau

maintained an office in all television and radio stations, press and publishing houses, film and theatre studios, and printing works throughout the country. Authorization was required even for such printed items as wedding invitations, obituary notices, or stationery. Access to photocopiers and printing machines was closely controlled, and all purchases of paper in bulk required a permit. Censorship of the foreign mail was routine. No sphere of information was immune, however distant from immediate political concerns; censors attempted not only to suppress material but also to mold all information at its source.

Despite the official controls, speech was not generally suppressed in Poland, and the highly literate Poles became masters at writing and reading "between the lines." Also, unofficial information was available from the foreign media, from within the Roman Catholic church, and in the vast realm of underground publishing. State censorship of the press was abandoned in 1990, leading to the appearance of a wide range of publications.

(J.A.K./A.H.D./N.D.)

For statistical data on the land and people of Poland, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

THE PIAST MONARCHY

The early state. The terms Poland and Poles appear for the first time in medieval chronicles in the late 10th century. The land that the Poles, a West Slavic people, came to inhabit was covered by forests with small areas under cultivation where clans grouped themselves into numerous tribes. The dukes (*dux*) were originally the commanders of an armed retinue (*drużyna*) with which they broke the authority of the chieftains of the clans, thus transforming the original tribal organization into a territorial unit. Two tribes, the Polanie, centring around the fortified settlement (*castrum*) of Gniezno, and the Wiślanie, who lived around Kraków, expanded to bring other tribes under their control. Exposed to some missionary activities linked with St. Methodius, the state of Wiślanie fell under the rule of Great Moravia—which was destroyed by the Magyar invasion of the early 10th century—and came eventually under the rule of Mieszko I, the first ruler of the Polanie mentioned in written records. He is regarded as the founder of the Piast dynasty, the beginnings of which are clouded in legend. Creating what a contemporary traveler described as the most powerful of the existing Slav states, Mieszko accepted Roman Catholicism via Bohemia in 966. A missionary bishopric directly dependent on the papacy was established in Poznań. This was the true opening of Polish history, for Christianity was a carrier of Western civilization with which Poland was henceforth associated.

The reign of Mieszko I

Facing the crucial problem of Poland's relationship to the two pillars of medieval Christendom, the Germanic Holy Roman Empire and the papacy, Mieszko battled the expansive tendencies of the former while he sought reliance on Rome, to which he subordinated his state in a curious document, the *Dagome Iudex* (*c.* 991). Poland alternately competed and cooperated with neighbouring Bohemia and Hungary as well as with the principality of Kievan Rus. At Mieszko's death the Polish state stretched from the Baltic Sea to the Carpathian Mountains, resembling in shape post-World War II Poland.

Because the principle of primogeniture was unknown in the country, every succession led to internal strife. Mieszko's successor was Bolesław I the Brave. Commanding a huge military force, he sought hegemony in east-central Europe. In 1000 he received in Gniezno Emperor Otto III, who dreamed of restoring a universal Roman empire and who recognized the sovereign status of the Polish duke. Moreover, Otto agreed to an independent Polish ecclesiastical organization that added an archbishopric in Gniezno and bishoprics in Kraków, Wrocław, and Kołobrzeg to the already extant bishopric in Poznań. Given the role of the church in medieval statehood, this was a great achievement. Paying their respects to the martyred St. Adalbert (Vojtěch), the two rulers sought to coordinate their missionary activities in the pagan Slav lands between the Elbe and Oder rivers. This area, home of the so-called Polabian Slavs, formed a kind of buffer between the two states and was the object of their respective expansion.

The influence of Otto III

The successors of Otto pursued German objectives rather than imperial mirages and struggled with Bolesław, who briefly occupied Bohemia and intervened in Kievan Rus. Polish-German strife continued intermittently until 1018. In 1025 Bolesław assumed the royal crown, which made him the equal of the other monarchs of Europe.

Collapse and restoration. The virtual collapse of the state under Bolesław's son Mieszko II, who was even obliged to renounce his kingly status, showed how much the political fortunes of a state were bound to the personality of its ruler. Mieszko's successor, Casimir I, had to flee the country, which was torn by internal strife. A pagan reaction against Christianity combined with revolt against fiscal and administrative burdens to bring about a popular uprising. Casimir had to be restored by the emperor, Conrad II, who wished to preserve a balance of power in the region. Known later as "the Restorer," Casimir eventually succeeded in bringing under his sway most of the Polish lands, reviving the ecclesiastical organization, and making Kraków his capital.

Casimir's son and successor, Bolesław II the Bold, sought to revive the great power designs of the first Bolesław. Skillfully exploiting the great Investiture Controversy between the empire and the papacy that affected most of Europe, Bolesław II sided with Rome and gained the royal crown in 1076. Bolesław was later drawn into a conflict with Stanislaus (Stanisław), the bishop of Kraków, whom the king ordered killed in 1079 under circumstances still debated by historians. Bolesław then fled to Hungary, where he died. The cult of St. Stanislaus, who was canonized in 1253, became widespread in Poland and was invoked to defend the freedom of religion against the state and ethics against power.

Under Bolesław's brother and successor, Władysław I Herman, claims to the royal crown and to a more ambitious foreign policy were abandoned. Efforts to maintain centralized power clashed with the ambitions of the rising magnate class. Poland became torn by internal conflicts out of which Bolesław III the Wry-Mouthed emerged as the sole ruler (1102–38). Promoting Christianity, he expanded his sway over Western Pomerania, whose towns and harbours were already important centres of trade and crafts. Eastern, or Gdańsk, Pomerania came under direct Polish administration. After an invasion by Emperor Henry V was repelled, peace prevailed with the empire, and Bohemia renounced its claims to Silesia.

The period of divisions. *Collapse of Bolesław's governing system.* The awareness of centrifugal trends and external dangers led Bolesław III to establish in his testament of 1138 a system meant to assure greater stability. He divided the state among his sons; the oldest became the senior duke, whose domain included the capital in Kraków and who had general powers over military, foreign, and ecclesiastical matters. By the early 13th century, however, the efforts of the grand duke to exert real controls came to naught. The entire system was characterized by disputes, subdivisions, and fratricidal strife into which the neighbouring powers were frequently drawn.

During the period of divisions, lasting almost 200 years (until the rule of Casimir III), Poland underwent transformation in almost every sphere of life. The centrally controlled early Piast monarchy had been based on a system of fortified settlements from which an official called the castellan acted as administrator, military commander, judge, and tax collector and tended to the ruler's domain. Around some settlements there arose so-called service villages, in which artisans produced objects needed by the dukes and their retinues. The emerging social pyramid positioned the duke and his officials and leading warriors on top, with various categories of freemen, part-freemen, and slaves at the bottom. Between the 10th and the 12th centuries this system slowly began to break down. Improved cultivation methods (notably the three-field system) enhanced the value of the land with which the ruler endowed the church and compensated his nobles, warriors, and officials. Estates cultivated by a semi-serf population grew

significantly. The old *drużyna* changed into a smaller personal guard, the armed force being composed of nobles performing military service as landholders.

Cultural developments, 11th–13th centuries. The church was the principal proponent of learning and art. Romanesque and then Gothic architecture made their way into Poland. Religious orders such as the Benedictines arrived in the 11th century, the Cistercians in the 12th, and the Dominicans and the first nuns in the 13th. Cathedral and, later, parish schools appeared. The first two chronicles, that compiled by a Benedictine monk known as Gallus Anonymous and that of Wincenty Kadlubek, were written in the early 12th and at the end of the 12th and beginning of the 13th centuries, respectively.

Social and economic developments. The 13th century marked a turning point in the history of medieval Poland. The agricultural boom was accompanied by the development of salt mining in Little Poland and of silver and gold mining in Silesia. The Polish lands were brought more fully into the European economy, participating in the west-east trade as well as in that of the Baltic region in the north and that along the Danube in the south. The growth of large landed estates was partly the cause and partly the consequence of surplus production. It became profitable to cultivate the land through free tenant farmers rather than serfs, and large groups of settlers were attracted from as far away as the Rhineland and the Low Countries. Demographic trends in western Europe facilitated this "colonization." The settlers—assured of personal freedom, fixed rents, and some measure of self-administration and operating under the so-called German Law—founded new villages and towns or reorganized old ones. Towns received formal charters (Wrocław in 1242, Poznań in 1253, Kraków in 1257) that provided for autonomy and self-government modeled on that of the German city of Magdeburg—hence the term Magdeburg Law.

Although the burgher population became largely German or German-speaking, the extent of settlement by Germans was restricted except in Silesia and Pomerania. Otherwise, most of the countryside remained Polish. Another alien group, however, began to play an important role in the country's economy—namely, the Jews escaping persecution in the west. Bolesław V the Chaste of Great Poland granted to the Jews the Kalisz Privilege (1264), which provided personal freedom, some legal autonomy, and safeguards against forcible baptism.

The Kalisz Privilege

Feudalism. Economic and social transformation led to some forms of feudalism and organization of estates. A system based on contractual personal arrangements between superiors and inferiors (lords and vassals)—with land being the traditional means of reward for services (fiefs)—did not really prevail in Poland. Nor did a typical feudal pyramid exist. Nevertheless, vassalage of sorts and customs of chivalry and knighthood developed. In view of the weakening of the rulers, the landowners, both ecclesiastical (the church in the 12th century) and lay (the nobility in the 13th century), succeeded in obtaining so-called immunities—*i.e.,* exemptions for their estates from taxes, services, and the legal jurisdiction of the state.

During that period the church functioned as the only structure that transcended the divisions. Although the Silesian duchies gravitated toward Germany, the archbishopric of Gniezno continued to include the diocese of Wrocław. Several archbishops were active proponents of reunification of Poland, notably Jakób Świnka. The concepts of Corona Regni Poloniae, as divorced from the actual ruler, and of *gens polonica*—an early form of nationalism that identified the state with the Polish people and implied its indivisibility—began to make their appearance.

The arrival of the Teutonic Knights. The chances of reunification were dim, as the various branches of the Piast dynasty pursued their vested interests and further subdivided their lands. Western Pomerania, with its native dynasty, and Eastern Pomerania were already largely severed from Poland and threatened by the expansive margravate of Brandenburg. In the north the pagan Lithuanians, Prussians, and Jatvingians were harassing Mazovia. In 1226 Conrad of Mazovia called in the German crusading order, generally known as the Teutonic Order, provided them with a territorial base, and assumed that after a joint conquest of the Prussian lands (later known as East Prussia) they would become his vassals. The Teutonic Knights, however, tacitly secured imperial and papal recognition and forged Conrad's acquiescence to their independent status. After a series of ruthless campaigns, Prussia was conquered, was colonized with German settlers—the old Prussian population being virtually wiped out—and became a powerful state of the Teutonic Knights. While German historians have traditionally stressed the civilizing and organizational achievements of the Knights, the Poles have emphasized their ruthlessness and aggressiveness. The arrival of the Teutonic Knights changed the balance of forces in that part of Europe and marked the beginning of the rise of Prussia as a great power.

In 1241 Little Poland and Silesia experienced a disastrous Mongol (Tatar) invasion. The duke of Silesia, Henry II the Pious, who had been gathering forces to reunite Poland, perished in the Battle of Legnica (Liegnitz) in 1241, and the devastation wrought by the Mongols may have contributed to the above-mentioned colonization.

Revival of the kingdom. *The Czech dynasty.* In the late 13th century Bohemia emerged as the leading country in east-central Europe, and King Otakar II (Přemysl Otakar II; reigned 1253–78) even sought the imperial crown. His son Wenceslas II profited from the chaos prevailing in the Polish duchies—a bid for unification by Przemysł II of Great Poland (crowned king in 1295) was cut short by his assassination—to become king of Poland in 1300. Establishing an administration based on provincial royal officials (*starosta*)—a permanent feature of Polish administration in the centuries to come—he temporarily pacified the country. Wenceslas's grandiose plans to rule all of east-central Europe ended with his death in 1305, followed a year later by the assassination of his son Wenceslas III. This meant the end of the native Czech Přemyslid dynasty, and John of Luxembourg claimed the thrones of Bohemia and Poland. His pursuit of the latter was opposed by one of the minor dukes, Władysław the Short, who had earlier battled the two Wenceslases and their supporters. Allying himself with the new king of Hungary, Charles I, he withstood the enmity of Bohemia, the Teutonic Knights, rival Polish dukes, and the mainly German patriciate of Kraków. At one point the struggle assumed the character of a Polish-German national conflict.

John of Luxembourg

Władysław I. Władysław was crowned king of Poland in 1320, but he no longer controlled Silesia—whose dukes opted for John and which henceforth came under the Bohemian crown—and the Teutonic Knights seized Eastern Pomerania. The massacre the Knights perpetrated in Gdańsk in 1308 entered Polish folklore. Thus, the reunited Polish kingdom was deprived of two of its most developed provinces and was effectively cut off from the Baltic Sea. Cooperating closely with Hungary, Władysław sought unsuccessfully to regain Pomerania through lawsuits and papal arbitration, but the Knights ignored the verdicts. A major battle with the invading Knights fought at Płowce in 1331 was a Pyrrhic victory for Władysław.

Casimir the Great. Under Władysław's son Casimir III the Great (reigned 1333–70), the only Polish ruler to bear this epithet, peace was made with John of Luxembourg, who gave up his claims to the Polish crown at the meeting of the kings of Poland, Hungary, and Bohemia at Visegrád, Hungary, in 1335. Casimir's simultaneous renunciation of Silesia was somewhat equivocal, and he sought later to regain the Silesian duchies by diplomacy and force. In 1343 Poland signed a peace treaty with the Teutonic Knights through which it recovered some land but retained only formal suzerainty over Pomerania. That policy of compromise was a tactical necessity on the part of a state still much weaker than the Teutonic Knights, Bohemia, or Hungary. Between 1340 and the 1360s, however, Poland expanded by roughly one-third, acquiring a larger part of Halicz, or Red, Ruthenia (the future eastern Galicia), which was also coveted by Hungary and Lithuania. That acquisition marked an expansion beyond ethnic Polish territory. Casimir's international prestige was evidenced by his acting as arbiter between the Luxembourgs, Angevins, and Habsburgs and subsequently hosting an in-

ternational conference in Kraków in 1364 that dealt with general European issues.

Domestic achievements

The domestic achievements of Casimir may be subsumed under the slogan "One king, one law, one currency." His rule uncontested, Casimir presided over a process of unification and codification of laws in the mid-14th century for Great and Little Poland that is often called the Statute of Wiślica. In need of trained lawyers, he founded a university in Kraków (1364) modeled largely on that of Bologna.

The introduction of a new currency, the Kraków *grosz,* stimulated the economy and assisted the development of international trade. Many brick and stone structures arose in the country as well as a large number of fortified castles. The population and its density increased. In view of a new wave of Jewish immigrants, the 1264 privilege was extended throughout the kingdom, and the town of Kazimierz, adjacent to Kraków, became a Jewish centre. The privileged condition of the Jews, although they were resented as competitors by the burghers (who staged anti-Jewish riots), eventually resulted in Poland becoming the home of the largest Jewish population in Christendom.

Louis I. Casimir designated as his successor his nephew Louis I the Great of Hungary, who gained the support of influential nobles by granting them certain privileges in 1355. Louis' rule in Poland (1370–82), with his mother acting as regent, proved disappointing. Despite earlier promises, he definitely abandoned Silesia and Pomerania and sought to make Halicz Ruthenia directly dependent on Buda. Anxious to secure the succession to the Polish crown for one of his daughters, he granted privileges to the Polish nobility in the Pact of Koszyce (Hungarian: Kassa) in 1374. Among those privileges was the guarantee of a minimum tax, which meant that any future increase would have to be negotiated with the nobles as an estate. Thus, the principle of representation was established but did not become operative for decades to come.

The marriage of Jadwiga. After Louis' death the lords of Little Poland selected his younger daughter, Jadwiga (Hungarian: Hedvig), over her sister Maria (the wife of King Sigismund of Hungary). Preventing Jadwiga's marriage to Wilhelm Habsburg, the lords chose for her husband Jagiełło (Lithuanian: Jogaila), the grand duke of Lithuania. There were several reasons for this momentous move that opened a new chapter in Polish history by linking the relatively small kingdom with a huge and heterogeneous Lithuania, which then comprised most of Ukraine and Belarus. The threat posed by the Teutonic Knights to both Poland and Lithuania, and the aspiration of the Poles to achieve status as a great power, figured prominently in the calculations. The church and Jadwiga, who was later beatified, attached great importance to the extension of Christianity. The prospects of opening vast regions in the east for trade and settlement appealed to the lords and merchants of Kraków. In 1385 the negotiations were finalized through the Union of Krewo. Jagiełło accepted Roman Catholicism for himself and Lithuania proper—the other duchies were already Christian (Eastern Orthodox)—and promised to join (*applicare*) his Lithuanian and Ruthenian lands to the Polish crown. He became the king of Poland under the name of Władysław II Jagiełło upon marrying Jadwiga, with whom he at first ruled jointly.

The Union of Krewo

THE STATES OF THE JAGIELLONIANS

The waning of the Middle Ages. *The rule of Jagiełło.* The Polish clergy played a major role in the long process of Christianization, and Polish knights assisted Lithuania in its military campaigns; nevertheless, the Lithuanians were determined not to tolerate Polish interference. It was obvious that a simple incorporation of Lithuania into Poland was not possible. Jagiełło's cousin Vytautas (Polish: Witold), who eventually controlled the various Lithuanian duchies, assumed the title of grand duke and made Lithuania a virtually independent state. He even aimed at a royal crown for himself. The defeat he suffered at the hands of the Tatars at Worskla in 1399, however, destroyed his plans. The union with Poland was renegotiated on the basis of partnership of two sovereign states under the reign of Władysław II, king and supreme duke.

The continuing struggles with the Teutonic Knights seeking to master eastern Lithuanian Samogitia (Polish: Żmudź)—on the pretext of Christianizing its inhabitants—led to the great war in which Poland and Lithuania joined forces. The result was a crushing defeat of the Knights at Tannenberg (Grunwald) in 1410. The victory had no immediate sequel, for the Knights ceded only and temporarily Samogitia, but it marked the beginning of their decline; the Prussian nobles and towns secretly opposed the ruthless rule of the Teutonic Order. Polish tolerance was manifest at the Council of Constance (1414–18), where the prominent theologian and rector of Kraków University, Paweł Włodkowic (Paulus Vladimiri), denounced the Knights' policy of conversion by the sword and maintained that the pagans also had their rights. Similarly, the Poles were sympathetic to Jan Hus of Bohemia, who was condemned as a heretic by the council, and lent discreet support to his followers, the Hussites, in their struggles against the Holy Roman empire and the papacy.

Because of his concerns over dynastic succession, Władysław II, who had no children with Jadwiga, granted new privileges to the *szlachta* (all those of noble rank). Called *neminem captivabimus* (comparable to habeas corpus), the measure guarded against arbitrary arrest or confiscation of property and distinguished between the executive and judiciary. In the magnate-dominated Lithuania, the Polish example began to affect its internal evolution. The lesser boyars, envious of the position of their Polish counterparts, favoured closer unity. At the Union of Horodło (in 1413), Polish nobles offered their coats of arms to a number of Lithuanians as a gesture of solidarity.

The Union of Horodło

Only Władysław's fourth wife, Sophia Hołszańska, bore him a male heir. One of their sons, Władysław III Warneńczyk, ruled Poland (1434–44) under the regency of the powerful Zbigniew Cardinal Oleśnicki; the other son, Casimir, was the grand duke of Lithuania. Largely because of Oleśnicki, Władysław III was elected king of Hungary, became active in crusades against the Turks, and, after initial victories, died at the Battle of Varna in 1444. Casimir subsequently became the ruler of both Poland and Lithuania.

Casimir IV. The long and brilliant reign of Casimir IV Jagiellonian (1447–92) corresponded to the age of "new monarchies" in western Europe. By the 15th century Poland narrowed the distance separating it from western Europe and became a significant factor in international relations. The demand for raw materials and semifinished goods stimulated trade (producing a positive balance) and contributed to the growth of crafts and mining. Townspeople in Poland proper constituted about 20 percent of the population—roughly the European average. Divisions between the nobles, burghers, and peasants were still somewhat fluid. Coexistence of the ruler and the estates was relatively smooth and stable. Cultural progress was striking, the reconstituted and enlarged University of Kraków playing a major role. Humanist trends found a promoter at Kraków in the Italian scholar Filippo de Buonacorsi, known as Callimachus. From the pen of Jan Długosz came the first major history of Poland.

Casimir's foreign policy centred on the conflict with the Knights and succession in Hungary and Bohemia. When the rebellious Prussian towns and nobility turned to Casimir, he decreed an incorporation of the Knights' state into Poland (1454). Unable to decisively defeat the Teutonic Order during the Thirteen Years' War (1454–66), he had to sign the compromise Treaty of Toruń in 1466. Gdańsk Pomerania, renamed Royal Prussia and endowed with far-reaching autonomy, became Polish once again. This opened the route to the Baltic. The other territories (most of the future East Prussia) remained with the Knights, albeit as a Polish fief.

Casimir's dependence on the noble levies in wartime enabled the *szlachta* to extract new concessions. They culminated in the Privilege of Nieszawa (1454), which gave the provincial diets (*sejmiki*) the right to declare the levies and raise new taxes. In 1493–96 a bicameral general diet (Sejm) marked the beginning of Polish parliamentarism. The representatives of the *sejmiki* formed the lower house, while the king's appointees constituted the senate.

The question of succession in Bohemia and Hungary was resolved toward the end of the 15th century, when one of Casimir's sons, Vladislas II, was elected to the throne of Bohemia in 1471 and Hungary in 1490. His other sons, John I Albert and Alexander, succeeded each other in Poland and Lithuania from 1492 to 1506. A Jagiellonian bloc had come into existence, but its effectiveness was marred by the fact that the four countries were guided by divergent interests and faced different problems.

The golden age of the Sigismunds. *Political developments.* Under the last two Jagiellonians Poland reached its apogee. The king was the source of law (usually in tandem with the Sejm), supreme judge, chief executive, and supreme commander, free to declare war and peace. He ruled Lithuania as a hereditary domain. Royal administration was quite effective, and Casimir's youngest son, Sigismund I the Old (ruled 1506–48), tried to improve the nation's finances and taxation. An inadequate financial base and an undersized standing army limited the actual power of the king.

Magnate-gentry competition

Domestic politics under Sigismund—and even more so under his son and successor, Sigismund II Augustus (1548–72)—centred on a contest between the fast-growing magnate oligarchy and the dynamic gentry, with the rulers generally favouring the former. The option of relying on the burghers, as was done by western European rulers, was not available because the towns (Gdańsk and some Royal Prussian towns excepted) allowed themselves to be eliminated from political struggles. The reformers among the lesser nobles focused on the program of the "execution" (enforcement) of laws that prohibited the transfer of crown lands and the accumulation of offices that profited the oligarchy. Wishing to emancipate itself from the magnates' political tutelage, the gentry strove for real partnership in government. The Nihil Novi constitution (1505) achieved some of these aims, but it also stipulated that no new laws could be passed without the consent of the Sejm. The way was opened for parliamentary predominance that would eventually undermine the existing system of checks and balances. The growing political and economic power of the landowners had pernicious effects on the lot of the peasantry. Beginning with edicts issued in 1496 and repeatedly throughout the 16th century, the peasants' mobility was curtailed, labour obligations (corvée) were increased, and subjection to the lord's jurisdiction was affirmed. The degree of evasion of these burdens, however, was probably high. The impoverishment of the peasantry from the late 16th century became pronounced, while the barriers between the burghers and the *szlachta* became more rigid.

Foreign affairs. Sixteenth-century foreign policies had to take into consideration an alliance between the Habsburgs, Moscow, and the Teutonic Order that was directed against Poland. Muscovite expansion threatened Lithuania, and only a major victory at Orsza in 1514 averted a catastrophe. The victory allowed Sigismund I to detach the Habsburgs from Moscow through the Vienna accords of 1515. Providing for dynastic marriages, those accords opened the way for Habsburg succession in Bohemia and Hungary should the Jagiellonians die out. Eleven years later Louis II, the Jagiellonian king of Hungary and Bohemia, perished at Mohács fighting the Turks. The Jagiellonian bloc was no more.

One year before Mohács the matter of the Teutonic Knights found a controversial resolution. The grand master of the order, Albert Hohenzollern, became a Lutheran and, isolated from the empire and papacy, offered to secularize his state as a vassal of the king of Poland. His act of homage in 1525 seemed to leave the way open for the eventual absorption of Ducal Prussia (as East Prussia was thereafter called) into Poland. However, subsequent concessions to the Hohenzollerns allowed them to rule both Prussia and Brandenburg, on the flanks of the "corridor" that provided Polish access to the Baltic.

Polish concern with Baltic issues led to the creation of a small navy and to wars with Muscovy over Livonia (present-day Latvia and Estonia), which was controlled by the Knights of the Sword and coveted by Ivan IV the Terrible. Eventually the region was partitioned. The union of 1561 brought the southern part to Poland and Lithuania and established a duchy of Courland ruled as a Polish fief by the former grand master of the secularized Knights of the Sword. Meanwhile, Sweden expanded in the north.

Since Sigismund Augustus was childless, the future of the Polish-Lithuanian union became a paramount issue. Lithuanian socioeconomic, legal, and administrative structures came to resemble those of Poland, but the Lithuanians still opposed incorporation. Their gentry, seeking the privileges of the Polish *szlachta,* wanted a real union, but the powerful magnates opposed it. Fear of discrimination on religious grounds on the part of the Orthodox gentry (in Ukraine and the region that would become modern-day Belarus) was dispelled by granting them equality. The opposition of the magnates was finally broken by the king, who detached Podlasie, Volhynia, and the Kiev and Braclaw regions from Lithuania and incorporated them into Poland. Facing complete annexation, the Lithuanian opposition gave in. The Union of Lublin (1569) established a federative state of two nations with a jointly elected mutual king-grand duke and legislature (a unique feature in Europe) and a customs union but with separate territories, laws, administrations, treasuries, and armies.

The Union of Lublin

Social and cultural developments. Polish culture continued to flourish. Renaissance art and architecture, promoted by Sigismund I's wife, Bona Sforza, became the style for numerous churches and castles. From Kraków University came Nicolaus Copernicus, who revolutionized astronomical concepts. After 1513 a large number of books were printed in Polish, including translations of the Bible. Mikołaj Rej became the father of Polish literature, and Jan Kochanowski one of its greatest poets. The Renaissance and the Reformation had a major impact on Lithuania, marking its absorption into western European culture.

Under the tolerant policies of Sigismund II, Lutheranism spread mainly in the cities, and Calvinism among the nobles of Lithuania and Little Poland. The Sandomierz Agreement of 1570, which defended religious freedom, marked the cooperation of Polish Lutherans and Calvinists. The Polish Brethren (known also as Arians and Anti-Trinitarians) made a major contribution by preaching social egalitarianism and pacifism. In 1573 the *szlachta* concluded the Warsaw Confederation, which provided for the maintenance of religious toleration. These victories for the Reformation, however, were gradually cancelled by the Catholic Counter-Reformation under the leadership of Stanisław Cardinal Hozjusz. In the 1560s the Jesuits arrived in Poland (their greatest preacher was Piotr Skarga), and their network of schools and colleges included the future University of Wilno (now Vilnius, Lithuania), founded in 1579.

THE COMMONWEALTH

Báthory and the Vasas. *Social and political structure.* Called Respublica, which in Polish (Rzeczpospolita) means "Commonwealth," the dual Polish-Lithuanian state was one of the largest in Europe. While Poland in the mid-16th century occupied an area of about 100,000 square miles (260,000 square kilometres), with some 3.5 million inhabitants, the Commonwealth at its highest point in the early 17th century comprised nearly 400,000 square miles and some 11 million inhabitants. A multiethnic country with Poles, Lithuanians, Ukrainians, Belarusians, Germans, Jews, and small numbers of Tatars, Armenians, and Scots, it also contained Roman Catholics, Protestants, Eastern Orthodox, Jews, and Muslims. Certain communities lived under their own laws; the Jews, for example, enjoyed self-administration through the Council of the Four Lands.

The term Poland was used both for the entire state and the strictly Polish part of it (though the latter was officially called the Crown). This could be confusing. A supranational term like "British" was missing. The Commonwealth gradually came to be dominated by the *szlachta,* which regarded the state as an embodiment of its rights and privileges. Ranging from the poorest, landless yeomen to the great magnates, the *szlachta* insisted on the equality of all its members. As a political nation it was more numerous (8–10 percent) than the electorate of most European states even in the early 19th century.

The *szlachta*

Throughout most of Europe the medieval system of estates evolved into absolutism, but in the Commonwealth it led to a *szlachta* democracy inspired by the ideals of ancient Rome, to which parallels were constantly drawn. The *szlachta* came to see in its state a perfect constitutional model, a granary for Europe, and a bulwark against eastern barbarism. Its inherent weaknesses in finance, administration, and the military were ignored.

The end of the Jagiellonian dynasty meant the beginning of unrestricted election to the throne. The first king elected *viritim* (*i.e.*, by direct vote of the *szlachta*) was Henry of Valois, the brother of the king of France. On his accession to the throne (1573–74), which he quickly abandoned to become Henry III of France, he accepted the so-called Henrician Articles and Pacta Conventa. Presented henceforth to every new king as a contract with the noble nation, the former document provided for free election (but not during the reigning monarch's lifetime), religious peace, biennial meetings of the Sejm (with a standing body of senators active in the interval), and the right to renounce the allegiance to the king should he break the contract.

Stephen Báthory. In 1576 the prince of Transylvania, Stephen Báthory (Batory), became king. A brilliant soldier, he closely cooperated with Jan Zamoyski, chancellor of the Crown and grand hetman (commander in chief). The most spectacular achievement of Báthory's reign was a series of military victories (1579–81) over Ivan IV the Terrible of Russia. Yet, it is likely that the king's eastern policies were inspired by the ultimate goal of liberation of Hungary, which was not necessarily a Polish concern.

Sigismund III Vasa. The long reign of his successor, Sigismund III Vasa (1587–1632), raised hopes of a union with Sweden that would strengthen Poland's standing in the north. Sigismund was the grandson of the legendary Swedish ruler Gustav I Vasa, but, as an ardent Roman Catholic and champion of the Counter-Reformation, he was unable to hold on to the crown of Lutheran Sweden, and a 10-year succession struggle ensued. His attempts to secure the throne involved Poland in a series of wars with Sweden. Although one of Lithuania's great military commanders, Jan Karol Chodkiewicz, triumphed at Kirchholm (1605) and the navy defeated the Swedish fleet near Oliwa (1627), the truce that followed was inconclusive. The same was true for most settlements in foreign and domestic affairs. Although Poland remained neutral in the Thirty Years' War (1618–48), Sigismund stealthily supported the Habsburgs, a policy that contributed to a war with Turkey. Poland suffered a major defeat at Cecora in 1620 but was victorious at Chocim (now in Moldova) and negotiated peace a year later.

There was, however, no real peace with Muscovy, then going through its Time of Troubles. The support extended by some Polish magnates to the False Dmitry (who claimed to be the son of Ivan IV the Terrible) eventually embroiled Poland in hostilities. The victory at Klushino in 1610 by hetman Stanisław Żółkiewski resulted in a Polish occupation of Moscow and the election by Moscow's boyars of Sigismund's son Władysław as tsar. Sigismund's veto wasted this opportunity and instead left a residue of Russian hatred of Poland.

The 1606–08 *rokosz*

Suspicions that Sigismund's policies were guided by his dynastic interests contributed to the 1606–08 *rokosz* ("rebellion"). Accusing the king of absolutist designs, the *rokosz* brought together sincere reformers (who demanded the "execution" of the laws), Roman Catholics, and Protestants, as well as magnates pursuing their own ends. Although the royal forces triumphed in battle, both the king and the reformers were losers in the political realm to the magnates posing as defenders of freedom.

Władysław IV. Władysław IV Vasa (1632–48) continued his father's policy of strengthening the monarchy and of pursuing the Swedish throne. Some of the bellicose plans he formulated to increase his power were thwarted by the Sejm and by international circumstances. The anti-Turkish crusade he planned, however, in which Cossacks were to play a major role, contributed to the upheaval that shook the Commonwealth between 1648 and 1660—the uprising in Ukraine and war in the northeast.

Transferred as a result of the Union of Lublin from the grand duchy of Lithuania to the more ethnically homogenous Crown, Ukraine was "colonized" by both Polish and Ukrainian great nobles. Most of the latter gradually abandoned Orthodoxy to become Roman Catholic and Polish. These "little kings" of Ukraine controlled hundreds of thousands of "subjects" and commanded armies larger than those of the regular Crown troops. In 1596 the Union of Brest-Litovsk subordinated the Eastern Orthodox church of the Commonwealth to the papacy by creating the Eastern-rite (Uniate) church. Politically, this was intended to cement the cohesion of the state vis-à-vis Moscow; instead it led to internal divisions among the Orthodox. The new church became a hierarchy without followers while the forbidden Orthodox church was driven underground. Władysław's recognition of the latter's existence in 1632 may have come too late. The Orthodox masses—deprived of their native protectors, who had become Polonized and Catholic—turned to the Cossacks.

The Cossacks. The Zaporozhian Cossacks were frontiersmen who organized themselves in a self-governing centre at modern Zaporizhzhya, Ukraine, first to resist Tatar raids and then to plunder as far away as Constantinople (modern Istanbul). Their prowess was recognized by Sigismund Augustus and Báthory, who "registered" a number of Cossacks for military duty. This privileged status was coveted by other Cossacks and all those diverse groups of settlers or tenants whom the lords tried to turn into serfs. The heavy-handed behaviour of the "little kings," bent on realizing maximum profits and employing Jews as middlemen and overseers, was resented even by small nobles and burghers. Growing socioeconomic antagonisms combined with religious tensions.

Polish-Turkish war of 1620–21

In the Polish-Turkish war of 1620–21, the victory in the Battle of Chocim had been largely due to the participation of some 40,000 Zaporozhian Cossacks, whom Petro Konashevych-Sahaydachny had brought to aid the Poles. Nonetheless, some 12 years later, Cossack demands to be placed on an equal footing with the *szlachta* were contemptuously rejected by the Sejm. The king and the magnates needed the Cossacks in wartime but feared them as an unruly and seditious group that was embroiling the Commonwealth in hostilities with Turkey and the Tatars. Complaints about the enlargement of the military register and about mistreatment led to several Cossack uprisings. After that of 1638 was put down by Polish troops, Cossack privileges were greatly curtailed.

The undertaking of an anti-Turkish crusade opened new vistas. There was talk of massive Cossack participation provided that some 20,000 men be "registered," social grievances redressed, and a military border free of Polish troops established. Whatever the exact encouragements proffered by Władysław IV, the Sejm and the *szlachta* were adamantly opposed and frightened lest the king use the Cossacks for his own ends.

Bohdan Khmelnytsky. In 1648 Bohdan Khmelnytsky, whom contemporaries likened to Oliver Cromwell, assumed the leadership of the Zaporozhian Cossacks and, allied with the Tatars, defeated the troops of the Commonwealth and some magnate contingents. Khmelnytsky became the master of Ukraine, and peasants, townspeople, and even lesser noblemen were among his followers. The city of Kiev hailed him as prince and defender of the Orthodox faith. His objective became the creation of a separate Ukraine under the direct rule of a king. In Poland, where the sudden death of Władysław IV left the country leaderless, a policy of compromise represented by the chancellor, Jerzy Ossolinski, and the last Orthodox senator, Adam Kisiel (Kysil), clashed with warlike operations of the leading "little king," Prince Jeremi Wiśniowiecki. The nature of temporary agreements, which intervened between the Commonwealth and the Cossacks, varied depending on the changing fortunes of war. The Polish victory at Beresteczko in 1651 was followed by the pact of Biała Cerkiew, which the Cossacks found hard to accept. In 1654 Khmelnytsky submitted to Tsar Alexis in the Pereyaslav Agreement. Russian historiography characterizes that agreement as the reunification of Ukraine with Russia; the Ukrainians interpret it as an alliance based

on expediency. At any rate, war began between Muscovy and the Commonwealth, and Alexis' armies drove deep into Lithuania. In 1655 they occupied its capital, Wilno. For the first time in nearly two centuries an enemy invasion had taken place, and, when it was followed by a Swedish aggression, a veritable "deluge" overtook the Commonwealth.

John II Casimir Vasa. Charles X Gustav of Sweden worried lest Muscovy upset the balance of power in the Baltic, which he aimed to turn into a Swedish lake. The refusal of King John II Casimir Vasa, the successor and brother of Władysław IV, to give up his claims to the Swedish crown offered a good pretext for resuming hostilities with the Commonwealth. Aiming originally to seize Polish and Prussian harbours, Charles Gustav saw, after the first successes, the possibility of gaining the Polish crown and the mastery of the Commonwealth.

Swedish intervention

The magnates and gentry of Great Poland capitulated to the Swedes in July 1655. Prince Janusz Radziwiłł, a leading Calvinist and the greatest magnate of Lithuania, hard-pressed by the Russians, broke off the union with Poland and signed one with Sweden. His motives were a combination of Lithuanian and Protestant interests coloured by his own ambition to rule the grand duchy.

The nearly bloodless conquest of the huge Commonwealth came as a shock to many. Yet, Polish resistance to what turned out to be a regime of brutal occupation developed very quickly. The successful defense of the fortified monastery of Jasna Góra (now in Częstochowa) became a rallying point and provided a symbolic religious-ideological banner. Although the Poles were seldom a match for the Swedish professional troops, they excelled at partisan warfare and at winning minor battles. Not only the *szlachta* but also the peasants fought the foreigner and enemy of Roman Catholicism. Stefan Czarniecki became the hero of the war. Returning from exile in Silesia, John Casimir built an international coalition against the Swedes, whose successes were upsetting the balance of power. A cease-fire intervened on the Russian front, and the Cossacks were neutralized by the Tatars, while the Habsburgs, Denmark, and Brandenburg-Prussia came to Poland's aid. The Swedes were gradually driven out of the Commonwealth, despite an armed intervention on their side by Transylvania's Prince György II Rákóczi, who aspired to the Polish crown. The war ended with the Treaty of Oliwa (1660), which restored the territorial status quo before the Swedish invasion and brought the final renunciation of John Casimir's claim to the crown of Sweden.

The real winner in the conflict proved to be Frederick William, the elector of Brandenburg and duke of Prussia. Adroitly maneuvering between Sweden and Poland and extracting a price for his collaboration from both sides, the "Great Elector" finally switched his support to John Casimir, thereby receiving the recognition of full sovereignty over Prussia for himself and his male descendants through the treaties of Wehlau (Welawa) and Bromberg (Bydgoszcz) in 1657.

Eastern wars still continued for Poland for several years. In Ukraine the Hadziacz agreement of 1658 with Khmelnytsky's successor provided for the creation of a Ukrainian state as a third member of the Commonwealth with its own offices and army, as well as mass ennoblements of Cossacks and the suspension of the Union of Brest-Litovsk. But, denounced by a pro-Russian faction in Ukraine and nullified, the accord led to a renewal of hostilities with Muscovy, which ended in 1667 with the Truce of Andrusovo. Restoring the occupied parts of Lithuania to the Commonwealth, the truce divided Ukraine along the Dnieper River. Together with the Treaty of Oliwa, that agreement was a turning point that marked the decline of the Commonwealth's international standing.

The 17th-century crisis. *Social and economic changes.* The two decades of war and occupation in the mid-17th century, which in the case of Lithuania gave a foretaste of the 18th-century partitions, ruined and exhausted the Commonwealth. Famines and epidemics followed hostilities, and the population dropped from roughly 11 to 7 million. The number of inhabitants of Kraków and Warsaw fell by two-thirds and one-half, respectively. Wilno had been burned down. The Khmelnytsky uprising decimated the Jews in Ukraine, even if they recovered fairly rapidly demographically. With many farm buildings and farming implements as well as cattle destroyed, and as a result of labour shortages, the productivity of agriculture diminished dramatically. The dynamic network of international trade fairs collapsed. Grain exports, which had reached their peak in the early 17th century, could not redress the unfavourable balance of trade with western Europe. Losses in art treasures—the Swedes engaged in systematic looting—were irreplaceable.

Economic collapse

The Commonwealth never fully recovered, unlike Muscovy, which had suffered almost as much during the Time of Troubles. Twentieth-century Marxist historians blamed the manorial economy based on serf labour for pauperizing the masses and undermining the towns, yet the Polish economy was not unique in that respect. Moreover, some attempts to replace serfs with rent-paying tenants did not prove to be a panacea. The economic factor must therefore be treated jointly with other structural weaknesses of the Commonwealth that militated against recovery. The 17th-century crisis—a European phenomenon—was basically a crisis of political authority. In the Commonwealth the perennial financial weakness was the central issue. The state budget in the second half of the century amounted to 10–11 million złotys as compared with the equivalent of about 360 million in France or 240 million in England. About nine-tenths of it went for military purposes, compared with half in Brandenburg and more than three-fifths in France and Russia. Equating a large army with royal absolutism and extolling the virtue of noble levies, the *szlachta* was unwilling to devise defensive mechanisms. This was true even after the chastising experience of the Swedish "deluge." Most nobles contented themselves with invoking the special protection of St. Mary, symbolically crowned Queen of Poland, as a sufficient safeguard.

Political stagnation. Those wishing to reform the state without strengthening the monarchy wanted to make the Sejm an effective centre of power. The *szlachta*, however, refused to accept the notion that liberty could be better preserved in a stronger state. In 1652 the notorious and often misunderstood practice of liberum veto (free veto) appeared: a single negative vote by a member of the Sejm was sufficient to block the proceedings. It was argued that unanimity was essential for passing laws, for deputies as representatives of the local *sejmiki* were bound by instructions. Moreover, a majority could disregard local interests and be corrupted by the administration. Hence, liberum veto came to be regarded as the kernel of liberty and a safeguard against tyranny. In reality, the dissenting deputy was usually an instrument of a magnate or of the king.

The liberum veto could paralyze the functioning of the state, and in the 17th century it was used sparingly. The weakening of the Sejm meant that some of its functions, notably in matters of taxation, had to pass to local *sejmiki*. Without a central bureaucracy and with a dual structure of offices in the Crown and Lithuania, the fragmentation of sovereignty became increasingly ominous. The attempts at reform by John Casimir and his wife, Marie Louise, may have been ill-conceived, but, given the factional strife within the oligarchy, it was difficult for the monarch to find a stable base of support. The *szlachta,* ever suspicious of anything that could smack of absolutism, was naturally opposed. The royal plans were defeated by a *rokosz* in 1665–66 led by Marshal Jerzy Lubomirski. Two years later the frustrated John Casimir abdicated and settled in France, having warned the Sejm that Poland would fall victim to its rapacious neighbours unless it reformed its ways.

Cultural changes. The prevalent mentality in the Commonwealth in the 17th century manifested itself in Sarmatism. The name came from alleged ancestors of the *szlachta* (Sarmatians), and the concept served to integrate the multiethnic nobility. Representing a symbiosis of a political ideology and a lifestyle typical of a landowning, rather provincial, tightly knit, and increasingly xenophobic culture, Sarmatism extolled the virtues of the *szlachta* and contrasted them with Western values. An Orientalization of Polish-Lithuanian culture was occurring. Ro-

Sarmatism

man Catholicism was Sarmatized in its turn, assuming a more intolerant posture toward other denominations. The struggles against Lutheran Swedes and Prussians, Orthodox Russians, and Muslim Turks and Tatars strengthened the belief in Poland's mission as a Catholic bastion. The expulsion in 1658 of Polish Brethren—accused of collaboration with the Swedes—when taken together with the virtual elimination of non-Catholics from public offices, was the first harbinger of the Pole-Catholic syndrome (the notion that a true Pole must be a Catholic).

Decline and attempts at reform. The Lubomirski *rokosz* was barely over and truce with Muscovy newly signed when the Cossacks in the Polish part of divided Ukraine submitted to Turkey and called for Tatar aid against Poland. Victories won by hetman Jan Sobieski only temporarily forestalled the threat, and in 1672 the Commonwealth faced a major invasion by Turkey. The fall of the key border fortress Kamieniec Podolski was followed by the humiliating Peace of Buczacz. The Commonwealth lost the provinces of Podolia and Bratslav and part of Kiev, and it had to pay a tribute to the Sublime Porte. Sobieski's victory over the Turks at Chocim in 1673 was not exploited, because of the lack of financial means, but it paved the way for Sobieski's election to the Polish throne. His predecessor, who had followed John Casimir—Michael Korybut Wiśniowiecki—reigned for four years only (1669–73) and proved utterly incapable. Sobieski, ruling as John III (1674–96), at first considered conquering Prussia in alliance with France. But that plan did not succeed. With the papacy and the Habsburgs preparing for all-out war against Turkey, John reverted to an anti-Turkish policy and concluded an alliance with Austria. In 1683 he led a relief army to a Vienna besieged by the Turks and as commander of the allied forces won a resounding victory that marked the beginning of Turkish withdrawal from Europe. The Commonwealth, however, did not share in the subsequent victorious Austrian campaigns. Poland became a secondary partner, and, when the final peace with Turkey was concluded in Carlowitz (modern Sremski Karlovci, Yugos.) in 1699, the Poles recovered only the lost Ukrainian lands. By that time John was no longer alive, and Augustus II, the elector of Saxony, had succeeded him on the throne (1697–1733).

The withdrawal of the Turks

The Saxons. The "Saxon Era" lasted for more than 60 years and marked the lowest point in Polish history. Research since the 1980s has somewhat corrected the largely negative picture of Augustus II and Augustus III by stressing the context in which they operated: anarchic political life dominated by factions of struggling oligarchs and open to meddling by the neighbouring powers. The latter signed agreements among themselves to promote weakness within the Commonwealth, as for instance the Austro-Russian accord of 1675 and the Swedish-Brandenburg pacts of 1686 and 1696, which were followed by others in the 1720s. Foreign interlopers corrupted politicians and fomented disorder. During the reign of Augustus II, 10 out of 18 Sejms were paralyzed by liberum veto. When a Protestant-Catholic riot in Toruń in 1724 was followed by capital sentences imposed on the Protestant officials, Prussian and Russian propaganda spoke of a "bloodbath" and denounced Polish intolerance. Posing as a protector of non-Catholics, St. Petersburg was in fact using them as a political instrument. Polish politics, ways, and manners were increasingly pictured as exotically anachronistic. The Polish nobles became the laughingstock of Europe. Because the promises made by John Casimir during the darkest days of Swedish invasion to improve the lot of the peasantry had remained empty, the oppressed peasants were largely alienated from the nation.

Augustus II. A personal union with Saxony, where Augustus II was a strong ruler, seemed at first to offer some advantages to Poland. A king with a power base of his own might reform the Commonwealth, which was still a huge state and potentially a great power. But such hopes proved vain. Pursuing schemes of dynastic greatness, Augustus II involved unwilling Poland in a coalition war against Charles XII of Sweden that proved disastrous. In 1702 Charles invaded, forced Augustus out, and staged an election of the youthful Stanisław I Leszczyński as king. The country, split between two rival monarchs, plunged into chaos. The slowly proceeding demographic and economic recovery was reversed as the looting armies and an outbreak of bubonic plague decimated the people. A crushing defeat of Sweden by Peter I the Great of Russia at Poltava, Ukraine, Russian Empire, in 1709 eventually restored Augustus to the throne but made him dependent on the tsar. Having failed to strengthen his position by war and territorial acquisitions, Augustus contemplated domestic reforms while his entourage played with the idea of a coup backed by Saxon troops. Peter intervened as an arbiter between the king and his noble opponents. A settlement at the "silent Sejm" surrounded by Russian troops removed Saxon contingents from Poland, but it brought about certain reforms. Subsequent attempts by Augustus to mount a coalition against the rising might of Russia foundered on the distrust of the king's motives. He was even suspected of plotting partitions of the Commonwealth. During the remaining years of his reign, Augustus' main preoccupation was to ensure the succession of his son.

The "silent Sejm"

Augustus III. Upon Augustus' death in 1733, Stanisław I, seen this time as a symbol of Poland's independence and supported by France, was elected once again. The counter-election of Augustus III followed, and Russian troops drove Stanisław out of the country. He abdicated, receiving as compensation (after the so-called War of Polish Succession) the duchy of Lorraine.

The reign of Augustus III (1733–63)—during which 5 out of 15 Sejms were dissolved while the remainder took no decisions—witnessed the nadir of Polish statehood. The Commonwealth no longer could be counted as an independent participant in international relations; the king's diplomacy was conducted from Dresden, in Saxony. Poland passively watched the once-Polish Silesia pass, as a result of the War of Austrian Succession, from the Habsburgs to Prussia. The latter, under Frederick II the Great (whose grandfather had already been recognized in 1701 as "king in Prussia" by Augustus II), was becoming a great power. During the Seven Years' War (1756–63) Austrian and Russian troops marched through Poland, and Frederick flooded the country with counterfeit money. The Commonwealth was being treated as a wayside inn.

And yet there were also the first signs of economic recovery and population growth, the beginnings of a transition from Sarmatism to Enlightenment, and the appearance of a reformist political literature. In 1740 Stanisław Konarski, a member of the Roman Catholic Piarist teaching order, founded the Collegium Nobilium, which was to train the future elite. A network of Piarist schools followed, while the ideas of the Enlightenment were being spread often through Freemasonry. Konarski's writings, as well as those coming from the circle of Stanisław I in Lorraine, attacked the liberum veto and advocated an improvement of the lot of towns and peasantry. After the 1740s two major camps were emerging: the so-called Familia, led by the Czartoryskis, and the Republicans, with the Potockis and Radziwiłls at their head.

Reforms, agony, and partitions. *Reform under Stanisław II.* The election of Stanisław II August Poniatowski as the last king of Poland (1764–95) was the work of the powerful Familia. Rising from middle nobility (though his mother was a Czartoryska), the candidate was handpicked by Catherine II the Great of Russia not only because he had been her lover but because she felt that he would be completely dependent on her. The Czartoryskis in turn saw him as their puppet. Thus, from the beginning Stanisław II, a highly intelligent man, a patron of arts, and a reformer in the spirit of the Enlightenment, had to operate under most difficult conditions. The magnates resented him as an upstart; the conservative *szlachta* viewed him as Catherine's tool and as a threat to their liberties. The king's adroitness and personal charm allowed him in time to win over some of his adversaries, but he lacked strong will and showed none of the military inclination so cherished by the Poles.

The role of Catherine II

The reforms that accompanied the election were limited. Stanisław sought to reform the state by strengthening the monarchy; the Czartoryskis wished to reform it by

strengthening the Sejm. The king embarked on a vast program of modernization, encouraging initiatives in the economic, financial, and military spheres. But above all he waged a nationwide campaign using the press, literature, and the new National Theatre to change the conservative mentality of the *szlachta*. In 1765 Stanisław established the Knights' School, the first truly secular college, which promoted civil virtues and religious toleration, and criticized the treatment of towns and peasantry. The king's policies, however, were constantly undermined by neighbouring powers. Frederick II's view that Poland ought to be kept in lethargy was shared by St. Petersburg, which sought to isolate Stanisław by encouraging both religious dissenters (*i.e.,* non-Catholics) and the conservative circles to form confederations. The presence of Russian troops terrorized the Sejm, and Russia formally guaranteed as immutable such principles of Polish politics as liberum veto, elective monarchy, and dominance of the *szlachta*.

The First Partition. In 1768 the Confederation of Bar was formed. Its antiroyalist and anti-Russian program mingled patriotic and conservative overtones with religious objectives (namely, the defense of the privileged status of Roman Catholicism vis-à-vis the religious and political equality for non-Catholics advocated by Russia). Civil war erupted, lasting until 1772. Royal troops assisted the Russians, and France and Turkey helped the confederates. The movement strengthened Polish national consciousness and produced the first martyrs sent to Siberia, but at the same time it created such chaotic conditions that St. Petersburg began to listen when Frederick repeatedly proposed partitioning Poland. With Russia and Austria on the brink of war over Turkish matters, Berlin suggested a resolution of the eastern crisis through mutually agreeable compensations at Poland's expense. Austria, which had opposed the scheme, unwittingly created a precedent by annexing some Polish border areas.

As a result of the First Partition (1772), Poland lost almost one-third of its territory and more than one-third of its population. Russia received the largest but least important area, in the northeast. Austria gained the densely populated Little Poland (renamed Galicia). Prussia's share was the smallest, but the annexation of Eastern Pomerania (without Gdańsk) cut off Poland from the sea and allowed Frederick to put a stranglehold on the Polish economy. The helpless Sejm, fearing additional territorial losses, ratified the partition. Despite some British concern about Gdańsk and the Baltic trade, the European powers reacted to the partition with utmost indifference. The British political philosopher Edmund Burke was alone in not only criticizing the immorality of the act but also seeing in it the beginning of a revolutionary change in the European balance of power.

European indifference

Social and economic changes. During the two decades that separated the First and Second Partitions, the country experienced a remarkable revival. The dissolution of the Jesuit order in 1773 allowed a complete reorganization of the Polish educational system under one of the first ministries of education in Europe. Cut off from the Baltic, Poland reoriented its trade toward the Black Sea. Producing for the national market, early manufacturing concerns grew on both royal and magnate land. Many estates began to operate with tenant farmers rather than serfs. Banks and joint stock companies appeared, canals were built, and roads improved. The position of the towns began to change, and Warsaw with its 100,000 inhabitants became a centre radiating into the country. Under the king's patronage arts and literature flourished. Learning made important strides. The satiric poet Bishop Ignacy Krasicki headed a long list of important authors. Political literature reached its summit with the writings of Stanisław Staszic (a burgher) and Hugo Kołłątaj. There was discussion of a reform of towns (including a Jewish reform) and changes in the status of the peasantry by extending to them rights and representation as well as state protection. The newly created Permanent Council, a collegial body composed of five ministries, was the first executive organ for both the Crown and Lithuania. Although it was seen as a channel for Russian influence and was attacked by the oligarchic opposition, who believed it strengthened the position of the king, the council marked progress in financial, police, and administrative fields. Stanisław II was convinced that only close collaboration with St. Petersburg constituted a guarantee against further partitions; thus, reforms had to meet with Russian approval. The failure of a projected new code of laws reforming the social system and state-church relations showed the limits of tolerated reform. St. Petersburg seemed to regard its tutelage as firm enough to withdraw its troops from the country in 1780.

The constitution of 1791. A Russo-Turkish war that began in 1787 created a situation that both the king and the magnate opposition tried to exploit. With Prussia proposing an alliance to the Poles (signed in 1790) and Austria becoming preoccupied with the French Revolution, the so-called Great Sejm, which met in 1788, embarked on sweeping reforms. The king aspired to constitutional monarchy, the "patriots" preferred a republic presided over by a monarch, while the die-hard conservatives ("false patriots") opposed all modernization and change. It took the Sejm four years of heated debates, in the course of which the example of the American Revolution was frequently invoked, to enact a new system. The king, although fearful that the Sejm, by going too far, would antagonize Russia, eventually joined with the patriots in approving on May 3, 1791, the first modern written constitution in Europe.

The new constitution was a revolutionary document that created a constitutional parliamentary monarchy and gave a new meaning to the term "political nation." It combined Polish traditions with the ideas of the Enlightenment. Dynasties, not individuals, would henceforth be elected, beginning with the house of Wettin. The king's decrees had to be countersigned by ministers responsible to the Sejm, which was partly elected on the basis of property qualifications. Since burghers gained some political rights and the poorest gentry—clients of the magnates—lost some of theirs, birth alone would no longer be a determinant of citizenship. The liberum veto was abolished, as was discrimination on religious grounds. An army 100,000 strong was to be raised. Royal towns recovered their autonomy, but the peasantry was only taken under the protection of the law. Additional laws applying to social and economic problems were to follow. Although the constitution was passed through a quasi-coup (1791–92), Stanisław gained for it the approval of most of the *sejmiki*. It was, however, unacceptable to Russia and Prussia, both of which were fearful lest a revived, strong Poland reclaim its lost lands. Driven by pride and doctrine, a number of die-hard conservatives—among them high dignitaries such as Stanisław Szczęsny Potocki, Seweryn Rzewuski, and Ksawery Branicki—formed the Confederation of Targowica (in St. Petersburg) to overthrow the May constitution. Acting as guarantor of the old Polish regime, Catherine ordered her armies to invade Poland in 1792. There they fought the outnumbered Polish troops under Prince Józef Poniatowski and General Tadeusz Kościuszko.

The "political nation"

The Second and Third Partitions. Intimidated, the king and the government capitulated; the May constitution was abolished; leading patriots emigrated. All this did not prevent the Second Partition in 1793 by Russia and Prussia. In 1794 Kościuszko, returning from abroad, raised the banner of insurrection in the rump Commonwealth. It may have been a hopeless undertaking, but the Poles could not see their state destroyed without making a last stand. Kościuszko, assuming the title of chief (*naczelnik*), ignored the king, but crowds in Warsaw, inspired by revolutionary France, summarily executed a number of Targowica leaders. Offering emancipation measures to the peasants, Kościuszko brought a large number of them under his banner. After winning the battle of Racławice and capturing Warsaw and Wilno, the insurrectionists were defeated by Russian and Prussian forces. Kościuszko was taken prisoner, and Aleksandr Suvorov's Russian army carried out a wholesale massacre of the population in the Warsaw suburb of Praga. The Third Partition followed in 1795, and, as in the preceding cases, the Polish Sejm was obliged to give its consent. Stanisław abdicated and left for St. Petersburg, where he died. In the final count Russia annexed 62 percent of Poland's area and 45 percent of the population, Prussia 20 percent of the area and 23

percent of the population, and Austria 18 and 32 percent, respectively. The three monarchs engaged themselves not to include Poland in their respective titles, thus obliterating its very name. But, while Poland disappeared, the Polish question, which affected European diplomacy and the growth of nationalism, was born.

The Polish question

PARTITIONED POLAND

The legions and the Duchy of Warsaw. The 123 years during which Poland existed only as a partitioned land had a profound impact on the Polish psyche. Moreover, major 19th-century developments such as industrialization and modernization were uneven in Poland and proved to be a mixed blessing. Growing Polish nationalism was by necessity that of an oppressed nation and displayed the tendency of "all or nothing." Compromise became a dirty word, for it implied collaboration with the partitioners; a distrust of authority grew. The tradition of the Polish nobles' republic militated against submission and engendered an attitude of revolutionary defiance.

Beginning with the Kościuszko Insurrection the Poles staged uprisings in 1806, 1830, 1846, 1848, and 1863 and a revolution in 1905. Defeats were followed by "organic work" that aimed at strengthening the society and its economy by peaceful means. This other major trend of nationalist aspiration was linked with Positivism, while the insurrectionary tradition became closely connected with Romanticism, but it is an oversimplification to identify the former with realism and the latter with idealism.

The survival of the Polish nation, which during the 19th century absorbed the peasant masses, was due in no small degree to a culture that continued to be all-Polish and to the Roman Catholic church, whose role in maintaining "Polishness" was very important. Numerous writers, from the Romantic poets to Henryk Sienkiewicz, winner of the 1905 Nobel Prize for Literature, shaped the Polish mentality. For a stateless nation, ideas and imponderables acquired special importance. A rebirth of statehood, however, could be achieved only under the conditions of a major European upheaval, which would mean a collapse of the partitioning powers; this did not happen until 1918.

Politically conscious Poles did not reconcile themselves to the loss of independence. Conspiracies and attempts to exploit the differences between the partitioning powers arose. Émigrés looked up to revolutionary France for assistance, and General Jan Henryk Dąbrowski succeeded in 1797 in persuading Napoleon Bonaparte to create auxiliary Polish legions. In their headquarters the future Polish national anthem—"Jeszcze Polska nie Zginęła" ("Poland Has Not Yet Perished")—was sung for the first time. Hopes placed on a French victory over Austria that would open the Polish question were, however, quashed by the Treaty of Campo Formio. Polish legionnaires also fought French battles in Germany and in Santo Domingo, but Poland gained no political commitments. Yet their struggles did have a meaning in the long run, keeping a democratic Polish spirit alive and furnishing cadres to a future Polish army under Napoleon.

The pro-French military option had a counterpart in the ideas of Prince Adam Jerzy Czartoryski. Appointed Russian foreign minister by Tsar Alexander I, the prince advocated redrawing the map of Europe to take into account national feelings and reconstitute Poland in union with Russia. This approach failed when Alexander committed to a struggle against France on the side of Prussia. After Napoleon's victories over Prussia in 1806, French troops entered the Prussian part of Poland. Responding to somewhat vague promises by Napoleon, Dąbrowski organized Polish forces that played a significant role in the campaigns that followed, and Napoleon could not avoid making some gesture to the Poles. As a result of the compromise peace with Alexander at Tilsit, Prussia (now Sovetsk, Russia), in 1807, a small state was created out of the Prussian shares in the First and Second Partitions and called the Duchy of Warsaw. Its ruler was the king of Saxony, Frederick Augustus I. Gdańsk was made a free city.

The creation of the Duchy of Warsaw

The Duchy of Warsaw, so named in order not to offend the partitioners, appeared to the Poles as a nucleus of a revived Poland. Doubled in size after a victorious war against Austria in 1809, it numbered more than four million people and had within its borders Warsaw, Kraków, and Poznań. The constitution imposed by Napoleon was comparable to his other authoritarian constitutions but took into account Polish traditions and customs. The ruler was absolute but used his powers with discretion and later delegated them to his ministers. The Napoleonic Code was introduced, and the constitution abolished "slavery." But this was interpreted to imply only the personal emancipation of the peasants without land reform. Hence servile obligations for those who stayed on the land continued in practice.

Napoleon regarded the duchy as a French outpost in the east, which required the maintenance of a disproportionately large army. The costs of maintaining it, together with the adverse effects of the Continental System, brought the duchy's economy to the brink of ruin. The emperor then took some Polish troops on his payroll, and they fought in Spain, where the charge of the light horse guards at Somosierra in 1808 passed into national legend.

Napoleon's invasion of Russia in 1812, in which nearly 100,000 Polish soldiers participated, seemed to promise the re-creation of Poland. Napoleon encouraged the Poles to proclaim the restoration of their country but did not commit himself to that goal. In reality, the emperor waged war not to destroy Russia but to force the tsar back into a policy of collaboration with France. Only in his exile at St. Helena did Napoleon speak of the key importance of Poland. His defeat in Russia brought Russian troops into the Duchy of Warsaw. While other allies of Napoleon were abandoning the sinking ship, Prince Józef Poniatowski, who commanded the Polish army, remained loyal and died fighting at the Battle of Leipzig (1813).

From the Congress of Vienna to 1848. The victory of the anti-Napoleonic coalition led to a redrafting of the map of Europe at the Congress of Vienna (1814–15). The Congress paid lip service to Poland by enjoining the partitioning powers to respect the national rights of their Polish subjects (insofar as was compatible with the state interests) and by providing for free trade and communications within the borders of the old Commonwealth. The latter turned out to be a dead letter. The territorial issue caused dissent among the powers, but eventually a compromise arrangement left the former Duchy of Warsaw, minus Poznania (which went to Prussia) and Kraków (made a free city), to Tsar Alexander under the name of the Kingdom of Poland. The tsar now controlled about two-thirds of the old Commonwealth—both the area commonly called Congress Kingdom, or Congress Poland, and the former Commonwealth (Lithuanian, Ukrainian, and Belarusian) provinces that had been annexed during the partitions.

Congress Poland

Early Russian rule. Endowed with a liberal constitution, which was increasingly violated in practice, the settlement satisfied neither the Poles nor the Russians. The former hoped for the kingdom to be united with the eastern "lost lands" and to become a junior partner of the empire. Alexander, who played with the idea, abandoned it under the pressure of Russian circles that were unwilling to give up any of the annexed provinces. The Wilno educational district, which comprised most of them, had been seen as a model for educational reform in Russia. The university in Wilno was the largest in the empire. In 1823 it came under attack; students accused of sedition were jailed or exiled. One of the victims was the great Polish poet Adam Mickiewicz. Thus, basic disagreement about the territorial question was augmented by the anomalous union of an autocratic empire with a liberal kingdom. In the long run a confrontation may have been inevitable, but it was hastened by a gradual deterioration of the position of the Poles.

The post of viceroy did not go to Prince Czartoryski, by then estranged from Alexander, but to a servile political nonentity, General Józef Zajączek. The tsar's brother Constantine, the brutal grand duke, was made commander in chief. Together with a special representative of the tsar, the intriguing and unscrupulous Nikolay Novosiltsev, they dominated the kingdom while usually at odds with one another. Alexander, autocratic by temperament, was revolted by the phenomenon of a liberal opposition in the



The partitions of Poland, 1772–95.

Sejm, which he regarded as ingratitude. Out of Freemasonry, which the tsar at first patronized, there grew a secret Polish Patriotic Society whose aims could hardly be qualified as treason. Nevertheless, its leader, Major Walerian Łukasiński, became a national martyr when he was thrown into prison, where he languished half-forgotten for more than 40 years until his death. Other conspiracies of more radical character began to spread. The economy of the kingdom, however, developed, and its finances were put in order by Prince Ksawery Drucki-Lubecki. He showed that Congress Poland was not a burden on the empire.

The Decembrist uprising in Russia in 1825, which accompanied the succession to the throne of Nicholas I, had repercussions in Congress Poland. A public trial exonerated the Polish leaders of complicity but made Russo-Polish relations tense. Revolutions in Belgium and France in 1830 hastened the outbreak of the November Insurrection. After its inception as a conspiratorial act at the cadet school in Warsaw (Nov. 29, 1830), this uprising developed into a national revolt, marked by the dethronement of the Romanovs in Poland and the onset of a full-fledged Russo-Polish war. Hostilities spread into Lithuania and lasted until September 1831. Russian victory was followed by severe reprisals. The kingdom's constitution was suspended, which meant the end of a separate Polish Sejm, government, and army. The universities of Wilno and Warsaw were closed. Cultural Russification in the empire's former Polish provinces involved the liquidation of the Uniate church in 1839 and the abolition of the statute that had preserved the Lithuanian code of law. The Uniate church continued to exist only within the Congress Kingdom (until 1875) and in Galicia (until 1945).

The November Insurrection

Emigration and revolt. Several thousand Poles, including the political and intellectual elite, emigrated. When they passed through Germany, these émigrés were hailed as champions of freedom, and many of them came to believe in the idea of the solidarity of nations. The émigrés, settling mainly in France, splintered into many factions but grouped mainly around two figures: the moderate conservatives followed Prince Czartoryski, and the leftists were led at first by the great historian Joachim Lelewel. Later these leftists took a more radical stance as the Polish Democratic Society. Czartoryski concentrated on seeking the support of Britain and France for the Polish cause against Russia. The democrats, distrusting governments and blaming the conservatives for the defeat of the November Insurrection, preached a national and social revolution in cooperation with other peoples that would emancipate the peasantry. The Polish Democratic Society, whose program was embodied in the Poitiers Manifesto of 1836, became the first democratically run, centralized, and disciplined political party of east-central Europe. Karl Marx regarded its concept of agrarian revolution as a major Polish contribution to European revolutionary thought. The writings of the émigrés were imbued with an intense patriotic message. The three greatest Polish Romantic poets—Mickiewicz, Juliusz Słowacki, and Zygmunt Krasiński—were the national "bards" (*wieszcz*) who influenced entire generations of Poles. They were followed by the much-later-discovered poet Cyprian Norwid. In music the emigration was epitomized by the compositions of Frédéric Chopin. A Messianic conception of the Polish nation arose, which in its most extreme and mystical form characterized Poland as the Christ of nations, redeeming all oppressed peoples through its suffering and transcendence.

The Christ of nations

In partitioned Poland, conspiratorial activities were inspired by émigré emissaries. After the failure of several other attempts, an uprising was planned for 1846. Stanched by arrests in Poznań, it got off the ground only in Kraków (where a national government was proclaimed) and in the neighbouring districts of western Galicia. The Kraków rising was put down by Austrian troops, and the city was annexed; elsewhere peasant antagonism toward the landowners was channeled by Austrian officials against

the mostly noble rebels. A jacquerie (peasant revolt) developed, in the course of which many manors were burned down and landowners killed. This came as a shock to Polish democrats who had extolled the people (*lud*) as the backbone and the hope of the nation and to conservatives who had warned against a social upheaval.

The liberal and democratic Revolutions of 1848, which spread over most of Europe, raised hopes for the revival of the Polish cause. Poles were in the forefront of numerous struggles, and General Józef Bem became a hero of the Hungarian Revolution. While the tsar threatened to punish the revolutionaries in Germany and in the Habsburg monarchy, liberals in Berlin and Vienna saw the advantage of a Polish buffer state against Russia. In Prussian Poland the authorities tolerated a virtual Polish takeover of Poznania, including the formation of an armed militia. However, when the Russian danger receded, Polish nationalism appeared as the main threat, and the militia was crushed by Prussian troops. The Germans had opted for "healthy national egoism," which meant that henceforth the Polish strife with the Prussian officialdom would become a nationalist German-Polish struggle.

Some circles in revolutionary Vienna seemed to consider giving up Galicia to a revived Poland. The governor of Galicia was interested mainly in assuring control over the province. Forestalling Polish plans, he abolished serfdom and used the nascent Ruthenian-Ukrainian movement in eastern Galicia to oppose national aspirations. A limited Polish resistance was broken by bombardments of Kraków and Lemberg (now Lviv, Ukraine).

After the Revolutions of 1848, which revealed the sharpness of national conflicts, the Poles began to realize that a Poland within the pre-partitions borders—a smaller Polish state was out of the question not only for Poles but also for Marx and Friedrich Engels—might have to be a federation of distinct nationalities and no longer a unitary country. The emancipation of the peasantry in Galicia made the peasant question a central issue—namely, if the peasants could be absorbed into the Polish national fabric or if their first loyalty would be to the partitioning monarchs. The issue became acute in the Russian partition, which had remained passive in 1848.

The January 1863 uprising and its aftermath. After humiliating defeats in the Crimean War, the Russian Empire under Tsar Alexander II embarked on major liberal reforms. For Congress Poland this meant political amnesty, conciliatory measures in cultural and religious matters, and the creation of an Agricultural Society to tackle the peasant question. Simultaneously, Alexander II warned the Poles against political "daydreaming." The Agricultural Society, a union of reformist landowners headed by the popular Hrabia (count) Andrzej Zamoyski, debated changes in the agrarian sector but found it hard to avoid politics. A patriotic movement later known as the Whites grew around and partly out of the society. It included landowners and members of the bourgeoisie (often of German or Jewish origin), such as the banker Leopold Kronenberg. At this time a Polish-Jewish dialogue promoted close cooperation.

The Whites and the Reds

On the other side of the political spectrum there developed a number of conspiratorial groups composed of students, younger army officers, artisans, and members of the lesser gentry. Subsequently called the Reds, these radicals acted as a pressure group on the Agricultural Society and staged demonstrations commemorating Polish patriots or historic events. In 1861, the year of the peasant emancipation decree in the Russian Empire, demonstrators in Warsaw clashed with Russian troops, and several were killed or wounded.

The Russians, determined to be firm with the radicals, sought a dialogue with the upper classes. But Zamoyski demanded a return to the guarantees of the 1815 constitution. Such demands were rejected, and Zamoyski was eventually ordered to leave the country. The Russian viceroy turned to Zamoyski's rival, Margrabia (margrave) Aleksander Wielopolski, whose program of limited concessions (Polonization of education, restoration of local self-government, transformation of the peasants into tenants, and emancipation of the Jews) was acceptable to St. Petersburg. Wielopolski's contempt for public opinion and high-handed methods—especially the disbanding of the Agricultural Society and a showdown with the Roman Catholic church—estranged him from the Poles. Tension grew after a massacre of demonstrators near the castle square.

Wielopolski, appointed the head of government in 1862, introduced reforms that were not insignificant but that did not include peasant emancipation. He was viewed as an enemy both by the Reds, who created an underground National Committee, and by the Whites, who also set up a clandestine organization. Wielopolski decided to break the Reds by drafting large numbers of them into the Russian army. In January 1863 the National Committee, left with no choice but to take up the challenge, called on the peoples of Poland, Lithuania, and Rus (Ukraine) to rise, decreed peasant emancipation, and appealed for support from the Jews ("Poles of Mosaic faith").

Thousands responded to the call; however, because the insurgents had failed to establish a stronghold, the National Committee, transformed into the National Government, had to operate underground. In the spring the Whites joined the uprising, contributing finances and international contacts but also seeking control. Fighting extended into Lithuanian and Belarusian lands but not into Ukraine. In some instances the peasantry participated in the struggle, in others they cooperated with the Russians. France hinted that the blood of the insurgents would mark the boundaries of an independent Poland. But in practice France, Britain, and Austria did not go beyond joint diplomatic démarches in St. Petersburg. Prussia sided with Russia. The insurgents, equipped with primitive weapons, fought doggedly as partisans in small detachments and succeeded in keeping the rising going until the autumn of 1864, when its last and most prominent leader, Romuald Traugutt, was captured and executed.

The decades that followed the January Insurrection opened a new phase in the history of partitioned Poland. Harsh reprisals in the kingdom—now called the Vistula Land—were designed to reduce it to a mere province of Russia, denied even the benefits of subsequent reforms in Russia proper. Large garrisons and emergency legislation kept the Poles down. Many involved in the rising were executed or deported to Siberia; thousands of landed estates were confiscated. The Uniate church was abolished, and the Roman Catholic hierarchy was harassed. A huge Orthodox church emerged in the centre of Warsaw. The government believed that it could resolve the Polish question by winning over the peasantry (emancipated in 1864) and pitting them against the *szlachta* and the Catholic church, as well as by eradicating the historical ties between the "western provinces" and Poland. Catholics could no longer buy land there. In Lithuania the brutal governor Mikhail Muravyov was nicknamed "the hangman." The post-1863 period marked the beginning of a final parting of the Poles and the Lithuanians and Ukrainians (the latter also were undergoing a national revival), but in the long run Russian policies did not accomplish their aims.

Russian reprisals

The emancipated peasantry, coming into direct contact with the Russian officialdom and antagonized by anti-Catholic and Russification policies, became more self-consciously Polish. The dispossessed gentry moved to towns, transmitting their values to a growing intelligentsia, which assumed national leadership. As the Industrial Revolution penetrated Congress Poland, the growth of a bourgeoisie and of an industrial proletariat was accelerated. The fastest and greatest development was in textiles and was centred on Łódź—the Polish Manchester—the population of which increased 10-fold between 1865 and 1897. Mining, metallurgy, and food-processing industries followed suit. Vistula Land became the most developed part of the Russian Empire, but its development was uneven and its modernization partial. Moreover, its reliance on the eastern markets made the country dependent on Russia.

Socioeconomic progress contrasted with political stagnation. The Polish question largely disappeared from the European agenda after 1870. Blaming romantic idealism for the catastrophic uprising, people rejected political ac-

tivities and extolled the value of "organic work," progress, and modernization. Warsaw Positivism, deriving its name and inspiration from the thought of Auguste Comte, provided the rationale for these views.

Accommodation with the ruling governments. Uprisings also were condemned by conservatives in Galicia, where the Kraków historical school critically reinterpreted Poland's history. The conservatives were willing to cooperate with Vienna in exchange for concessions, and, as the Habsburg monarchy transformed itself into a dual Austro-Hungarian Empire in 1867, Galicia obtained local autonomy. From the 1860s the province was largely Polonized. Persecuted elsewhere, Polish culture could flourish there; the universities of Kraków and Lwów (Lemberg) and the Academy of Arts and Sciences became cultural beacons radiating across the partition borders. There was less progress in the socioeconomic field. Ruled by conservative landowners, Galicia remained poor and backward. In its eastern part nascent Ukrainian nationalism clashed with that of the Poles.

Kulturkampf

The situation for Poles in Prussia at times appeared critical. Anti-Polish policies pursued by German Chancellor Otto von Bismarck culminated in the Kulturkampf, designed to strengthen the cohesion of the newly created German Empire. The Polish and Roman Catholic character of Poznania and West Prussia was continuously threatened by Germanization. A colonization commission was set up in 1886. Eight years later a society for the promotion of German interests in the east came into being. The Poles called it Hakata, after the initials of its founders. The Polish response took the form of credit unions, cooperative associations, and self-help institutions. Showing great solidarity and organizational talents, working hard, and raising socioeconomic standards, Prussian Poles developed characteristics that distinguished them from their countrymen under Russian or Austrian rule.

In the post-1863 decades, prevailing political attitudes took the form of Triple Loyalism, the belief that material and cultural progress in each part of divided Poland was predicated on loyalty to the ruling governments. This policy seemed to produce beneficial results only under Austria. The pursuit of riches was represented as essentially patriotic even if realized under the harsh conditions of early capitalism. For the rapidly increasing masses, living conditions were deplorable. This led to their radicalization on the one hand and to a sizable emigration on the other. In the period 1870–1914 about 3.6 million people, mostly peasants, emigrated from Polish lands to the United States.

A reaction to that situation developed in the 1890s that had both a nationalist and a socialist character. The National Democratic movement originated with a Polish League organized in Switzerland; by 1893 the organization transformed into the clandestine National League, based in Warsaw. It stressed its all-Polish character, rejected loyalism, and promoted national resistance, even uprisings, when opportune. Its nationalist ideology tinged with populism gradually evolved into "integral" nationalism, which placed national interest and national egoism above everything else. Affected by social Darwinist theories of survival of the fittest and natural selection, Polish nationalism advocated a struggle not only against the partitioning powers but also against the Ukrainians and the Jews, whose interests were seen as opposed to those of the Poles. The father of this integral nationalism was Roman Dmowski, whose writings stressed the need to create a modern Polish nation deriving its strength from the ethnically Polish masses.

Polish socialism, which in its early manifestations was purely a class movement with an emphasis on internationalism, began by the 1890s to stress an indissoluble connection between social revolution and Poland's independence. In 1892, the Polish Socialist Party (PPS) was brought into existence. Illegal under Russian rule, it had a counterpart in Galicia in the Polish Social Democratic Party led by Ignacy Daszyński. The dominant figure in the PPS was Józef Piłsudski, who saw the historic role of socialism in Poland as that of a destroyer of reactionary tsardom. Doubly oppressed (nationally and socially), the Polish proletariat was to be the force to carry the struggle for social justice and national liberation. Opposing such views was the Social Democracy of the Kingdom of Poland and Lithuania, the forerunner of Polish communism. Its leading theorist, Rosa Luxemburg, argued that national independence would not promote the interests of the proletariat, who were integrated economically into the three partitioning states.

Rosa Luxemburg

The outbreak of the Russo-Japanese War in 1904 created a tense atmosphere and sharpened the basic differences between the major political trends. While Dmowski and his supporters sought to extract concessions from the tsarist regime, Piłsudski promoted revolutionary nationalistic tactics. Both politicians went to Tokyo, where they presented their opposite programs to the Japanese. At the beginning of 1905, just as the revolution began to sweep Russia, Congress Poland responded to events with a school strike and a general workers' strike, while Piłsudski's PPS squads battled with Russian troops and police. The government offered limited concessions to the Poles in Congress Poland and the western provinces. Dmowski's larger hopes, bound with the creation of the Russian Duma—in which the Poles were mainly represented by National Democrats—proved unfounded. The PPS, in turn, suffered internal splits, Piłsudski moving increasingly in an insurrectionary (national), as opposed to a revolutionary (social), direction.

In the first decade of the 20th century a mass political culture developed in Polish lands. The Russian Revolution of 1905 contributed to the growth of a civil society in Congress Poland (with legal political parties and trade unions), though it was constantly undermined by Russian rule. In Austria the introduction of universal manhood suffrage in 1907 widened the political involvement of the masses. The Polish peasant movement that had risen in Galicia in the 1890s was beset by schisms. In 1913 there emerged from it the Polish Peasant Party-Piast led by Wincenty Witos. In the German partition, a Polish national revival in Upper Silesia led by Wojciech Korfanty, and one on a lesser scale in East Prussia, affected for the first time regions that had not been part of the pre-partition Commonwealth.

POLAND IN THE 20TH CENTURY

The rebirth of Poland. The outbreak of World War I was accompanied by two major political trends among the Poles. Piłsudski, distancing himself from socialist politics, became a military leader and commander of a brigade that fought on the Austrian side. His cooperation with the Central Powers was tactical, part of his pursuit of the goal of complete independence. Expecting a collapse of the three partitioners, he prepared for a Polish fait accompli. In 1915 the Germans and the Austrians drove out the Russians from Congress Poland, and on Nov. 5, 1916, they issued the Two Emperors' Manifesto proclaiming the creation of the Polish kingdom. Its status and borders remained undefined, but the document internationalized the Polish question. Piłsudski, who refused to raise Polish troops without binding political commitments from the Central Powers, came into conflict with them and in 1917 was imprisoned in Germany.

Dmowski's alternate policy of linking the Polish cause with the Franco-Russian alliance appeared promising when the first formal offer of Polish autonomy and unification came from the Russian commander in chief, Grand Duke Nicholas, on Aug. 14, 1914. Subsequent moves by the Russian government, however, revealed the hollowness of such promises. Russian concessions to the Poles, culminating in the tsar's Christmas Day 1916 order, were made only in reaction to the Central Powers' initiatives and victories. The chances of Polish independence increased radically in 1917 when the United States entered the war and two revolutions shook Russia. U.S. President Woodrow Wilson, to whom the great Polish patriot and pianist Ignacy Paderewski had gained access, already spoke of a united and autonomous Poland in a January 1917 address. The Russian Provisional Government, somewhat ambiguously, and the Petrograd (St. Petersburg) Soviet of Workers' and Soldiers' Deputies more explicitly, declared their recognition of Poland's right to independence in March 1917. At the Brest-Litovsk conference (Dec.

Russian concessions



Changes in Poland's territory from 1031 to 1945.

22, 1917–March 3, 1918) the Bolsheviks denounced the Central Powers' handling of the Polish question. On Jan. 8, 1918, Wilson's Fourteen Points appeared. Point 13 declared that an independent Polish state should be composed of indisputably Polish inhabitants, with a secure access to the sea. The Inter-Allied conference (June 1918) endorsed Polish independence, thus crowning the efforts of Dmowski, who had promoted the Polish cause in the West since 1915. In August 1917 he had set up a Polish National Committee in Paris, which the French viewed as a quasi-government. Under its aegis a Polish army composed mainly of volunteers from the United States was placed under the command of General Józef Haller.

With the end of the war on Nov. 11, 1918, Piłsudski, released by the German revolutionaries, returned to Warsaw. The German-appointed Regency Council handed over its powers to him, and Piłsudski successfully negotiated a German evacuation of the kingdom. A leftist government in Lublin headed by Daszyński resigned in his favour, but not Dmowski's Polish National Committee, representing the Polish political right. The danger of two rival governments was avoided through the mediation of Paderewski. Under a compromise implemented in January 1919, Piłsudski remained chief of state and commander in chief; Paderewski, who became premier and foreign minister, and Dmowski represented Poland at the Paris Peace Conference.

At that stage the Polish government controlled only Congress Poland and western Galicia. In the east the Ukrainians, having proclaimed their own republic, battled the Poles. Farther east the Poles clashed with the Bolsheviks who were advancing into Belarusian and Lithuanian lands. A Polish uprising in Poznania led to a partial seizure of the province, but the fate of Prussian Poland lay in the hands of the peacemakers.

The struggle between Piłsudski and Dmowski

From the Treaty of Versailles to the Treaty of Riga. The Polish program at the Paris Peace Conference was affected by the Piłsudski-Dmowski dualism. Piłsudski's approach was "federalist," Dmowski's "incorporationist." The former strove to establish a bloc of states corresponding to pre-partition Poland. The latter postulated a centralized Polish state, with its eastern border determined by the Second Partition but also including Upper Silesia and parts of East Prussia transferred from Germany. France favoured strengthening Poland at Germany's expense, but Britain opposed that approach. Wilson occupied a middle position. The borders drawn under the Treaty of Versailles (June 1919) roughly corresponded to Polish-German frontiers before the partitions, except that Gdańsk became the Free City of Danzig, and plebiscites were held in parts of East Prussia and Upper Silesia to determine which nation these regions wished to join. The East Prussian plebiscite of July 1920 (at the height of the Russo-Polish War) was won by Germany. In the Silesian plebiscite of March 1921—preceded and followed by three Polish uprisings—682 communes voted for Poland and 792 for Germany. The region was formally divided in October 1921.

The drawing of the southern border under the Treaty of Saint-Germain (September 1919) was preceded by an armed Czech-Polish clash in January 1919 in the Duchy of Cieszyn. In July 1920 the area was divided, leaving a sizable Polish minority in Czechoslovakia. As for the embattled eastern Galicia, the Allies authorized a Polish administration and military occupation in 1919. Final recognition of Polish sovereignty came only in 1923, the delay being due to the Russian situation.

An armed struggle between the Bolsheviks and Poland resulted from Russian attempts to carry the revolution westward and from Piłsudski's federalist policy. The Great Powers failed to pursue either an all-out intervention against the Bolsheviks or a policy of peace. An Allied proposal for a temporary border between Bolshevik Russia and Poland (called the Curzon Line) was unacceptable to either side. Except for an alliance with the Ukrainian leader Symon Petlyura (April 1920), whose troops accompanied the Poles as they captured Kiev in May, Poland fought in isolation. An offensive by the Red Army drove the Poles back to the outskirts of Warsaw, but the country

was saved from catastrophe by Piłsudski's counterattack on August 16 (the "Miracle of the Vistula"). In the compromise Peace of Riga (March 1921), the Bolsheviks abandoned their plans to communize Poland, but the Poles had to abandon their federalist concepts. The new border, which corresponded roughly to the 1793 frontier, cut across mixed Ukrainian and Belarusian territories. In the north it included Wilno, captured by General Lucjan Żeligowski, a move that dug a chasm between Lithuania and Poland.

The Second Republic. Interwar Poland with an area of about 150,000 square miles and more than 27 million inhabitants (more than 35 million by 1939) was the sixth largest country in Europe. Devastated by the years of hostilities, the state had to be reconstructed of three parts with different political, economic, and judicial systems and traditions. More than three-fifths of the population was dependent on agriculture that was badly in need of structural change: agrarian reform and redistribution of land that would relieve the demographic pressure (*e.g.,* hidden unemployment) and modernization of production that could alleviate the disparity between agrarian and industrial prices ("the price scissors"). Industrialization was essential, but local capital was insufficient, and foreign investors did not always operate in Poland's interests. Nonetheless, the Polish economy made important strides in the mid-1920s through the reforms of Władysław Grabski. After being severely hit by the Great Depression of the 1930s, it began to recover under the guidance of Eugeniusz Kwiatkowski, whose earlier achievements included the building of a new port and town of Gdynia.

Dependence on agriculture

Economic difficulties were interwoven with and augmented by political problems, among which the minority issues ranked high. Ukrainians (some 16 percent of the total population, according to recent estimates), Jews (about 10 percent), Belarusians (about 6 percent), and Germans (about 3 percent) lived in a state that, although multiethnic, was based on single-nation ideology. The Ukrainians never fully accepted Polish rule, and Ukrainian extremists engaged in terrorism to which the Poles responded with brutal "pacifications." In the case of the large and unassimilated Jewish population, concentrated in certain areas and professions, anti-Semitism was rampant, especially in the 1930s, though Poland never introduced anti-Jewish legislation.

Interwar politics centred to a large extent on the search for a constitutional model that would reconcile traditional Polish strivings for liberty with the need for a strong government. Piłsudski gave up his provisional powers to a Sejm elected in January 1919 but continued as the head of state under a provisional "Little Constitution." The Sejm quickly became an arena of interparty strife, with the right grouped around the National Democrats, the left around the PPS and radical Populists, and the centre represented mainly by the Polish Peasant Party-Piast. The illegal Communist Party, formed in 1918, was of marginal importance. The constitution of 1921 made the parliament supreme vis-à-vis the executive. The proportional system of universal suffrage (which included women) necessitated coalition cabinets, and the left and the right almost never cooperated. The assassination of the first president of the republic, Gabriel Narutowicz, in 1922 by a nationalist fanatic showed the extent of blind partisanship.

In May 1926 Piłsudski (who had held the title of marshal since 1920) came out of his three-year retirement. Demanding moral and political cleansing (*sanacja*), he staged an armed demonstration intended to force President Stanisław Wojciechowski to dismiss the government. Fighting in Warsaw ensued and ended in victory for Piłsudski. His candidate, Ignacy Mościcki, became president and remained in office until World War II. Piłsudski rejected fascism and totalitarianism but promoted an authoritarian regime in which his former legionnaires played a key role. Worshiped by his supporters and hated by his opponents, he became a father figure for large segments of the population. The pro-Piłsudski Non-Party Bloc of Cooperation with the Government (BBWR) became his political instrument, used at first against the opposition rightist National Democrats. In 1930 Piłsudski responded to the challenge of the centre-left opposition (Centrolew) by ordering the arrest and trial of its leaders, including three-time premier Witos. The brutal Brześć affair (named for the fortress in which the politicians involved were imprisoned) was seen as a blot on the Piłsudski regime, even though the sentences were light and some of the accused emigrated.

The Brześć affair

Following the 1930 elections BBWR had a majority in the Sejm. In April 1935 it was able to push through a new constitution, which placed the president above all other branches of government. An electoral law undercut the political parties that boycotted the 1935 parliamentary elections. In May Piłsudski died, leaving the country as a dictatorship without a dictator. His legend could not be bequeathed. A decomposition of the *sanacja* regime (sometimes called that of the colonels) ensued. Attempts to pass on Piłsudski's mantle to the new commander in chief, Marshal Edward Śmigły-Rydz, were unsuccessful, as was the artificial creation of a governmental party—the Camp of National Unity. The peasant parties (now united); the increasingly chauvinist National Party (as the National Democrats were by then known), with its fascist splinter party, the National Radical Camp; and the socialists all opposed the regime and showed success in municipal elections. Socioeconomic tension was translated into peasant strikes in the countryside and riots in towns.

Political and socioeconomic difficulties contrasted with the richness of intellectual, artistic, and scholarly life of the period. Twenty years of independence had given the Poles a new confidence that proved essential in the trials of World War II. Poland's international position between an inimical and revisionist Germany (which constantly denounced the "corridor" separating it from East Prussia) and the Soviet Union was dangerous from the start. The tasks of Polish diplomacy in the interwar period were exceedingly difficult. The only option was to remain neutral in regard to its two giant neighbours while concluding alliances (in 1921) with France and Romania. An alliance with Czechoslovakia foundered on basic differences of approach to international relations, particularly when Colonel Józef Beck became Piłsudski's foreign minister in 1932.

In 1932 Poland succeeded in signing a nonaggression pact with Soviet Russia and in 1934 a declaration of nonaggression with Nazi Germany. The enmity of the Nazis for the Soviets seemed to preclude a rapprochement (such as the Russo-German agreement at Rapallo, Italy, in 1922). The French alliance was maintained, though its value to Poland was diminished by the treaties of Locarno (1925) and subsequent Franco-German cooperation. Warsaw vainly sought to encourage Paris to adopt a strong line against Nazi Germany. But the French did not react forcibly even to the German remilitarization of the Rhineland (1936). Poland continued its policy of balance, but, in profiting from the German action against Czechoslovakia by gaining the disputed part of Cieszyn (October 1938), it gave the impression of being in collusion with Adolf Hitler. However, when confronted with German demands for an extraterritorial road through the "corridor" and the annexation of Danzig, as well as with an invitation to join the Anti-Comintern Pact, Beck knew that his country's independence was at stake. Accepting British Prime Minister Neville Chamberlain's guarantee of March 1939 and turning it into a full-fledged alliance with Britain, Warsaw rejected German demands. On Sept. 1, 1939, Hitler, having secured Soviet cooperation through the German-Soviet (Molotov-Ribbentrop) Nonaggression Pact a week earlier, launched an all-out attack against Poland.

The German-Soviet pact

World War II. The Poles, fighting alone against the Wehrmacht's overwhelming might, particularly in air power and armour, were doomed. On September 17 the Red Army invaded Poland, and on September 28 Hitler and Stalin agreed on a final partition, the Soviets taking eastern Galicia and lands east of the Bug River (*i.e.,* more than half of the country, where the Poles constituted about two-fifths of the population). After farcical plebiscites these territories were incorporated into Soviet Ukraine and Belorussia. Between 1940 and 1941 about 1.5 million people

were deported to the U.S.S.R. Wilno was handed over to Lithuania, which by 1940 had become one of the Soviet republics. While the Soviets singled out class enemies, the Germans—who split the area they occupied into a central region called the General Government and territories annexed to the Reich—emphasized race. The Holocaust claimed the lives of some three million Polish Jews, herded into ghettoes and killed in extermination camps, of which Auschwitz (Oświęcim) was but one. Thousands of Jews died fighting, as in the Warsaw Ghetto Uprising in 1943. The Nazis also engaged in mass terror, deporting and executing Poles in an attempt to destroy the intelligentsia and reduce Polish culture to the lowest level.

A large Polish underground resisted the Nazis from 1939 through a veritable secret state and a Home Army (AK) loyal to the Polish government in exile. The latter was a legal successor of the government that on Sept. 17, 1939, had crossed into Romania and was interned there. Set up in Paris and moved after the collapse of France to London, it was led by the premier and commander in chief General Władysław Sikorski. Under his command Polish troops, organized in the west, fought in Europe and North Africa. Polish pilots played a disproportionately large role in the Battle of Britain (1940–41), and the small Polish navy distinguished itself. A major Polish contribution to the war effort lay in discovering and passing on to the Allies the secret of the German ciphering machine Enigma.

The German attack on the Soviet Union in June 1941 changed drastically Poland's position, for one of its foes now became a member of the Grand Alliance. Under British pressure the Polish government in exile reestablished relations with the Soviet Union through the Sikorski-Maysky accord, accepting the annulment of the Ribbentrop-Molotov treaty without an explicit Soviet renunciation of annexed Polish territory. The Soviets promised to release the deported Poles and agreed to the creation of a Polish army—more than 230,000 Poles had been prisoners of war since 1939—under the command of General Władysław Anders. Difficulties appeared almost from the start, however. The Soviets sought British and U.S. approval for their territorial gains. Friction developed regarding the Polish army in Russia, which in 1942 was evacuated to the Middle East. Meanwhile, the Soviets were promoting Polish communist activity both in the U.S.S.R. and in occupied Poland, where a Polish Workers' Party (PPR) emerged in 1942 with its own small People's Guard.

The emergence of the Polish Workers' Party

British Prime Minister Winston Churchill, not appreciating fully Stalin's hegemonic designs, believed that timely territorial concessions to the U.S.S.R. would preserve the internal independence of postwar Poland. During three visits to Washington, D.C. (1941–42), Sikorski outlined his ideas about postwar security in east-central Europe, including a Czechoslovak-Polish confederation; however, U.S. President Franklin D. Roosevelt regarded Polish issues as secondary. For him, as for Churchill, the importance of the Soviet Union as an ally was crucial, and neither leader was prepared to see relations with Stalin founder on the Polish rock. This became apparent when they were undeterred by the German announcement on April 13, 1943, of the discovery in the Katyn Forest of mass graves of more than 4,000 Polish officers who had been captured by the Red Army. Only in 1992 did postcommunist Moscow publicly acknowledge its guilt in this matter.

Sikorski's death in a mysterious plane crash in Gibraltar (July 1943) was a great blow to the Poles at a time when Soviet offensives after the victories of Stalingrad and Kursk brought the Red Army closer to the prewar Polish borders. The new prime minister and Peasant Party leader, Stanisław Mikołajczyk, could not rival Sikorski's standing and was at odds with the new commander in chief, General Kazimierz Sosnkowski. The Soviets demanded as the price for reestablishing relations with the Polish government not only territorial concessions but also the dismissal of several of its members. Their backing of Polish communists—the Union of Polish Patriots in Moscow and the National Committee of the Homeland, headed by Bolesław Bierut and set up in Poland in December 1943—was evident. At the Tehrān Conference late in 1943, Churchill's proposal that the Soviet-Polish border coincide with the Curzon Line (roughly similar to the Ribbentrop-Molotov line) and that Poland be compensated at Germany's cost was accepted by Roosevelt and Stalin. The Mikołajczyk government, which was opposed to such a territorial deal, was not informed.

Roosevelt suggested to Mikołajczyk that the AK show its goodwill by cooperating with the Red Army. Such cooperation, however, when attempted in areas that had been part of prewar eastern Poland, was followed by arrests and deportation or conscription into the Soviet-sponsored Polish Kościuszko Division commanded by General Zygmunt Berling. On Aug. 1, 1944, just as Mikołajczyk, prompted by the British, came to Moscow, the AK, under the supreme command of General Tadeusz Bór-Komorowski, rose in Warsaw against the retreating Germans. The Warsaw Uprising constitutes one of the most tragic and controversial events of the war. The AK planned to capture the capital and act on behalf of Mikołajczyk's government as host to the entering Red Army. It was assumed that the Soviets would not dare to disregard this demonstration of the Polish right to self-determination. In the absence of Soviet military assistance, the rising was doomed, yet had the AK not risen it would have been accused of inactivity by the communists. The insurgents fought alone for 63 days, because the Soviets not only halted their own offensive but also refused to allow Allied planes to help resupply the AK. When Warsaw capitulated, the city had been almost totally destroyed, and 200,000 civilians and more than 10,000 combatants had perished.

The Warsaw Uprising

Stalin had no interest in assisting the Polish underground and did not hesitate to defy world public opinion when, in March 1945, he had 16 leaders of the underground arrested and tried in Moscow. Their elimination was linked to the process of building a communist-dominated Polish state. In July 1944 a Polish Committee of National Liberation was set up in Moscow, issued its Lublin Manifesto (July 22), and signed a secret territorial accord with the U.S.S.R. Mikołajczyk, caught between British pressure and the resistance of his government, resigned in November 1944. Ignoring the socialist Tomasz Arciszewski, who succeeded Mikołajczyk as premier, Roosevelt and Churchill agreed with Stalin at the Yalta Conference (February 1945) to create a Provisional Polish Government of National Unity. Its core was the Lublin Polish Committee of National Liberation (already recognized by Stalin as the government), to which some politicians from Poland and abroad were added. Britain and the United States recognized that government on July 5, 1945, simultaneously withdrawing recognition from the government in London. A large Polish political emigration emerged as a voice of a free Poland and remained active during the next 40 years.

Communist Poland. The postwar Polish republic, renamed in 1952 the Polish People's Republic, occupied an area some 20 percent smaller than prewar Poland, and its population of almost 30 million rose to nearly 39 million in the following four decades. After the Holocaust, the expulsion of several million Germans, and population transfers with the U.S.S.R., Poland was virtually ethnically homogenous. The expulsion of the Germans was approved by the Potsdam Conference, but the final decision regarding the new German-Polish border along the Oder-Neisse Line was left to a future peace conference. The U.S.S.R. cleverly capitalized on its status as the sole guarantor of this border, which gave Poland a long seacoast, with such harbours as Szczecin and Gdańsk, and such natural resources as coal and zinc in Silesia. But, if Poland became potentially wealthier, it nevertheless was devastated (cities like Warsaw, Wrocław, and Gdańsk lay in ruins), and social conditions bordered on chaos. Huge migrations, mainly to the ex-German "western territories," added to the instability. Fighting against the remnants of the Ukrainian Liberation Army was followed by the mass relocation of the Ukrainians (Operation Vistula) in 1947. Persecutions of the AK and political opponents (the National Party was outlawed) by the communists led to armed clashes that continued for several years. It was under these conditions that a Jewish pogrom occurred in Kielce in June 1946, claiming more than 40 lives.

Operation Vistula

Bierut, who was formally nonpartisan but in fact was

an old communist, became president. In a cabinet headed by a socialist and dominated by communists and fellow travelers, Mikołajczyk became deputy prime minister. He successfully re-created a genuine Polish Peasant Party (PSL), which was larger than the PPR and its socialist and democratic satellite parties (the PPS and SD, respectively). Supported by all enemies of communism, Mikołajczyk sought to challenge the PPR in the "free and unfettered" elections stipulated by the Yalta accords. His opponents included the ruthless secretary-general of the PPR, Władysław Gomułka, a "home communist," and the men in charge of security (Jakub Berman) and of the economy (Hilary Minc), who had returned from Russia.

The Sovietization of Poland, accompanied by terror, included the nationalization of industry and the expropriation of privately owned land parcels larger than 125 acres (50 hectares). Yet in some areas (namely, matters concerning the church and foreign policy) the communists trod lightly during this transition period. The test of strength between Mikołajczyk and the PPR first occurred during the referendum of 1946—the results of which, favourable to Mikołajczyk, were falsified—and then in the general elections of 1947, which were hardly "free and unfettered." Mikołajczyk, fearing for his life, fled the country. The victorious communists completed their monopoly of power in 1948 by absorbing the increasingly dependent PPS to become the Polish United Workers' Party (PUWP).

Over the next few years the Bierut regime closely followed the Stalinist model in politics (adopting the Soviet-style 1952 constitution), economics (emphasizing heavy industry and collectivization of agriculture), military affairs (appointing the Soviet Marshal Konstantin Rokossovsky as commander of Polish forces and adhering to the Warsaw Pact of 1955), foreign policy (joining the Communist Information Bureau, the agency of international communism), culture, and the rule of the secret police. Political terror in Poland, however, did not include, as elsewhere, show trials of fallen party leaders—Gomułka, denounced as a "Titoist" and imprisoned in 1951, was spared such a trial. Moreover, the primate of Poland, Stefan Wyszyński, could still negotiate a modus vivendi in 1950, though he was arrested in September 1953.

The death of Stalin in March 1953 opened a period of struggle for succession and change in the U.S.S.R. that had repercussions throughout the Soviet bloc. The interlude of liberalization that followed culminated in the Soviet leader Nikita Khrushchev's denunciation of Stalinism at the 20th Party Congress in February 1956. With the sudden death of Bierut, anti-Stalinists in Poland raised their heads; a violently suppressed workers' strike in Poznań in June 1956 shook the whole country. Gomułka, who believed in a "Polish road to socialism," became a candidate for the leadership of the party. What appeared as his confrontation with Khrushchev, who came to Warsaw in October and threatened intervention, made Gomułka popular throughout Poland. In reality the Polish leader convinced Khrushchev of his devotion to communism and of the need for a reformist approach to strengthen its doctrine. Important changes followed, among them Polish-Soviet accords on trade and military cooperation (Rokossovsky and most Soviet officers left the country), a significant reduction of political terror, an end to forced collectivization, the release of Cardinal Wyszyński (followed by some concessions in the religious field), and contacts with the West, including freer travel. Gomułka's objective, however, was to bridge the gap between the people and the party, thereby legitimizing the latter. Hence, the period of reform known as "Polish October" did not prove to be the beginning of an evolution of communism as revisionists at home and politically motivated émigrés had hoped.

The Gomułka regime

Within a decade economic reform slowed down, the activity of the church was circumscribed, and intellectuals were subjected to pressures. Demonstrations by students in favour of intellectual freedom led to reprisals in March 1968 that brought to an end the so-called "little stabilization" that Gomułka succeeded in achieving. Ever more autocratic, Gomułka became involved in an "anti-Zionist" campaign that resulted in purges within the party, administration, and army. Thousands of people of Jewish origin emigrated. Also in 1968 Polish troops joined the Soviet-led intervention in Czechoslovakia. In 1970 Gomułka registered a foreign policy success by signing a treaty with West Germany that involved a recognition of the Oder-Neisse border. In December 1970, however, major strikes in the shipyards at Gdańsk, Gdynia, and Szczecin, provoked by price increases, led to bloody clashes with police and troops in which many were killed. Gomułka was replaced as first secretary by the more pragmatic Edward Gierek.

The Gierek decade (1970–80) began with ambitious attempts to modernize the economy and raise living standards. Exploiting the East-West détente, he attracted large foreign loans and investments. But initial successes turned sour as the world oil crisis and mismanagement of the economy produced huge budget deficits, which Gierek tried to cover through increased borrowing. The policy of consumerism failed to strengthen the system, and new price increases in 1976 led to workers' riots in Ursus and Radom, which once again were brutally suppressed.

A Workers' Defense Committee (KOR) arose and sought to unite the intelligentsia, which had been isolated in 1968, and the workers, who had received no support in 1970. Dissidents such as Jacek Kuroń and Adam Michnik became internationally known. Other committees appeared that claimed the legality of their activity and protested reprisals as contrary to the 1975 Helsinki Accords. The PUWP responded with measures of selective intimidation. In 1978 the election of Karol Cardinal Wojtyła, the archbishop of Kraków, as Pope John Paul II gave the Poles a father figure and a new inspiration. The coalition of workers and intellectuals, operating largely under the protective umbrella of the church, was in fact building a civil society. The pope's visit to Poland in 1979 endowed that society with national, patriotic, and ethical dimensions. The barrier of fear was broken as another strike was staged at the Gdańsk shipyard that forced an accord with the government on Aug. 31, 1980. Out of the strike led by a charismatic electrician, Lech Wałęsa, there emerged the almost 10-million-strong Independent Self-Governing Trade Union Solidarity (Solidarność), which the government was forced to recognize. Here was an unprecedented working-class revolution directed against a "socialist" state, an example to other peoples of the Soviet bloc.

The pope's visit

A huge movement that sought not to govern but rather to ensure freedom through a "self-limiting revolution," Solidarity could not have been homogenous. The opponents of communism ranged from those who opposed the system as contrary to liberty and democracy, to those who saw it as inimical to national and Christian values, to those who felt that it had not lived up to its socioeconomic promises. These three attitudes all resurfaced after the fall of communism and explain a good deal about the developments in the 1989–96 period.

Gierek did not survive politically the birth of Solidarity, and he was replaced by Stanisław Kania, who was followed by General Wojciech Jaruzelski. By the autumn of 1981 Jaruzelski had become premier, first secretary of the party, and commander in chief. His decision to break Solidarity through the introduction of martial law in December 1981 may well have stemmed from a conviction that the constant tug of war between Solidarity and the government was leading the country toward anarchy that had to be ended by Polish or by Soviet hands. It is likely that he could not conceive of any Poland except a communist one. Martial law effectively broke Solidarity by paralyzing the country and imprisoning virtually all of the movement's leadership, Wałęsa included. It did not, however, destroy it. After the lifting of martial law in 1983, the government could not establish its legitimacy. A political deadlock was worsened by a catastrophic economic situation. In 1984 a popular priest, Jerzy Popieluszko, was murdered by the secret police, but for the first time in such a case state agents were arrested and charged with the crime.

The advent of Mikhail Gorbachev as the leader of the Soviet Union in 1985 and his policies of reform (glasnost and perestroika) started a process that led to the collapse of communism in eastern Europe and the disintegration of the U.S.S.R. The Jaruzelski regime realized that broad reforms were unavoidable and that a revived Solidarity

had to be part of them. The roundtable negotiations under the auspices of the church—Józef Cardinal Glemp succeeded Wyszyński as primate—resulted in a "negotiated revolution." Solidarity was restored and participated in partly free elections in June 1989 that brought it a sweeping victory.

Poland since 1989. Detaching the satellite (populist and democratic) parties from the PUWP, Wałęsa negotiated a compromise by virtue of which Jaruzelski was elected president, and Wałęsa's adviser, Tadeusz Mazowiecki, became premier. The first government led by a noncommunist since World War II, it adopted a "shock therapy" program of economic reform (the Balcerowicz Plan) in 1990 that was meant to arrest Poland's financial and structural crisis and rapidly convert the communist economic model into a free-market system. Although successful, it came with enormous social costs. The new government achieved two major successes: a formal recognition of the Oder-Neisse border by the reunited Germany and, after the dissolution of the Warsaw Pact in 1991, the evacuation of Soviet troops from the country in 1992.

"Shock therapy"

Poland gradually reoriented itself toward western Europe. By 1996 the country had become a member of the Council of Europe, had established economic ties with the European Union, and was admitted to the Organisation for Economic Co-operation and Development. In 1999 Poland (along with the Czech Republic and Hungary) joined the North Atlantic Treaty Organization. Russia's unsettled political situation during the 1990s cast a shadow on Polish foreign policy and complicated its options. Nevertheless, Poland signed accords with Ukraine and Lithuania, and limited regional cooperation was established with the formation of the Visegrad Group, whose members were Poland, the Czech Republic, Slovakia, and Hungary.

By the mid-1990s, the Polish economy was making important strides, including significant reductions in the annual inflation rate and the budget deficit. Moreover, Poland's annual growth rate was the highest in Europe. But progress was geographically uneven, and some economic sectors (*e.g.*, the coal-mining and the building industries) experienced slumps. The gap between the rich and the poor grew, contributing to bitterness and frustration.

Solidarity's disintegration became apparent in the 1990 election, in which Wałęsa defeated Mazowiecki for the presidency. Voters expressed their dissatisfaction by supporting the dark-horse candidate Stanisław Tyminski, a Polish émigré businessman from Canada, who finished second in the balloting. The centrist Freedom Union (UW), which bore the brunt of the transition to democracy, failed to communicate its vision to the public and remained largely a party of the intelligentsia. The rightists, split into several groups, accused Wałęsa and other roundtable negotiators of selling out to the communists. The communists, who profited from the collapse of the economy, reorganized as the Social Democracy of the Republic of Poland (SdRP). The party exploited the increased frustration over the inequalities of a capitalist economy and the political infighting of the Solidarity camp. The SdRP organized a broader coalition, the Democratic Left Alliance, which included a postcommunist version of the All Poland Trade Unions Alliance and other smaller groupings. Well-organized and disciplined, they won the legislative election in alliance with the PSL in 1993 and the presidency in 1995. Wałęsa's narrow defeat by the young, dynamic "post-communist" Aleksander Kwasniewski—whose campaign asked voters to look to the future rather than the past—may have been symptomatic of a generational change that was visible also in the attitude toward the church. The high prestige of the church suffered as its efforts to influence politics and to be a national rallying point in the increasingly secularized postcommunist society occasionally backfired. This was particularly evident on the issue of abortion.

The election of Kwasniewski

The new government continued to introduce, albeit more cautiously, pro-market reforms after 1993. It failed, however, to reform the obsolete welfare state that had been inherited from communism.

A new constitution was adopted in 1997, despite opposition by the extra-parliamentary right, since 1996 united in a loose coalition known as the Electoral Action Solidarity (AWS). In September the AWS defeated its postcommunist foes and formed a ruling coalition. The new government of Prime Minister Jerzy Buzek included, among others, the leader of the UW and the architect of the shock therapy reforms, Leszek Balcerowicz, as the deputy prime minister and minister of finance. The government continued economic reforms and introduced a three-tier system of administration and local self-government. The health-care, pension, and education systems also underwent reform beginning in 1999. The government pursued entry into the European Union, and by the end of the 1990s, Poland ranked as one of the top candidates for membership. (P.S.Wa/K.Ja.)

For later developments in the history of Poland, see the BRITANNICA BOOK OF THE YEAR.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 921, 923, 962, 963, 971, and 972.

BIBLIOGRAPHY

Physical and human geography. *Poland: A Handbook* (1977; originally published in Polish, 2nd ed., 1977), is a comprehensive reference source written by Polish authors and published in Poland for readership outside the country. GLENN E. CURTIS (ed.), *Poland: A Country Study*, 3rd ed. (1994), provides a balanced treatment. R.H. OSBORNE, *East-Central Europe* (1967), on geography, includes a chapter on Poland. GRZEGORZ WECŁAWOWICZ, *Contemporary Poland: Space and Society* (1996), discusses the changes since 1989. ZBIGNIEW LANDAU and JERZY TOMASZEWSKI, *The Polish Economy in the Twentieth Century*, trans. from Polish (1985), offers an uncritical treatment. Aspects of cultural life are dealt with in BOLESŁAW KLIMASZEWSKI (ed.), *An Outline History of Polish Culture*, trans. from Polish (1983), covering the main cultural trends from medieval times to 1982; and STANISŁAW LORENTZ, *Guide to Museums and Collections in Poland* (1974; originally published in Polish, 1971). *The Polish Review* (quarterly) focuses on current cultural events. (J.A.K./A.H.D./N.D.)

History. The history of Poland is presented in W.F. REDDAWAY *et al.* (eds.), *The Cambridge History of Poland*, 2 vol. (1941–50, reissued 1971); ALEKSANDER GIEYSZTOR *et al.*, *History of Poland*, 2nd ed., trans. from Polish (1979); NORMAN DAVIES, *God's Playground: A History of Poland*, 2 vol. (also published as *A History of Poland, God's Playground*, 1981); ADAM ZAMOYSKI, *The Polish Way: A Thousand-Year History of the Poles and Their Culture* (1987, reissued 1994); PAUL W. KNOLL, *The Rise of the Polish Monarchy: Piast Poland in East Central Europe, 1320–1370* (1972); J.K. FEDOROWICZ, MARIA BOGUCKA, and HENRYK SAMSONOWICZ (eds.), *A Republic of Nobles: Studies in Polish History to 1864*, trans. from Polish (1982); HARRY E. DEMBKOWSKI, *The Union of Lublin: Polish Federalism in the Golden Age* (1982); PIOTR S. WANDYCZ, *The Lands of Partitioned Poland, 1795–1918* (1974, reprinted 1984), and *Polish Diplomacy, 1914–1945: Aims and Achievements*, ed. by KEITH SWORD (1988); R.F. LESLIE (ed.), *The History of Poland Since 1863* (1980); ANDRZEJ WALICKI, *Philosophy and Romantic Nationalism: The Case of Poland* (1982, reissued 1994); ANTONY POLONSKY, *Politics in Independent Poland, 1921–1939: The Crisis of Constitutional Government* (1972); JAN KARSKI, *The Great Powers & Poland, 1919–1945: From Versailles to Yalta* (1985); JÓZEF GARLIŃSKI, *Poland in the Second World War* (1985); and JACEK JEDRUCH, *Constitutions, Elections, and Legislatures of Poland, 1493–1977: A Guide to Their History* (1982). WŁADYSŁAW CZAPLIŃSKI and TADEUSZ ŁADGÓRSKI (eds.), *The Historical Atlas of Poland*, 2nd ed. (1986; originally published in Polish, 4th ed., 1977); and GEORGE J. LERSKI, *Historical Dictionary of Poland, 966–1945*, ed. by PIOTR WRÓBEL and RICHARD J. KOZICKI (1996), are useful companions to historical readings.

Communism and its collapse and aftermath are examined in these works: JAKUB KARPINSKI, *Countdown: The Polish Upheavals of 1956, 1968, 1970, 1976, 1980*, trans. from Polish (1982); JAN JÓZEF LIPSKI, *KOR: A History of the Workers' Defense Committee in Poland, 1976–1981* (1985; originally published in Polish, 1983); MICHAEL CHECINSKI, *Poland: Communism, Nationalism, Anti-Semitism*, trans. from Polish (1982); JADWIGA STANISZKIS, *Poland's Self-Limiting Revolution*, ed. by JAN T. GROSS (1984); LAWRENCE WESCHLER, *The Passion of Poland, from Solidarity Through the State of War* (1984); TIMOTHY GARTON ASH, *The Polish Revolution: Solidarity, 1980–82*, rev. and updated ed. (1991); NEAL ASCHERSON, *The Polish August: The Self-Limiting Revolution* (1981); GEORGE BLAZYCA and RYSZARD RAPACKI (eds.), *Poland into the 1990s: Economy and Society in Transition* (1991); JANE LEFTWICH CURRY and LUBA FAJFER (eds.), *Poland's Permanent Revolution: People vs. Elites, 1956 to the Present* (1996); GRZEGORZ EKIERT, *The State Against Society: Political Crises and Their Aftermath in East Central Europe* (1996); and RAYMOND TARAS, *Consolidating Democracy in Poland* (1995). (Ha.Ro./K.M.Ś./P.S.Wa./K.Ja.)

Police

The term police is now used primarily to denote a body of people organized to maintain civil order and public safety, to enforce the law, and to investigate breaches of the law.

All societies need some way to maintain order. In the smallest societies, informal sanctions discourage deviation. In the simplest forms of state organization, informal sanctions are supplemented by agents of the ruler who enforce his decisions. Although the police function is universal in society, it is only in the larger and more complex states that full-time officials are appointed with special police responsibilities regulated by politics, tradition, and law. Most police in such societies are enrolled in state, provincial, or municipal forces concerned with criminal investigation, traffic regulation, and preventive patrol. A central government generally has personnel responsible for the collection of political information, counterespionage, and so forth, not all of which is considered policing.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 543 and 737, and the *Index*.

This article is divided into the following sections:

Police and society 957
- The diversity of police activities 957
- Centralization versus decentralization 958

The history of policing 958
- Anglo-Saxon beginnings 958
- The French police system 959
- The formation of the English police 960
- Early police in the United States 960
- Detective policing in England and the United States 961
- Late 19th-century English and American policing 961
- Developments in Australia and Canada 961
- 20th-century policing 962
- Centralized police organizations 963
- Decentralized police organizations 964
- Secret police 965

Police work and law enforcement 965
- Routine police activities 965
- Police and courts 965
- Crowds and riots 965
- Police work as an occupation 966
- International organization 966

Police technology 966
- Modern surveillance systems 967
  - Patrol systems
  - Police dogs
  - Electrical protection systems
  - Technical surveillance
- Crime reporting and records 968
  - Fingerprint classification
  - Modus operandi
  - Investigation and reference files
  - Computerized records
- Communications 969
- Interrogation methods and equipment 969
- Weapons and related devices 969
  - Standard weapons
  - Chemical and other nonlethal weapons
  - Restraining and protective devices
- Laboratory procedures 970
  - Facilities and equipment
  - Evidence and its examination

Bibliography 972

Police and society

THE DIVERSITY OF POLICE ACTIVITIES

There is a remarkable historical, geographical, and organizational diversity in the activities of persons who are, or have been, counted as police. Within any one country the work of police today is very different from what it was 200 years ago. There are also major differences between countries—policing New York City is bound to have little in common with policing the Solomon Islands. The diversity is so great that the onlooker may wonder if the different kinds of activity and organization have sufficient in common to be classified together.

Police accountability and adaptation

The principle of police accountability is helpful in explaining the diversity. The problem is summarized in a Latin question: *Quis custodiet ipsos custodes?* ("Who guards the guardians?"). From one point of view, it is desirable that a police force be as efficient as possible. This consideration favours the establishment of national police forces, which can take advantage of the economies of scale in training, promotion, organization, and so on. But it often has been felt that the existence of a national force places too much power in the hands of those who direct it; that there is a danger that the government will use its control of the police to keep itself in office; and that the police will not be accountable to the public. Therefore, some countries favour a local basis for police organization.

A second principle that helps explain the diversity of police activities in different societies is that of police adaptation to cultural traditions. If the society is agricultural, as in rural India, the police will be concerned with the state of the crops, the irrigation of fields, the conditions of roads and paths, private feuds and quarrels, the registration of births and deaths, feasts, fairs, and all manner of private and public events. If local landowners are influential, as they were in 19th-century England, much police effort will be expended in the enforcement of laws against poaching and much less in the investigation of ways in which the behaviour of employers infringes the rights of employees. If some offenses are committed by criminals using vehicles, the police must be partially motorized. If criminals use firearms, the police are likely to be armed. Police activity must be adapted to the society that is being policed.

Social control in small communities. Most people obey most laws willingly, whether or not there is a police officer looking on. They comply with the laws because they believe them right and perceive that in the long run it is in everybody's interest to observe them. In small communities in which everyone knows everyone else, people who live up to the shared ideals are rewarded by the esteem of their fellows. If they fall short of other people's expectations, life is made more difficult because others are no longer so ready to help them in times of trouble. This pattern of informal rewards and punishments is, in all kinds of society, the most potent force in law enforcement, but it is stronger in the small community. The forces that order the life of the small community make the task of the police much easier. Instead of imposing requirements, they have only to lubricate the mechanisms inherent in social relations, and police action is needed only when the informal controls have proved insufficient.

Social control in large societies. The bigger a society becomes, the weaker its self-policing elements are likely to be and the stronger the legal means of social control becomes. In the first place, when people are involved with those whom they are unlikely ever to meet again, there may seem to be less reward for honest dealing. In the second place, with technological progress there comes a

need for new laws. With the growth of automobile traffic, for example, many new laws governing vehicle construction, licensing, and manner of driving become necessary. Because many such laws do not have the same moral significance as laws penalizing personal violence or dishonesty, the reporting of offenses cannot be left to members of the public. Controls are needed in the form of trained inspectors, officials, and traffic police. When the legislature declares that certain actions are in the future to count as offenses, it increases the amount of crime. There is a danger that people who are convicted of one offense—say, a motoring offense—may feel aggrieved and in the future be less willing to cooperate with the police or to obey the law when they are not being observed. In the third place, as societies grow it becomes more difficult to relate public and private interests. If an employer catches a group of employees committing an offense within the workplace, he may not wish to call in the police. He may believe that, if he deals with it in his own way, production and profit may suffer less; the public interest, however, may not be served thereby.

CENTRALIZATION VERSUS DECENTRALIZATION

For these and other reasons, a tension exists between the forces of society that tend to centralize the public police and those that tend to decentralize it. On the one hand, it would appear that communities cannot wholly be left to police themselves, and that a central government should impose its standards upon local communities. On the other hand, if the central government itself is corrupt or self-seeking, the outcome will be worse than if communities were left to enforce law and order according to their own rights. Moreover, good intentions on the part of the central government are in themselves obviously not sufficient. Much depends upon whether it can induce its agents to use their powers in desired ways.

Although the belief still persists that decentralization favours police accountability, some centralization is required for economic and operational efficiency. As the period of training needed for effective police work has grown longer and work hours have been reduced, police manpower has become more expensive; and it has become important to see that police are not employed on tasks for which less highly trained employees could be used. Consequently, police in most industrial countries are now assisted by a host of auxiliaries: photographers, mechanics, fingerprinting technicians, forensic scientists, and so on. It has become advantageous to have national schemes for fixing the pay scales of both police and auxiliaries; for recruiting; for seeing that similar standards are employed throughout the nation; for establishing a common system of criminal records; for developing national facilities for higher training and the selection of top administrators; for forming specialist squads to investigate particular kinds of crime; for research; and, generally, for seeing that the police service keeps abreast of the changing pattern of criminal activity. Such tendencies make for increased centralization, and they operate in a similar fashion in all industrialized countries. As a result, regardless of the way in which police services have started out—whether, like the French, they have been highly centralized or, like the British, have taken community policing as their model—all have tended to develop some common features.

Police specialists

The classification of police as centralized or decentralized organizations is a relative concept. The English police, for example, are generally considered to be a decentralized force; however, when they are compared to police forces in the United States, they appear quite centralized. The organization of police forces around the world forms a wide spectrum, in which the police of the United States represent the decentralized end and the secret police of totalitarian states the centralized extreme.

(M.P.B./Ed.)

The history of policing

Forms of policing have existed for several thousand years, with religious, political, or military police wielding power as early as the time of Babylon. Early police were usually either military or semi-military organizations that developed from the personal bodyguards of rulers and warlords or from community organizations in which citizens banded together for mutual protection.

The duties of the military type of police usually consisted of keeping the public order and enforcing the religious or political mandates of those in power. Rome, under the emperor Augustus, had one of the earliest forms of organized policing. In 7 BC Augustus divided Rome into 14 *regiones* (wards), each divided into *vici* (precincts) overseen by *vicomagistri* responsible for fire protection, among other administrative and religious duties. In AD 6, after a particularly bad fire, Augustus expanded the city's fire brigade into a corps of *vigiles,* consisting of seven squads, or cohorts, of 1,000 freedmen each. Each cohort was responsible for fire and, especially at night, police protection in two *regiones.* To further impose order on the often violent streets of his city of nearly 1,000,000, Augustus created three cohorts of police, part of the army of the state, who were placed under the command of the urban prefect. These cohorts could, in turn, call upon the emperor's own bodyguard (the Praetorian Guard) for assistance.

Military and semi-military police forces developed independently in many countries of the world. For example, the shogun, ruler of 17th-century feudal Japan, devised an elaborate police system in which each castle town had a military samurai warrior who served as town magistrate, judge, and chief of police. He appointed other sword-carrying samurai (*yoriki* and *dōshin*) to serve as a patrolling police force.

In the early 1700s the Russian tsars also established a police system to enforce their laws. Tsar Nicholas I later expanded the powers of this police force and turned it into an early form of state political police—the dreaded Okhranka. After the Russian Revolution, this force gave rise to V.I. Lenin's powerful and highly organized Cheka, the political police that served as a model for Mussolini's OVRA and Hitler's Gestapo.

The other system of early policing, consisting of citizens banding together for mutual protection, was best evidenced by the frankpledge system of early England. This beginning in communal policing eventually led to the development of the Metropolitan Police Act and a British police system that served as the model for most modern police forces.

ANGLO-SAXON BEGINNINGS

As ubiquitous as the contemporary Western police system may seem, it has existed in its present organizational form only for the past 150 years. The earliest policing system in England predates the Norman Conquest. The Saxon frankpledge was a private system of social obligation in which all adult males were responsible for the good conduct of all others. To formalize this social obligation, all males were grouped into tithings headed by a tithingman. Each tithing, in turn, was grouped into a hundred. The hundred was headed by a hundredman, who served as both administrator and judge.

The frankpledge system

When crimes were observed, citizens were expected to raise an alarm, gather their countrymen, and pursue and capture the criminal. All citizens were obliged to pursue wrongdoers, and those who refused were subject to punishment. If a crime was committed with no witnesses, efforts to identify the criminal after the fact were the responsibility of the victim alone: no governmental agency existed for the investigation and solution of crimes.

The frankpledge method of policing continued unchanged until the Norman Conquest in 1066, when the Normans added the office of constable to the system. The office of constable was originally a post in the royal court, but by the late 13th century the position had evolved into a local office of individual manors and parishes. In addition to their frankpledge obligations, constables were responsible for overseeing the "watch-and-ward" system (the nightwatch). The primary responsibility of the watch and ward was to guard the city gates at night. Later, the duties of watchmen were expanded to include lighting lamps, calling time, watching for fires, and reporting other conditions. Despite the addition of constables, however,

the investigation and prosecution of crimes remained a private mattter to be handled by the victims.

The Statute of Winchester of 1285 codified this system of social obligation. It provided that (1) it was everyone's duty to maintain the king's peace, and any citizen could arrest an offender; (2) the unpaid, part-time constable had a special duty to do so, and in towns he would be assisted by his inferior officer, the watchman; (3) if the offender was not caught red-handed, a hue and cry must be raised; (4) everyone was obliged to keep arms and to follow the cry when required; and (5) the constable had a duty to present the offender at court tests.

In one form or another, this system remained in place in Anglo-Saxon countries until the 19th century. Essentially, it was an unpoliced system in which police functions were fulfilled by citizens who held rotating local offices. Constables and watchmen were supported by citizens, posses, and, in the case of riots, by the military or the yeomanry (a cavalry force largely composed of landowners). Victims who could not recover their property offered rewards for its return and often resorted to hiring "thieftakers." These precursors to modern bounty hunters were private citizens who, for a fee or reward, attempted to identify wrongdoers and return stolen property to its rightful owners.

The stipendiary police. The system of private rewards grew until the payment of fees to justices of the peace and constables was firmly established. When communities began paying private citizens for the capture and conviction of thieves, a standard set of fees was established and a "stipendiary" police system evolved. Sources of fees for this system included public rewards, insurance companies, commercial houses, prosecuting associations, and funds raised by subscription. Any citizen, not only constables and justices, could earn those fees and rewards by becoming a thieftaker or "common informer."

Jonathan Wild

Jonathan Wild is a notorious example of a private citizen who abused the system and turned it into a gigantic racket. He organized the London underworld and systematically arranged to have goods stolen so that he could sell them back to the original owners. Any thief wishing to remain independent of Wild's crime ring was delivered to the authorities (and ultimately the gallows). Finally, after a seven-year reign as the "Thief-Taker General," Wild, too, ended up on the gallows.

The stipendiary system was supported by a legal system that decreed Draconian punishments for crimes that, by today's standards, would be considered petty, and capital punishment and serious mutilation were prescribed for almost every conceivable crime. Such harsh punishments were handed out for two reasons—to deter wrongdoers from committing crimes and, failing that, to provide criminals with the opportunity to repent through punishment and save their souls.

The decline of constabulary police. Although the system of social obligation remained in place for more than 800 years and was transplanted to countries like Australia, Canada, and the United States, it had serious weaknesses that were amplified by industrialization and urbanization. The system had become corrupted, specially in the cities. The status of constables deteriorated through the years, and eventually the position became one of subservience to justices of the peace. As the office was degraded, persons of high social status no longer were willing to serve as constables. As a consequence, laws were established allowing persons to hire replacements to serve their terms as constables. While this did not create serious problems in small towns and agrarian areas, only the poor, aged, and infirm were willing to be constables in cities such as London, Boston, and New York City.

Although riots were an established means of political protest in agrarian England, they were an intolerable disruption of economic life in the urbanized England of the Industrial Revolution. The military and yeomanry, therefore, were called upon to quell rioting with increasing regularity. This presented two problems. First, calling in the military was seen as an easing of the restrictions on the political authority of the crown, and it was feared that once in place, the military could extend the king's political jurisdiction. Second, use of the yeomanry to control rioting exposed propertied classes to attack and created another threat to political stability.

The constable system ultimately failed because it was inefficient. Serious crimes and disorders in cities had reached intolerable levels, and the current system was under constant attack from political leaders and writers like Henry Mayhew and Charles Dickens.

Powerful forces opposed reform, however. Every alternative to the stipendiary system required substantial public funding, and raising taxes was not popular. Also, the idea of government becoming actively involved in policing violated the basic tenets of the dominant political philosophy of the era, which posited that the government that governed least governed best. Political opposition to the idea of a standing police force was substantial. Concerned about the threat of political centralization and aware of the political abuses of the French, or "continental," police, political leaders envisioned the political misuse of a standing police force. Finally, provincial leaders saw inefficiency and corruption as "London" problems and believed that the constable system worked satisfactorily in their areas.

THE FRENCH POLICE SYSTEM

Pre-Revolutionary French police. The pre-Revolutionary French police system was totally different from the English system. Organized, financed, and controlled by the government, the French police were actually the personal political police of the king. The system serves as a striking example of action by a central government to enforce its own standards of behaviour. One fact, however, is certain—the 18th-century citizen of Paris was far better protected from crime than the citizen of London. Police in Paris were controlled by the *lieutenant-général de police,* an appointee of the king. Officially under the control of the minister of Paris, the lieutenant general included in his authority the scrutiny of social, political, and economic activities of Paris, as well as a responsibility for controlling crime and maintaining order.

Paris' 20 districts were policed by 48 *commissaires de police,* officials holding both local administrative and judicial authority in the districts; 20 *inspecteurs de police,* investigators responsible for collecting and collating information from each of the districts; 50 *exempts de police,* officials responsible for maintaining order and who were supported by *sentinelles* (foot soldiers) dispersed throughout the city; 10 brigades of 10 archers each who patrolled throughout the city; a watch guard consisting of 220 horsemen and 400 foot soldiers who also patrolled the city; and *mouchards,* irregular spies, or secret police, who obtained information on the activities of all citizens. The police system also was supported by the military, which could be called upon at any time by the lieutenant general.

Political abuse of the early French police

Despite its use of preventive patrols, the system was characterized (and made infamous) by its undercover intelligence-gathering activities. *Inspecteurs de police* and *mouchards* gathered enormous amounts of criminal and personal information from spies, brothel keepers, employers, and interdicted mail. Indeed, at one time it was estimated that one-quarter of all the housemaids and lackeys in Paris were receiving money from the police in return for information about their employers and acquaintances. The police chief Sartine once boasted to Louis XV, "Sire, whenever three people speak to one another in the street, one of them will be mine." Each day this information was passed on to the king through the *bulletin moral* and was further distributed to political officials through the *bulletin politique ou d'espionnage.* The king could issue a *lettre de cachet* ordering the imprisonment of persons even if there was insufficient evidence to bring them to trial. Such letters allowed police to detain people without charge and to exercise a kind of summary justice.

At the same time the ordinary police took an active part in many aspects of the city's life. The lieutenants general of police built markets, paved and constructed streets, founded the Bourse (stock exchange), moved a cemetery, provided street lighting, established a system of state pawnshops, created a hospital for children, founded a veterinary college, built schools for children of the poor, sought work for the unemployed, and organized fire and

river rescue services. The police took stock of the city's food supplies and forbade wine sellers to have leaden counters and milkmen to use copper vessels. They censored posters and regulated the employment of foster mothers. They persecuted Protestants, controlled the sale of lottery tickets, and inspected drains. They had the streets cleaned. They supervised cabdrivers, made regulations about dogs, checked food prices, surveyed dangerous buildings, read people's letters in the course of post, and even stopped fishmongers from selling oysters between April 30 and September 10. The police inspected prisons, read theological works for signs of heresy, and had a financial agency for the encouragement of overseas trade.

National police (*maréchaussée*), originally created to control crimes committed by patrolling soldiers, eventually were used to deter civilian crimes as well. Immediately before the Revolution, the *maréchaussée* consisted of 3,000 men organized into 30 patrolling companies. Officers had both the privileges of soldiers and the powers of civil arrest.

Post-Revolutionary French police. After the Revolution, the Paris police was reconstructed, and its direction was entrusted to a conglomeration of Parisian committees. The clumsy committee arrangement failed, and in March 1800 control of the police was centralized in the *prefet de police*, an office with administrative but no judicial authority.

The French Ministry of Police was created in 1796 with a mandate to enforce laws and forestall challenges to the Revolution. When Joseph Fouché was appointed director in 1799, he consolidated police administrative powers into four *arrondissements*, each directed by a *conseiller d'État*. His councillors operated with his authority and had few legal restrictions on their power. Ultimately, Fouché's power in France became second only to Napoleon's.

In Paris, Fouché maintained the pre-Revolutionary orientation of the police by recruiting more than 300 spies, including many ex-convicts, into the *regulators de l'opinion à Paris*. Spying was so ubiquitous that private individuals, the military, and even the Jesuits organized their own secret police forces. Eighteenth-century Parisians believed that police agents and spies were everywhere, listening for the smallest hints of discontent. They may have been right, but the belief that nothing that happened in Paris would go unnoticed by the police was far more helpful to the authorities than a host of paid undercover agents. The potential of this system to be politically misused, to stifle dissent, was enormous. Police regularly disrupted, controlled, and even shaped political thought, until corruption and injustice had riddled every phase of the system.

Vidocq and the Sûreté

Perhaps the most famous manifestation of the French detective orientation of policing, at least with crimes of robbery and theft, was the Sûreté. The Sûreté was founded by François-Eugène Vidocq in 1810. A convict and a daredevil jailbreaker, Vidocq approached the police and offered to use his knowledge of the criminal world in exchange for his freedom. The police agreed, and eventually Vidocq became chief of the criminal police. Running the Sûreté with the philosophy that to know and capture criminals one had to be a criminal oneself, Vidocq directed a network of spies and informers in a war against crime that was completely successful. His knowledge of the underworld and his reliance on criminals to catch criminals were the reasons for his success. In 1817, with only 12 full-time assistants, he was responsible for more than 800 arrests.

THE FORMATION OF THE ENGLISH POLICE

The creation of the Thames River Police was one of the most significant police reform experiments of the time. The first regular professional police force in London, the Thames River Police was organized to curb the thefts that plagued the world's largest port. The West India Trading Company created these marine police in June 1798 and installed police reformer Patrick Colquhoun as director of a permanent staff of 80 men and an on-call staff of 1,120. Two features of the marine police were unique. First, patrols were preventive; officers patrolled visibly to prevent thefts. Second, officers were not stipendiary police; they were salaried and were prohibited from taking fees. The venture was a complete success, and reported crimes dropped appreciably. In July 1890 the House of Commons passed a bill making the marine police a publicly financed organization, which ushered in a new era of policing.

Debate about the creation of a standing police force in England raged during the early part of the 19th century. Confronted with political objections and fears of potential abuse Robert Peel (later Sir Robert Peel) sponsored the first successful bill creating a bureaucratic police force in England.

The Metropolitan Police Act was passed in 1829 as a political compromise. The jurisdiction of the bill was limited to the metropolitan London area, excluding the City of London and provinces. All police were to be uniformed; crime and disorder were to be controlled by preventive patrols; police were to be paid regular salaries; and no stipends were permitted for successful solutions of crimes or the recovery of stolen property. But crime prevention was not the only business of the new police force: they inherited many functions of the watchmen, such as lighting lamplights, calling time, watching for fires, and providing other public services. "Bobbies," named after Peel, were not immediately popular. Most citizens viewed constables as an infringement on English social and political life, and people often jeered the police.

The preventive tactics of the early metropolitan police were successful, and crime and disorder declined. Their pitched battles with (and ultimate street victory over) the Chartists in Birmingham and London proved the ability of the police to deal with major disorders and street riots.

Despite the early successes of the metropolitan police, the expansion of police forces to rural areas was gradual. The Municipal Corporations Act of 1835 ordered all incorporated boroughs to set up police forces under the control of a watch committee, but it was not until 1856 that Parliament mandated that provinces establish police forces.

The origin of modern English policing

The Metropolitan Police Act established the principles that shaped modern English policing. First, policing was to be preventive, and the primary means of policing was conspicuous patrolling by uniformed police officers. Second, command and control were to be maintained through a centralized, quasi-military organizational structure. Third, police were to be patient, impersonal, and professional. Finally, the authority of the English constable derived from three official sources—the crown (not the political party in power), the law, and the consent and cooperation of the citizenry.

EARLY POLICE IN THE UNITED STATES

The United States inherited England's Anglo-Saxon common law as well as its system of social obligation and constables. As both countries moved from rural agrarian to urban industrialized economies, urban riots, crime, and disorder followed. Yet Americans, like the English, were wary of creating standing police forces.

In the early 1800s, as the United States became more urban, patterns of ethnic diversity began to erode the social and political hegemony of the original English and Dutch settlers. Not only did rioting, crime, and disorder begin to flourish, but the basic life-styles of new German and Irish immigrants offended the moral and social sensitivities of the original settlers. Immigrant and urban behaviour was perceived as a threat to the social, economic, and political fabric of American life.

The American response to growing urban unrest was twofold. Versions of the constable and nightwatch system were tried, and voluntary citizens' groups were encouraged to try to solve urban problems. Reformers distributed religious tracts and Bibles, started Sunday schools, created organizations such as the Young Men's Christian Association, and presented themselves as moral examples to immigrants and the poor. By the mid-19th century, middle-class frustration with the deterioration of the cities led to the passage of laws regulating public behaviour and creating new public institutions of social control and coercion—penitentiaries, asylums, and police forces.

In 1844 New York City created the first police department in the United States, using the London Metropolitan Police as a model. Boston and Philadelphia followed. The

idea of police spread quickly, and in 10 years cities as far west as Milwaukee had created police departments. The model for American police originated in England: police were organized in a quasi-military command structure; there were no detectives; their task was the prevention of crime and disorder; and they provided a wide array of public services.

Each city created its own police department. Since Americans lacked a unifying symbol like the English crown, local politics and laws became the primary bases of police authority. The decentralization of the authority for policing was extended to political wards and neighbourhoods, which developed relatively autonomous police units. The police established intimate relations with neighbourhoods and neighbourhood leaders and initially did not even wear uniforms. Middle- and upper-class reformers believed that one of the primary tasks of the police was to reestablish political and social control over a population wracked by ethnic and economic rivalries. This tension between being closely linked to communities and being perceived of as an instrument for reform of those same communities inevitably resulted in struggle for political control of the police. This struggle for political control is one of the dominant themes in the history of police in the United States.

DETECTIVE POLICING IN ENGLAND AND THE UNITED STATES

The investigation of crimes was not a central function of the newly formed preventive police departments in England and the United States. In both countries, thief-takers, stipendiary police, and constables continued their investigations of criminal activities. Despite the high hopes of reformers when they created the police, the number of preventable crimes was limited. As victimizations continued, police were pressured into accepting public responsibility for investigations and creating detective units. The London Metropolitan Police created the first criminal investigation division (CID) in 1842. Many American cities followed with detective units: Boston in 1846; New York City in 1857; and Chicago in 1861.

Investigators usually were recruited from former thief-takers or constables who had continued their stipendiary investigative activities after the creation of police departments. Although they brought investigative skills to the police, they also brought the bane of stipendiary police—corruption. In 1877, when three of the four chief inspectors of the CID were found guilty of corruption, the CID was eliminated although later reorganized. Chicago disbanded its unit in 1864 and Boston its unit in 1870, and New York City suffered major scandals in 1877—all as a consequence of corruption. All of these cities soon reconstructed their investigative units, but improvement of the status of detectives had to wait until well into the 20th century.

LATE 19TH-CENTURY ENGLISH AND AMERICAN POLICING

After the Act of 1856 mandated police in the provinces, police departments spread throughout England. Provincial police were funded by both local and central governments. After the Home Office certified the quality of a provincial police department, the central government paid half of the cost of local policing, and local taxes paid the rest. The dominant methods of provincial policing were foot patrols and criminal investigations.

Strengths and weaknesses of decentralized policing

Policing in the United States during the late 19th century, however, was complicated by migration and immigration, which continually reshaped the ethnic and cultural makeup of cities, and by the radical decentralization of police authority within the cities. Decentralized policing had one major strength: it gave police close contact with the public. Police knew the local citizens, and often they were recruited from the very neighbourhoods they policed. Such close communities allowed police to spot troublemakers, identify local problems, and provide an array of public services to the residents, such as manning soup lines and providing overnight lodging in police stations for immigrants.

The decentralization of police jurisdictions to individual cities also created problems. Crime did not respect jurisdictional lines, but police organizations had to. Thus, although serious crimes were rife on the American frontier, no national or state police organization had jurisdiction over crimes committed there. A second problem arose from the lack of a national organization. Since no binding ideology of policing existed, each police organization was an entity unto itself. Each department formed local policy and had little or no communication or coordination with other police organizations. Unlike English policing, for which Sir Robert Peel had articulated a unifying vision, American policing was atomistic and without a coherent identity.

In response to the intra-jurisdictional crime waves of the latter half of the 19th century, states enacted laws giving many corporations the authority to create their own private police forces or to contract with established police agencies. The Coal and Iron Police of Pennsylvania was an example of a company police force that later became notorious for its antilabour vigilantism. The most famous independent police force was the Pinkerton National Detective Agency. Created by Allan Pinkerton, a political fugitive from Scotland, the Pinkerton agency provided a wide array of private detective services, specializing in protecting trains, apprehending train robbers, and participating in anti-labour union activities.

Attempts in the late 19th century to develop a coherent vision of an intercity police system were largely unsuccessful; the police theory that did develop, however, was formulated mainly in local political halls. As ethnic residents gradually gained political control of city wards and neighbourhoods, the link between the police and neighbourhood politics became closer. In some instances the relationship became so close that the police were actually seen as adjuncts of the local political machines. The linking of police and politics bred political and financial corruption and injustice. Police became involved in partisan political activity to ensure election of particular candidates; they received "gratuities" for not enforcing unpopular vice laws; and they excluded strangers from social and political life.

By the end of the 19th century, middle- and upper-class citizens attempted to centralize political power to end the ward-level political control of ethnic minorities. Reformers attempted to centralize services on a citywide basis, create a civil service to end political patronage, and transfer control of police to cities-at-large, or, if all else failed, transfer control of the police to the state government.

DEVELOPMENTS IN AUSTRALIA AND CANADA

Australia. Settled as a penal colony in 1788, Australia first used the British constabulary and watch-and-ward systems. Problems plagued these systems, however, because both constables and watchmen were often recruited from the ranks of convicts. Modeled after England's Metropolitan Police Act, the Sydney Police Act of 1833 led to the establishment of urban police forces. Police coverage was extended to rural areas in 1838 when each of the country's six states created its own state police agency.

Although the state police encountered a lack of public acceptance and many of the same problems of police in England and the United States, their task was complicated by additional responsibilities. Not only were they mandated to capture villains, but they also were responsible for inflicting corporal punishment on convicted criminals. Moreover, from their inception, Australian police duties included the enforcement of health and welfare provisions of the law.

Canada. Canada's earliest legal traditions can be traced to both French and English origins. Quebec City followed French traditions and created a watchman system in 1651. Upper Canada, later renamed Ontario, adopted English traditions and established both a constabulary and a watch-and-ward system. The English system was imposed on French areas after 1759. Using England's Metropolitan Police Act as a model, Toronto created a police department in 1835, Quebec City in 1838, and Montreal in 1840. Provincial police forces were established in 1867 to provide police coverage for the vast rural areas in eastern Canada.

The North West Mounted Police

The North West Mounted Police (renamed the Royal

Canadian Mounted Police in 1920) were created in 1873 to police the vast western plains. The original 300 officers initially were assigned the task of eliminating incursions of whiskey-trading Americans who were inciting Canadian Indians to acts of violence; later the force spearheaded attempts to mold the Canadian frontier into an integral part of Canada.

The North West Mounted Police represented a significant departure from Anglo-Saxon policing traditions. Reminiscent of continental policing in organization, style, and method, they operated more like a military organization than the traditional police force. Strong leadership ensured that they operated with restraint and within Canadian political traditions. The Royal Canadian Mounted Police, along with city and provincial police forces, constitute the three-tier police system now in place in Canada.

20TH-CENTURY POLICING

The struggle for political control of police in the United States gave rise to a distinctive strategy of policing that was to influence policing throughout the Western democracies. This new style of policing integrated managerial techniques, sources of authority, innovative tactics, and a narrowed definition of police work.

The first two decades of the 20th century were tumultuous times for American police: progressives battled entrenched ward and "machine" leaders for political control; labour unrest and concerns about Communist influences preoccupied many politicians and local police organizations; political and economic corruption were widespread at all levels of government; and, the Prohibition movement was becoming a political force. Riots, mass demonstrations, and bombings were not unusual. Arrayed against these vast social problems were scattered local police forces and private detective agencies. So limited was their ability to handle such diverse national and local problems that federal and state governments were forced to create additional police agencies.

Federal and state police. The United States Secret Service was created in 1865 to prevent counterfeiting. Never numbering more than 30 or 40 agents during the 19th century, the agency was operated in the traditions of the previous century. During the 1890s the Secret Service occasionally was called upon to guard the president, but that was not made a permanent duty until 1901.

Federal Bureau of Investigation

The Bureau of Investigation—later the Federal Bureau of Investigation (FBI)—was created in 1908. In debates reminiscent of those in the British Parliament a century earlier, the U.S. Congress opposed the creation of a secret investigatory agency. In this case the concern was more than ideological—the Secret Service had been investigating corruption in Congress as well as in governmental agencies, and many congressmen were wary of increasing the president's investigatory powers. At Pres. Theodore Roosevelt's direction, Charles J. Bonaparte bypassed the reluctant politicians and created the bureau by executive order after Congress had adjourned. The Bureau of Investigation began with a modest mandate to investigate antitrust cases, several types of fraud, and certain crimes committed on government property or by government officials.

In 1920 the Treasury Department created the first sizable federal police agency. Charged with enforcing Prohibition, T-Men, as they came to be known, grew to a force of approximately 4,000 during the peak of the crusade against alcohol.

During the early 1900s states also began creating police forces. Although some states, such as Texas and Massachusetts, had small police forces earlier, in 1905 Pennsylvania established the first modern state police department. Formed with the professed purpose of fighting rural crime, state police in Pennsylvania (and later other states) were used primarily to circumvent corrupt or inefficient local police forces and to control strikes where city police were considered to be sympathetic to the unionists. Other states followed Pennsylvania's lead: New York in 1917; Michigan, Colorado, and West Virginia in 1919; and Massachusetts in 1920. Regarded as models of efficiency and honesty, state police were the prototypes of what came to be known as professional police organizations: highly disciplined, narrow in focus, and organized along rigid military lines of command and control.

Municipal police reform in the United States. Most efforts to reform the police system during the late 1800s originated from reformers who were outside the occupation of policing. During the early 1900s, pressures for reform started within the police system itself. The most notable and representative police reformer was August Vollmer. Beginning his career in 1905 as the head of a six-person police department in Berkeley, Calif., Vollmer ultimately offered a vision of policing around which the nation's police rallied.

Prohibition was the worst of times for police: corruption flared, fueled by vast amounts of bootlegging profits; public confidence declined; and the profession came to be symbolized by Hollywood's Keystone Kops, who portrayed police as inept and venal. Vollmer and his colleagues were concerned about the broad social issues of policing. Changes in morals, increases in crime and corruption, and, later, the Great Depression, all were seen by reform-minded police as symptoms of the erosion of the authority of basic social institutions like the family, church, schools, and neighbourhoods.

The police reforms of August Vollmer

Vollmer saw the police as the vanguard for socializing America's youth. In his view police should continue their traditional law enforcement role, and when necessary they should arrest and process delinquent youths through juvenile and adult courts. Arrest, however, was an undesirable outcome. Special juvenile bureaus should be created to handle problems of children and families; police should take a more active role in casework for social agencies; police should exploit their intimate knowledge of the community and place themselves at the hub of community activities with youth and families.

In addition to giving police an ideal to strive for, Vollmer also helped consolidate the International Association of Chiefs of Police (IACP) into a truly national police organization. Under its auspices he created the Uniform Crime Reports, which became an important indicator of the health of society and of the performance of police departments. Finally, through his work on the Wickersham Commission, Vollmer exposed many unconstitutional police practices to public scrutiny, especially the practice of detectives using the "third degree" in questioning suspects.

New models of policing. When J. Edgar Hoover took over the Bureau of Investigation in 1924, he laid the groundwork for a strategy that was to make the FBI one of the most prestigious police organizations in the world. The public's opinion of detectives was ready for change. Inspired by detective-heroes in the novels and short stories of Charles Dickens, Edgar Allan Poe, and Sir Arthur Conan Doyle, readers developed a new interest in real-life accounts of detectives' exploits. Hoover set out to make the fictional image of the detective into reality. He eliminated corruption by suspending bureau investigations requiring considerable undercover or investigative work (*e.g.*, vice and, later, organized crime) and by creating a strong bureaucratic organization that emphasized accountability. Educational requirements were established for new agents, and to ensure that they were well trained in police procedures, Hoover created the FBI Academy. He concentrated his resources on crimes that received great publicity and were relatively easy to solve, such as bank robberies and kidnappings. But most of all, Hoover assiduously cultivated publicity to create an image of the "G-Man" as the nation's incorruptible crime fighter.

O.W. Wilson's new police model

Vollmer's idealistic vision of police work, with its strong elements of social work, was soon replaced by Hoover's strategy. Ironically, Vollmer's protégé, O.W. Wilson, became the architect of the new police model. Instead of broadening police responsibilities as Vollmer had proposed, the new reformers narrowed them to concentrate on fighting serious street crimes. They also moved to sever the close ties between police and neighbourhoods: assignments were changed often; police no longer patrolled areas in which they lived; and, most importantly, police began to use automobiles to patrol. To insulate police from political influences, reformers created a civil service to hire and

promote officers and modified the basic police authority from law and politics to law only (especially criminal law). Finally, administrative decentralization within cities was abandoned, and police departments developed centralized citywide bureaucracies that restricted discretion at all levels of the organization. Since its introduction this model has dominated American policing. As the political influence of the United States grew after World Wars I and II, American policing methods spread to other nations.

Until the 1930s and '40s, police tactics were based on crime prevention by conspicuous foot patrol. As automobiles became a more important part of American life, and as reform-minded chiefs broke police officers' ties to neighbourhoods, more officers were assigned to automobile duty. The full motorization of American police was largely accomplished after the end of World War II.

The rationale for using automobiles in anticrime patrols was manifold. Police cars moving randomly and quickly through city streets would create a feeling of police omnipresence to deter criminals and reassure citizens of their safety. Rapidly patrolling police also would be able to spot and intercept crimes in progress. The use of radios in police cars increased the value of automobile patrols. The rapid response to calls for assistance would either deter criminals or help in their immediate apprehension. Police throughout the United States set an optimum goal to arrive at the scene of a crime within three minutes of its first being reported.

Contemporary policing. Wilson's strategy of policing came into fruition during the 1960s. Indeed, in 1967 the President's Commission on Law Enforcement and Crime, which was critical of the strategies of other criminal justice agencies, endorsed both preventive patrol and rapid response to calls. They agreed that improvement of policing would come as a result of fine-tuning police organizations, equipment, and personnel and that the basic strategy of policing was satisfactory. The commission noted that preventive patrol elicited hostility from some, especially minority, communities, but that its anticrime potential was so great that it had to be maintained. Police community-relations programs were proposed to offset the negative results of preventive patrol.

British police strategy

In Britain the police strategy of Sir Robert Peel was successful during the 20th century. Although influenced to some extent by the tactics of American police, the British approach appealed to the public and remained effective. Foot patrols continued in most cities, which lacked the suburban "sprawl" of American cities. Although "fire brigade" policing, as many British characterized the rapid-response orientation of American police, had some influence in Britain, it was counterbalanced by the continued emphasis on the neighbourhood bobby.

Aside from terrorism, which confronts police in all countries, the most serious contemporary problem facing British police surfaced during the 1980s—urban rioting by youths, often minorities, who believed that police tactics were brutal and unjust. The Scarman Report, which resulted from an official inquiry, suggested that police had become too remote from their communities, that local citizens should have more input into police policy-making, and that police tactics should be more sensitive to the growing cultural pluralism of Britain's major cities.

Research and new developments in policing. American policing strategy underwent a crisis during the 1960s. Crime continued to rise despite massive financial outlays for increased numbers of police. The depth of hostility that minority communities felt toward police surfaced during the riots and civil disturbances of the decade. Minority leaders not only resented the police tactics used to suppress civil disturbances, but many also believed that police actions actually caused the riots. With the fear of crime escalating more rapidly than crime itself, citizens refrained from using public and private facilities, took self-defensive measures, and abandoned city streets and parks; many fled the cities completely. Research studies were initiated to find a means of reversing these trends.

In the 1950s and '60s, both civilian and police groups assumed that the primary activity of police officers was dealing with crime and doing so with little discretion. Research on police functions conducted during this period, however, showed that when police activities, calls, and dispatches were analyzed, anticrime activities constituted less than 20 percent of patrol activities. The remaining patrol functions included resolving conflicts, providing emergency services, maintaining order, and providing other public services. Moreover, it was discovered that police officers regularly used discretion in handling events, criminal or otherwise, and that the use of discretion was an essential ingredient of police functions.

One of the earliest and most sophisticated research efforts concerning the effectiveness of police patrol was the Bright Foot Patrol Experiment in Britain in 1970. Although this experiment indicated that crime declined in areas covered by officers on foot patrol, efforts to further decrease crime by using additional foot patrols in the same neighbourhoods did not have a measurable impact.

American studies on the efficacy of preventive automobile patrols found that, despite the commitment of substantial amounts of police time to patrol, relatively little crime-related activity resulted from police initiatives. The vast majority of arrests (93 percent) resulted from citizens' requests for police action. Later studies suggested that preventive patrol by automobile did not effectively attain its goals to reduce crime, increase the public's satisfaction with police, or decrease citizens' fear of crime. Moreover, research studies on the efficacy of rapid response to calls for service suggest that it has little impact on crime prevention or criminal apprehension and that alternative approaches might produce greater levels of citizen satisfaction.

Introduced in the New York City Police Department by Patrick V. Murphy, team policing emphasizes decentralized patrol decisions by teams of officers policing a limited geographical area. Patrol decisions are made on the basis of relatively close communication with local leaders and residents.

Eliminating opportunities for crime

Another promising area of police activity focuses on eliminating the opportunities for crime. Certain conditions, such as leaving one's keys in a car, invite crime. By identifying conditions that increase the likelihood of criminal activity, police can help develop policies and procedures to alter conditions and thereby make crimes more difficult to commit. Automobile thefts can be reduced, for example, by making automobiles more difficult to steal.

In the 1970s and '80s strong community anticrime efforts also were developed in many countries, including the United States, Canada, and Britain. In a sense, these groups represented a return to the ancient tradition of social obligation, with each citizen obliged to come to the aid of others. Although police were instrumental in developing such groups, in cities such as Boston, Philadelphia, and Chicago, which had traditions of strong neighbourhoods, many groups developed on their own. Research suggests they are relatively effective in reducing crime and fear in neighbourhoods. Police experience and research suggests that police action, in itself, is limited in its ability to prevent and investigate crime. Police can be successful only to the extent that they work closely with citizens.

Indeed, the most promising development in American policing during the 1980s was community policing. Incorporating many of the ideas of team policing with research on foot patrols, community policing is an attempt to reintegrate the police officer into the community. Community policing has since proved so effective that most major U.S. cities, including New York City, Los Angeles, Chicago, Houston, and others, have either implemented the procedure or experimented with it. (G.L.Ke.)

CENTRALIZED POLICE ORGANIZATIONS

The United States and Britain typify countries that have developed decentralized police organizations. France and Italy, on the other hand, are examples of countries with highly centralized police forces.

France. The centralized police force of France's Revolutionary period gradually developed into two national police organizations: the Police Nationale and the Gendarmerie Nationale. The Police Nationale was formed in July 1966 by the amalgamation of the Prefecture of Po-

lice of Paris and the Sûreté Nationale, which had been responsible for policing the provincial towns and for certain specialized work. The director general's headquarters includes divisions covering general intelligence, judicial police, public safety, security, regulations, training, and personnel.

The Gendarmerie Nationale is composed of soldiers formed in strictly disciplined units and quartered in barracks. In the larger towns there are Gardiens de la Paix who serve as municipal police, responsible for preventive patrol and the control of traffic.

The mayors of certain communes have communal or rural policemen at their disposal; but, as a rule, police work in country districts and in towns with fewer than 10,000 inhabitants is undertaken by gendarmes.

These divisions within the French police service, and the separate ties that units have with local administrators, suggest that, though centralized, the French police is far from monolithic. There are many opportunities for one police body to check upon the activities of another to prevent corruption.

Italy. Italy, too, has several national police forces. The Public Security Guards (Guardia di Pubblica Sicurezza), which is part of the armed forces, is charged with the maintenance of internal public security and order, with the protection of life and property, and with preventing and checking crime and gathering evidence in criminal matters. Although the force is of a semi-military character, it performs all normal police functions and is responsible to the minister of the interior. The Corps of Carabinieri is a military-like body housed in barracks; it is responsible to the minister of defense in certain matters but to the minister of the interior in regard to police duties. As an executive organization for the detection of crime, the corps carries out duties assigned to it by law, under the direction of judicial authorities to whom it reports on investigated crimes. At every level, commanders of the corps must maintain liaison with the army and with political, judicial, and police authorities. A Guardia di Finanza is charged principally with protecting the national frontiers from activities connected with smuggling, counterfeiting, illegal entry, and tax evasion. In addition to the national forces, municipalities maintain their own police, the Vigili Urbani, which undertake certain kinds of municipal administration but are increasingly concerned with traffic control. Frequent attempts have been made to amalgamate the Vigili with the Public Security police, but the municipalities have resisted such proposals.

DECENTRALIZED POLICE ORGANIZATIONS

United States. The United States provides a picture of what may be the most decentralized police system in the world, accompanied by an extraordinary degree of duplication and conflicting jurisdiction. Because there is no federal government control to ensure uniform standards, every community is entitled to run its own police department but cannot exclude federal or state officials from investigating offenses over which they have jurisdiction.

Major types of U.S. police agencies

There are five major types of police agency: (1) the federal system, consisting of police officers attached to the Department of Justice (the Federal Bureau of Investigation, together with 700 U.S. marshals), the Bureau of Internal Revenue (investigating tax evasion), the U.S. Secret Service, the Drug Enforcement Administration, the Postal Inspection Service, and many others; (2) police forces and criminal investigation agencies established by each of the 50 states of the union; (3) sheriffs and deputy sheriffs in more than 3,000 counties, plus a few county police forces that either duplicate their police jurisdictions or displace them; (4) the police forces of about 1,000 cities and more than 20,000 townships and New England towns; and (5) the police of some 15,000 villages, boroughs, and incorporated towns. To this list must be added special categories, such as the police of the District of Columbia; various forces attached to authorities governing bridges, tunnels, parks, and parkways; university, or "campus," police forces; and some units for policing special districts formed for fire protection, soil conservation, and other diverse purposes. The usual estimate is that there are about 20,000 separate public police agencies in the United States and that the majority of them consist of one, two, or three men who are not employed full-time. Many are not paid wages but are compensated by the payment of fees for duties undertaken.

These police forces are controlled in different ways. The federal bodies are responsible to the federal authorities, and the state forces to the governors of their states. At local levels there are variations. States are divided into counties, and every county has its sheriff, usually elected and with responsibilities for the investigation of crime and the supervision of the custody of prisoners. He may have deputies performing duties as uniformed policemen. In some states there are county police forces under the control of county boards, whereas elsewhere, appointments are made by state governors on the recommendation of the county authorities. In some places county police forces are completely independent of the sheriff's office.

The existing American police structure to some extent reflects public opposition to any concentration of police power. It has been argued that the nation would suffer should all offenses become federal offenses and should power be transferred to Washington, D.C., and local government be enfeebled. Local problems require local remedies, and the weaknesses of big government are all too apparent. On the other hand, it is argued that the integration and consolidation of police forces could result in an improved service that would reduce costs, increase efficiency, and eliminate the problems of overlapping jurisdictions. The first steps in this direction have been taken in respect to training. Since 1960 no police officer in the state of New York has been able to assume his duties until he has received a certificate for successfully completing basic training. California has introduced grants-in-aid for training that are so attractive that most municipalities are taking advantage of them. Other states have adopted similar schemes. The federal government has made available to the police, through the Law Enforcement Assistance Administration, substantial funds for improving police resources and their management.

Australia. Certain other countries with a federal political structure have federal police forces complementing state forces, but the principles involved are no different from those exemplified by the United States. In Australia, for example, each of the six states has its own laws but does not legislate in matters pertaining to federal organizations and cannot pass laws at variance with those of the commonwealth. Each state has its own police force. The minister of administrative services oversees the Australian Federal Police, which enforces federal legislation. Other police forces under federal control are responsible for overseas territories, the Northern Territory of the subcontinent, and the capital territory.

Australian state police

Former Soviet Union. In many countries, changes in the structure of police administration have been in the direction of greater centralization; and it is of interest to consider a country where the change was in the reverse direction. After the death of Joseph Stalin the police force of the Ministry of State Security of the Soviet Union was absorbed into the force of the Ministry of Internal Affairs (MVD) under Laventry Beria, who was arrested less than a year later. This force was then divided. The organization dealing with state security was transferred to the Committee of State Security under the Council of Ministers of the U.S.S.R.; known by the initials KGB, it was understood to be primarily a counterintelligence agency. The body that carried out duties performed elsewhere by the normal uniformed police was the militia. Prior to 1956 it was centralized under the Ministry of the Interior, but in December of that year the administration was decentralized. Units were subject to control by the higher organizations of the militia, by the soviet of the local administrative area, by the local representative of the procurator general of the U.S.S.R., and by the courts. In 1960 the MVD was abolished, and in some areas the militia was itself replaced by Public Order Detachments made up of workers, builders, farmers, teachers, and students. Two and a half million citizens were enrolled into these detachments as voluntary militia. Two years later these

new detachments were being criticized in the Soviet press for using their powers in too high-handed a way. To the outsider it seemed as if the Soviet Union had found that a highly centralized police service was not accountable to the public and that better responsiveness required drastic decentralization, even to the extent of occasionally relying upon citizen volunteers. (M.P.B./Ed.)

SECRET POLICE

Secret police are a special category of police. Established by national governments to maintain political and social orthodoxy, they are generally clandestine organizations that operate independently of regular civil police.

Secret police not only have the traditional police authority to arrest and detain, but in some cases they are given unsupervised control of the length of detention, assigned to mete out punishments independent of the judiciary, and allowed to administer those punishments without external review. Civil police or militia may continue their traditional common policing roles and may even, at times, restrain secret police, but the tactics of investigation and intimidation permit secret police to accrue so much power that they usually operate with little practical restraint.

Abuses of secret police

Two notorious 20th-century secret police forces, the Nazi Gestapo and the Russian NKVD, were formed during the early days of unstable regimes. The Nazis, a minority party elected to power, quickly moved to cement their political control by taking control of the civil police and the military with their secret police organization. Their tactics included the arrest, imprisonment, torture, or assassination of political foes, as well as subversive activities aimed at discrediting political enemies and contenders for power.

Russian Communists, after a surprisingly successful political revolution, used secret police (then known as the Vecheka or Cheka) to intimidate and eliminate political enemies. Later Stalin used the force to eradicate all competing political forces and contenders for authority both inside and outside the Communist party.

In the former Soviet Union, Nikita Khrushchev's public revelations of the horrors of Stalin's NKVD is thought to have greatly reduced the organization's powers. Afterward, even though the Soviet secret police operated with more restraint, the potential for serious abuses still existed until the country's dissolution in 1991–92.

On occasion some police agencies in liberal democracies, like the CIA in the United States or the DGSE in France, also take on characteristics of secret police when they move to discredit political or social deviants. However, since most democracies establish governmental agencies to scrutinize and control secret police organizations, there is little comparison between their excesses and those of genuinely secret police like the Gestapo. (G.L.Ke.)

Police work and law enforcement

ROUTINE POLICE ACTIVITIES

Earlier reference was made to the way in which police activity is adapted to the kind of society that is to be policed. There are features common to the work of police in differing societies that are founded upon similar technology. Yet within the same society, at times within the same police force, there are variations in the policies of police chiefs. One police administrator, because of his personal beliefs or because of his perception of public opinion, may allocate more officers for the investigation of vice than of fraud, may allocate more or fewer officers for traffic control or crime prevention. A police chief may instruct officers not to bother about prostitution, provided it is confined to a given locality. Thus, the police in different neighbourhoods may follow different patterns of policing.

Within the framework of enforcement policy, police work is divided into various branches. The largest number of officers is usually allocated to uniformed patrol, either on foot or with motor transport. Studies of the activities of police on patrol indicate that only a small portion of their time goes to the making of arrests or to initiating formal actions under the criminal law. Moreover, whether one examines the types of calls for service that police receive, the calls to which police are dispatched, or the activities that police initiate on their own, it is clear that the majority of police activities consist of providing emergency services, maintaining order, resolving disputes, and providing other services.

POLICE AND COURTS

Judicial intervention

One of the most important ways in which police are held accountable for the manner in which they perform their duties is through the courts. In France (and in countries with similar juridical procedures) the police making inquiries are under the direction of the investigating magistrate, whose task it is to decide whether there is a case for trial and, if so, what it is. The French think it improper that police should inquire into a citizen's activities without his knowledge that they are doing so and without these investigations being under judicial control. The British and American procedure is very different. Anyone may prosecute. The judge listens to the prosecution's argument and evidence that the accused has broken a particular law, and he listens to the defense. Then he decides between them. If he finds the case proved, he then inquires into the accused's previous record and determines the sentence. Should it happen that the defendant has broken a law but not the one he is accused of having broken, the case must be dismissed.

If the accusatorial system is to function justly, the police must bring all cases of lawbreaking before the courts. It is for the court to decide whether they merit punishment. However, in practice, the police exercise considerable discretion as to whom they will prosecute.

There are three chief arguments in favour of police discretion. First, it has not been possible to draft and keep up-to-date a criminal code that unambiguously encompasses all conduct intended to be made criminal and none other; there are technical offenses or offenses that public opinion no longer regards as culpable. Second, those charged with enforcing the law do not have sufficient resources to enforce all the laws all the time, so that enforcement must be selective. Third, to bring a technical offender to court may, in practice, unduly damage that person's reputation.

The courts control police activities in other ways. In Britain a set of publicized "Judges' Rules" outlines safeguards for accused persons while under investigation. If in court it can be proved that the police failed at the proper points to warn a person that he was under suspicion of having committed a particular offense and that any statement he might make could be given in evidence, then the prosecution might fail. In the United States the Supreme Court has held that an accused need not answer police questions and that he has a right to consult a lawyer before questioning; if he cannot afford to hire a lawyer, one must be provided. In other decisions, police have been prevented from submitting at a trial any evidence that was obtained by unreasonable search and seizure.

CROWDS AND RIOTS

The London police have developed skillful methods of crowd control but have had little experience of urban rioting on the scale with which American police were confronted in the 1960s. Both crowds and riots require patrol officers to adopt a very different approach from that to which they are accustomed. Normally, police are free to handle members of the public in an individual manner. But the control of civil disturbances requires large numbers of disciplined personnel working as a team under unified command.

British methods of crowd control

From their beginnings in 1829, the English metropolitan police were careful not to overreact when handling crowds. In Britain, if the police fail to control a riot, recompense for the damage has to be paid by the police authority for the district. Were the police to use, say, water cannon, an injured onlooker could sue the police for assault. The London police use no riot squads, no protective clothing, no arms, no riot gas, and no barriers for crowd control. The police line has always to be in close contact with the crowd, preventing the stone throwing and violence that occurs when there is a gap between them and the citizenry. The advantages of the passive police stance encouraged by their legal accountability were exemplified by four

London demonstrations in 1967–68 against U.S. policy in Vietnam. At the fourth of these, in an atmosphere made tense by press apprehensions of revolution, some 50,000 people marched through the city. In February 1971 a demonstration of some 140,000 individual workers passed without an arrest.

Disturbances that erupted in more than 50 American cities in 1967 posed serious problems for the police. Most of these disturbances began during the evening watch, when a police department normally has about 13 percent of its uniformed force on duty. At this time, a city with a population of about 500,000, for example, would have had about 95 uniformed police officers on duty. The National Advisory Commission on Civil Disorders found that many of the disturbances began when allegedly abusive or discriminatory police actions set sparks among the grievances of blacks. In some cases the police were unable to control the resulting situation, and the National Guard, composed of private citizens who undertake part-time military duty, was summoned. The commission concluded that the National Guardsmen were inadequately trained for riot duties and that their performance in 1967 raised doubts about their capabilities for such assignments. Whereas in some cities National Guardsmen and police fired thousands of rounds of ammunition, in others regular troops restored order by calm discipline and little use of firearms. The experience of the 1967 disorders suggests that giving the police more destructive weapons is not sufficient to maintain order when a substantial section of the population is disaffected.

POLICE WORK AS AN OCCUPATION

As has been seen, some police bodies are constituted as part of the armed services. In some European countries it has been normal to recruit as police those who have served as regular soldiers or sailors. It has been even more common to appoint retired army, navy, or air force officers to command police forces. Since World War II the trend has been in the opposite direction.

Police work is seen as a specialist career, and it is argued that the senior positions should go to those who have learned the work by experience. This poses certain problems of recruitment. It is argued that the police should be able to offer appointments as cadets to young persons who want to become policemen. If they have to wait until they are 21 years or so before they are eligible for appointment, they may be lost to the police. On the other hand, it is said that a youth who has come straight into the police from secondary school has had too little experience of life and is likely to be too narrow in his outlook. Cadet training schemes therefore seek to give such entrants a wider experience by sending them to social-service agencies and other bodies.

Officer entry

A similar problem is that of officer entry. The French recruit graduates to the Police Judiciaire by examination, and in Norway only those with legal qualifications are eligible for appointment as police chiefs. In the United States many universities offer courses in police science, and those who obtain qualifications in this field have a certain but limited career advantage. In Britain, graduates are welcomed but have to compete with other entrants for promotion. The most able among the younger officers are selected for a 12-month course at the national Police College at Bramshill; they divide their time so that half is devoted to professional and half to general studies, by contrast with the rather shorter courses at the Federal Bureau of Investigation Academy at Quantico, Va., where all the teaching concentrates on police subjects.

INTERNATIONAL ORGANIZATION

The International Criminal Police Organization (Interpol) has a history going back to 1914, though its present constitution dates from 1956. Each member nation has established a bureau that maintains relations with the General Secretariat in Paris. The bureaus transmit criminal information that may be of interest to other countries; they undertake, within their own countries, inquiries, searches, and arrests requested by other countries; and they take steps to implement resolutions voted by the annual assembly. Interpol can act only within the framework of national laws; criminals can be returned only if an extradition treaty is in force and the offender is a national of the country requesting return. The International Association of Chiefs of Police, with headquarters near Washington, D.C., draws its members largely from the United States and is the leading voice in the United States for professional police standards. It is active in training, research, and public relations. The International Police Association was founded in Britain in 1950 as a social organization. Although it is most active in Europe, its members come from dozens of nations worldwide. The association grants scholarships for study travel and arranges annual conferences. (M.P.B./Ed.)

Police technology

Police technology covers the methods and techniques of, and the equipment available to, police agencies. Throughout the 20th century, improvements and innovations have been made in police transportation, communications, weaponry, laboratory facilities, and other areas, but these have scarcely more than kept pace with the pressures generated by denser urban populations, greater sophistication on the part of criminals, social disturbances, and new types of crime made possible by economic and technological developments—*e.g.*, aircraft hijacking and computer fraud.

Police technology increasingly involves borrowings from the physical and social sciences, as well as from the fields of jurisprudence and education. Although it was only in the 1960s that police technology became recognized as a distinct academic and scientific discipline, the proliferation of professional publications, educational programs, workshops, and international conferences of police reveals the growing importance attached to the technical aspects of police work.

Two influences serve to restrict the application of technology to police work. First, technical problems themselves may arise when the technology of one discipline is transferred to a different application in police science. Second, in the United States and other democracies, constitutional and legal issues may limit or prevent the use of feasible technologies in police work.

Many examples of an incipient police technology can be found in the ancient and medieval world. Long before the beginning of the Christian Era, Egyptians used detailed word descriptions of individuals, a concept known in modern times as "portrait parle." The Babylonians pressed fingerprints into clay to identify the author of cuneiform writings and to protect against forgery. Fingerprints were also used by the Chinese at an early date for purposes of identification. The concept of comparison and identification of handwriting was established in Roman times, as is evidenced by the Code of Justinian of the Eastern Roman Empire, first enacted in AD 529. Branding and mutilation were used from time immemorial to identify convicted criminals. Trial by ordeal and trial by combat were medieval methods of interrogation of suspects, with the survivors judged innocent. A more humane medieval method, and a step toward modern concepts, was compurgation, in which the friends and families of the disputant took oaths not on the facts but on the disputant's character.

Precursors of the modern lie detector

An Asian interrogation method involved scientific principles based on the lessening of salivation under nervous tension. The mouths of several suspects were filled with dry rice; the suspect exhibiting the greatest difficulty in spitting out the rice was judged guilty. Another precursor of the modern lie detector was employed in India. Suspects were sent into a dark room where a sacred ass was stabled and directed to pull the animal's tail. They were warned that if the ass brayed it was a sign of guilt. The ass's tail had been dusted with black powder; those with a clear conscience pulled the tail, whereas the guilty person did not, and an examination of the suspects' hands revealed the person with the guilty conscience.

Modern police methods were beginning to develop in Europe by the 18th century and in the 19th century were reinforced by a great number of technological advances.

When the British Parliament enacted the reforms of Robert Peel in 1829, the importance of keeping comprehensive police records was recognized. In principle, Peel stated that police records were indispensable for the proper distribution of police strength. His reforms laid the foundation for the development of the modern police communications system and provided for a coherent police organization.

Until the mid-19th century, however, police signal systems consisted of whistles, raps on the pavement with a nightstick or baton, and flashing of lights. In 1849 the central police station at Scotland Yard and the district police stations of London were connected by a telegraph line. Other cities followed, and telegraph boxes placed at points on patrolmen's beats permitted swift calls to headquarters when assistance was needed. The telephone was first installed in a police station in Washington, D.C., in 1878. Two years later Chicago took the important step of installing telephone boxes on patrolmen's beats.

Important advances in identification occurred in the late 19th century. In 1882, Alphonse Bertillon of France introduced a means of identifying suspects by using measurements of the body, head, and limbs. This system was largely supplanted by modern fingerprinting, which dates from about the same period. The Henry system of classifying fingerprints (by patterns and shapes) was adopted by Scotland Yard in 1901, and this system, or variants of it, soon became standard throughout the world.

The idea of tracing and identifying an individual typewriter by means of the peculiarities of its type characters probably first appeared in Arthur Conan Doyle's story "A Case of Identity," written in 1891, three years before the first mention of the subject in technical literature.

First use of the automobile

The automobile, first used in police work in Akron, Ohio, in 1899, became the basis of police transportation in the 20th century. Closely linked with it was the adoption of radio communication. Other innovations in communications were the teletypewriter and facsimile. The teletypewriter evolved from printing telegraphs and developed rapidly during World War I; by the 1920s it was used in police communications in most countries. Facsimile transmission (telephotography) dates from the early 1900s, and in 1908 a criminal was captured as a result of his photo being transmitted by facsimile from Paris to London. After World War I the use of this technique became widespread.

Perhaps the single most important invention in the development of police communications was the transistor, which, in the 1950s, led to the development of small, lightweight, reliable radio transmitter-receivers.

In the 1920s the technique of identifying firearms by the striated markings on bullet jackets was introduced. As these and other new techniques became available, the first police crime laboratories were instituted in Europe and the United States. One of the earliest was established in Lyon, Fr.

There have been few innovations in weaponry. The nightstick (club) and firearm, invariably adaptations of military weapons, remained virtually unchallenged in the police arsenal until the 1960s, when chemicals, tranquillizing darts, light- and sound-energy techniques, and other devices began appearing.

MODERN SURVEILLANCE SYSTEMS

Police use a number of surveillance techniques, but their chief reliance is almost universally placed on patrolling.

Patrol systems. Patrols vary in methodology, assignments, structure, and mobility. Their flexibility has been significantly enhanced by improved methods of communication, especially the two-way radio, and by the availability of numerous types of vehicles such as the automobile, the motorcycle, and, in many countries, the bicycle. The mobility provided by modern communication and transportation techniques is particularly useful in dealing with civil disorders.

The areas covered by metropolitan police patrols vary enormously in population density, character of commerce and industry, terrain, and roads. The initial enthusiasm for using helicopters in routine patrol work lessened when cost analyses showed that the true cost of helicopters was much higher than expected. High maintenance costs and excessive downtime because of a lack of pilots, aircraft disrepair, and poor flying weather resulted in helicopters being rarely used in routine patrols. Helicopters have come to be used primarily for rescue work and special tactical situations requiring surprise or high mobility. In some countries short takeoff and landing aircraft have been adopted for patrol and rescue work.

Police dogs. The training of dogs for police work was originally developed in Ghent, Belg., about 1900 and was soon copied elsewhere. Though certain breeds with especially keen senses are used for special purposes, such as detecting caches of illegal drugs or for tracking fugitives and missing persons, the most widely trained dog for regular patrol work is the German shepherd, or Alsatian. Other breeds that are sometimes used include boxers, Doberman pinschers, Airedale terriers, rottweilers, schnauzers, and bloodhounds. Selection involves the use of standard criteria related to breeding and physical measurements, and training is rigorous.

Electrical protection systems. Electrical protection (alarm) systems substantially reduce burglaries and contribute to the capture of many criminals. The development of alarm equipment and the evolution of operating procedures in the United States, Canada, and western Europe have been significantly influenced by the published standards of Underwriters Laboratories (UL). Central-station systems and local-alarm systems are the two basic categories of UL-regulated burglar-alarm systems.

Central stations

Central stations are commercially operated control centres serving metropolitan and surrounding areas. Burglar-alarm protective circuits and devices automatically transmit signals to these centres. Armed guards are on duty at all times and are dispatched to conduct immediate investigations of unauthorized entry of protected properties. Central stations generally have direct communication channels to police control centres. Some centralized operations in large European cities control several completely automatic satellite stations in nearby cities; the local police are dispatched by the central station.

Local burglar-alarm systems consist of sensitive warning mechanisms that are connected to loud alarms and, often, to a signal system in local police stations. Originally local burglar-alarm systems were used primarily in small towns where central station services were not available, but technological improvements and lower prices have led to their use in homes and small businesses in all types of cities. Many police departments discourage the use of these systems, however, because up to 95 percent of the alarms given by them are false.

Technical surveillance. The application of technology to surveillance operations has greatly expanded the capabilities of law-enforcement agencies and has led to the creation of a special field of police technology known as technical surveillance.

Audial surveillance. Audial surveillance, or electronic eavesdropping, became practical for obtaining evidence and investigative leads as a result of the intensive development of the magnetic tape recorder and solid-state electronic components during World War II and subsequent years. The field has two technically and legally distinct areas: the surreptitious interception of oral communications, commonly known as "bugging," and the surreptitious interception of wire communications, commonly known as "wiretapping." Typical devices used include miniature microphones, miniature radio transmitters, and a variety of radio-receiving and voice-recording equipment.

Microphones, of course, are always required for the interception of oral communications and may be connected by cables directly to recorders located at remote locations or may be connected to transmitters. Excellent quality microphones of very small size are available throughout the world. Some microphones located in the handsets of telephones can be modified to intercept nearby conversations.

Radio transmitters the size of a cigarette package, equipped with integral batteries and microphones, are commonly known as "wireless mikes." Most radio transmitters used for audio surveillance operate within the very high frequency (VHF) range.

The most common type of interception of wire communications consists of monitoring telephone conversations. Equipment used for telephone tapping ranges from the telephone repairman's handset to automatic voice recorders connected to target telephone lines.

Optical surveillance. Optical surveillance makes use of such devices as binoculars and telescopes; cameras with long-focal-length, or telephoto, lenses; closed-circuit television; and video tape recorders. Closed-circuit television is a versatile surveillance method that has become widely used both by police and by private security systems. From a central location, a supervisor can monitor events at one or several remote locations. Cameras may be fixed or they may oscillate to sweep a large area.

Optical surveillance is sometimes used as an evidence-gathering technique. In arrests involving driving and the use of alcohol, the physical condition of the driver is filmed or videotaped and later used as evidence. In an effort to reduce lawsuits concerning interrogation methods, some police agencies also film or videotape interrogation.

Devices for low light levels

Low-light-level devices have greatly extended passive optical-surveillance capabilities. Practical low-light-level television systems are manufactured for optical surveillance operations under light intensities as low as 10^{-8} lumens per square metre, or virtual darkness. Night-vision telescopes using two and three light-amplification stages are also used.

Tagging and tracking. Object tagging and signal tracking are a third category of police technical-surveillance operations. Devices emitting a unique signal are attached to vehicles or implanted in targeted objects. Compatible receiving or activating equipment is used to track the tagged objects. A typical device used in tracking is a miniature radio transmitter using distinctive tone modulation, commonly known as a "bumper beeper," which is surreptitiously attached to a vehicle. The target is discreetly tracked with one or more surveillance vehicles equipped with appropriate receiving equipment.

CRIME REPORTING AND RECORDS

The efficiency and effectiveness of a police department depend in large degree on complete, readily available, and up-to-date record keeping.

Information must be reported, collected, classified, analyzed, processed, and stored. Numerous files must be kept, of which the most important include fingerprints, modus operandi of criminals, and various investigation and reference files. There should be as much cross-filing as possible.

Fingerprint classification. Fingerprint records are among the most significant files that are maintained. Inked fingerprints from all 10 fingers of an individual person can be classified; *i.e.,* be put into narrow categories for file purposes and ultimate retrieval. Most countries use the Henry extension system, a modification of the Henry system, which was developed prior to 1900 in England. In the Henry extension system, each of the 10 individual fingerprints of a person is classified as to arches, loops, whorls, or composite style prints. The type of design, plus subclassification based on locating certain fixed points and counting the ridges lying between the points, is used to make the classification.

Fingerprints obtained from individuals are classified and compared with the fingerprints of other individuals having the same classification. With this system, personnel in a police centre need not compare a set of fingerprints with the millions of cards on file but only with the few in the same classification. (For the technique of fingerprint identification, see below *Laboratory procedures.*)

Modus operandi. If a criminal has been successful using certain methods and techniques, he tends to repeat the same procedure. Thus, evidence of characteristic behaviour and procedure at the crime scene can serve to identify the offender if the pattern is recognized by the police. Although classification of criminals according to operational methods and manners has lost a good deal of its effectiveness in the investigation of burglaries, large law-enforcement agencies still maintain modus operandi files to recognize patterns of behaviour, to associate a group of crimes with a single perpetrator, and to enable them to predict the next target of the criminal. The modus operandi file is most effective in crimes involving personal contact, such as felonies against persons, confidence games, and forgery. Enhanced computer storage and search capabilities have rekindled interest in modus operandi methods of investigation, even in such cases as burglaries and robberies.

Investigation and reference files. Informational files are maintained by law-enforcement agencies. The primary purpose of these records is to identify the source of evidence discovered in connection with a crime. For example, a file of automotive paint samples can assist in identifying the make and model of a vehicle involved in a hit-and-run case. Automotive registrations and records of firearms assist in establishing ownership. In addition to their own files, the police have access to the records of many governmental agencies that record titles, contracts, vital statistics, significant events, and official acts in general. Among other files maintained by the police the following prove exceptionally useful.

Missing persons. Each year many thousands of persons are reported missing to the police. In addition, the police must investigate cases of unidentified persons, living and dead. Descriptive files are maintained of missing persons under the age of approximately 19, unidentified dead, and unidentified persons who are physically and mentally afflicted.

Laundry and dry cleaners' marks. A valuable clue to the identity of an unknown person can be found in the mark or marked tag placed in his clothing by a laundry, dry cleaner, or tailor. The mark usually consists of a set of numbers or letters identifying the retail store or the wholesale processor. Samples of these marks are filed systematically by larger police departments for the purpose of tracing a lost, dead, or wanted person.

Lost or stolen property. Many cases of burglary and larceny are solved simply by tracing stolen property to the possession of the criminal. For this purpose, police maintain a central cross-indexed file of property lost or stolen and of articles pledged at pawnbrokers or sold to second-hand dealers. The common objects of larceny—automobiles, watches, jewelry, furs, and cameras—are identified by means of an accurate physical description, including serial marks and personal inscriptions.

Computerized records. The record-keeping tasks of police work were long handled by using a diversity of manual systems. Quick nationwide or international utilization of records and files was impossible. Now, however, electronic data processing has revolutionized information-handling techniques. Manual filing systems have been superseded by computers that operate from central geographic points, storing data and making it instantaneously and accurately available to many police constituents. Advances in computer networking permit the linking of many police and investigative data bases and make possible the development of national criminal investigation systems. Interpol is an example of an international computer network that coordinates the exchange of criminal intelligence to aid police agencies in searching for fugitives or missing persons or property.

Electronic data processing

In addition to serving as a depository and dispenser of information, the computer can provide statistical analysis for research and management, develop intelligence data bases and analysis systems, assist in the analysis of physical evidence, facilitate swift document retrieval, and help in such areas as identification of fingerprints. Using traditional manual methods, fingerprint examiners, for example, have to make visual comparisons of thousands of individual 10-digit fingerprint cards in order to make a positive identification. The development of a computerized fingerprint search system took more than 20 years of industry-wide research. Although the Federal Bureau of Investigation was a pioneer in the development of automated fingerprint files, the vast size of its collection of fingerprint cards and the high cost of converting its files delayed the completion of a high-speed fingerprint search system. The Japanese National Police Agency established the first practical system in Tokyo in the early 1980s.

A second generation of the Japanese system focused on

individual images instead of 10-digit cards. Since its introduction, this system has been used as a prototype for several computerized systems in the United States, including those in police agencies in Alaska, Massachusetts, San Francisco, and Suffolk County, N.Y. An interactive system established by the California Bureau of Criminal Identification is one of the largest, linking several major police departments and covering more than 90 percent of the population of the state. New Scotland Yard in London also has a computerized fingerprint system that makes some 1,000,000 comparisons a day.

COMMUNICATIONS

A total police communications system can be extremely complex, but its complexity varies with the needs associated with particular police organizations and their operational environmental situations. Each police function or situation calls for its own subsystem of communication.

In a large city the sources of information that must be fed into the central system are remarkably varied. Police may be on foot or motorcycles, in patrol cars, in patrol boats, or in aircraft. In addition, the system must encompass links to fire stations, ambulance services, traffic-control systems, rescue teams, alarm systems, emergency utility services, crime-information systems, and crime laboratories. A variety of fixed and mobile systems are used, including radios, television (including closed-circuit), telephone, teletypewriters, and facsimile transmission.

Police radio-communications systems, which profited considerably from the advances in electronics during and after World War II and from developments in the U.S. space program, also benefited from the development of computers, which made possible the quick retrieval of information on stolen property, wanted persons, and other police intelligence. Many urban centres now use police cars that have direct access to departmental computers and contain equipment capable of receiving printouts of orders and lists of wanted persons or missing property. The augmented flow of information has placed an added burden on existing communications systems, which have consequently grown more complex. Among numerous technological innovations, the development of the modern police communications centre merits particular mention.

Functions of a police communications centre

A communications centre performs several functions: (1) receives and processes police calls from the public; (2) audits and dispatches police resources in response to police calls; (3) receives and responds to information inquiries as well as to calls for assistance from police patrols; (4) maintains moment-to-moment information on police resources for command and control purposes.

The geographic distribution of police communications centres has been broadened to include locations in many countries of the world. These centres facilitate police work needed to deal with the international mobility of modern criminals. The extensive radio network maintained through Interpol serves to promote international cooperation among world police organizations.

INTERROGATION METHODS AND EQUIPMENT

Police and other investigators depend on interrogation as a principal means of determining facts and resolving issues. Reliance on interrogation, however, involves certain problems: ascertaining when a suspect or witness is telling the truth, evaluating memory, allowing for the physical and mental condition of a witness or suspect, and understanding the problems created by an individual's perspective.

Interrogation methods and equipment have evolved in response to these problem areas. The psychological, psychophysical, and physical sciences have played vital roles in police interrogation techniques.

The most dramatic gains in interrogation technology have come through the polygraph, or so-called lie detector. The polygraph monitors and records selected body changes that are affected by a person's emotional condition. The recorded changes are then studied, analyzed, and correlated in respect to specific questions or other stimuli.

The first modern polygraph was constructed in 1921 by John A. Larson, a medical student at the University of California, working with a member of the local police department. Larson's instrument was capable of continuously recording blood pressure, pulse, and respiration; since it recorded these conditions simultaneously, it was called a polygraph. A later development provided for the recording of the psychogalvanic skin reflex (electrodermal response), the flow of current between two different parts of the body. Research has continued on both the instrumentation and the psychological techniques necessary for its effective use by the police.

Even though the polygraph has been formally and successfully used in police intelligence and security investigation since 1924, there is still no complete agreement by psychologists on its validity. Furthermore, the results of a polygraph test are not always judicially acceptable. Despite these qualifications, the polygraph has proved an invaluable aid to police interrogation and the development of investigative leads.

WEAPONS AND RELATED DEVICES

To contend with the wide diversity of criminal situations, police require a variety of weapons and restraining and protective devices.

Standard weapons. Standard equipment varies. London police carry only a small baton; New York City police carry a large-calibre revolver and a baton; Paris police carry a small-calibre semiautomatic pistol, a baton, and a leaded cape; and Cairo police carry a submachine gun or automatic rifle.

The most universal physical weapon and certainly the most frequently used weapon over the years has been the baton, nightstick, truncheon, or billy. The baton has been the subject of study and research, and there now exists a family of batons that include, in addition to the original wood or plastic baton, devices that contain firearms, aerosols or tear-gas cartridges, and electric-shock systems. Some batons are made to break at a certain degree of impact in order to avoid serious injury.

Pistols, including the automatic pistol and the revolver, are the second most widely used police weapons. In the United States the revolver (.38 special to .45 calibre) is the primary pistol, whereas in Europe and Asia the automatic (.25 to .38 calibre 9 mm) is extensively used. Shotguns, rifles, and automatic weapons are usually included in the police arsenal but are issued only when essential.

Lacrimators and nauseators

Chemical and other nonlethal weapons. Chemical weapons include those that deliver compounds which cause either severe weeping or tearing at the eyes (lacrimators) or extreme discomfort or sickness (nauseators). The three most commonly used chemicals are known by the initials CN and CS, both lacrimators, and DM, a nauseator. Among the systems used to dispense the chemicals are riot shotguns, fog generators, and grenades. The most commonly used modern weapon is the handheld aerosol tear-gas projector (Mace). It can be carried in a holster, and it is easily aimed and instantly effective. Its range is 15 feet, and it delivers about 40 one-second bursts.

Other nonlethal weapons include sound and light devices that cause confusion, pain, or temporary blindness, as well as stun guns. Stun guns emit a high-voltage charge that temporarily renders a person unconscious. (Because some physicians object to its use, the stun gun must undergo widespread testing and evaluation before it is added to the standard police arsenal.) The water cannon is used, especially in Europe, for crowd control. Nonlethal projectiles include plastic shot and bullets made of rubber, wood, or fabric.

Restraining and protective devices. Criminal restraining and police protective systems have been subjects of much study and research. The developments in restraining devices and systems include lighter, more effective handcuffs with improved locking systems, lightweight thumb cuffs, pressure-twisting devices that are applied to wrists, and lighter and stronger straitjackets and leg irons.

A modern restraining device is the disposable handcuff. It consists of a nylon plastic strap with a slotted head at one end and a tapered tail at the other. A stainless steel barb in the head allows the tail to be pulled through the head in one direction and locks the strap so that it cannot be withdrawn. To remove plastic cuffs, the police officer

merely cuts them off. This device is particularly useful in mass arrests, as in riots.

Significant developments in the technology of protective devices include heavy plastic shields to ward off rocks and other missiles, lightweight nylon clothing resistant to pistol bullets, bulletproof shields, and bulletproof armour made of nylon and ceramic materials, which can withstand high-powered rifle bullets. Studies show that in their first five years of use in the United States, bulletproof vests saved the lives of more than 700 police personnel. Of all the police killed by guns during that period, only 10 percent were wearing vests, and virtually all of that number suffered head wounds.

LABORATORY PROCEDURES

Crime laboratories may be either specialized, such as those that handle principally documentary evidence, or broad in coverage, as is the case with most national police laboratories. The fundamental purposes include the collection and preservation of physical evidence, its examination, and the preparation and presentation of accurate, objective testimony regarding it.

Personnel include specialists such as document examiners, generalists such as criminologists, and technicians and other supporting personnel. They are usually employed by a government unit and must pass rigid examinations.

Facilities and equipment. Crime laboratories usually are equipped with an array of scientific instruments and facilities. Instruments are constantly being developed, and the resources of individual laboratories vary. Some are equipped with expensive and sophisticated analytical instruments, such as the mass spectrograph, the electron microprobe, and the scanning electron microscope. Close cooperation among laboratories makes possible the sharing of specialized equipment when needed.

Value of photography

The largest category of laboratory equipment is related to photography. The usefulness of the medium extends from straightforward photography of, for example, automobile accidents and sites of crimes to specialized work of various kinds. Colour photography is commonly used to obtain more realistic and informative representations of the crime scene. Ultraviolet and infrared photography aid in detecting stains, invisible writing, and alterations on documents. Radiography is also used to provide details that would otherwise be invisible. Photography through a microscope (photomicrography) is a common technique, a special case being the examination of bullets. Fingerprint photography is an area of specialization that may involve conventional methods or any of a number of special techniques to recover useful prints from different types of surfaces.

Among the more common facilities are chemical laboratories, with a range of reagents and analytical instruments, including spectrographs and chromatographs. Optical equipment includes a full range of magnifying devices, from magnifying glasses to specialized microscopes of various kinds, including stereomicroscopes and comparison microscopes.

Evidence and its examination. An investigation at the scene of a crime, particularly in crimes of violence, is the first phase of the laboratory function. Many jurisdictions have mobile crime laboratories, in which some of the regular laboratory tests can be performed by specially trained police personnel called evidence technicians.

Crime sites must be searched to locate pertinent physical evidence, which may range from latent fingerprints to bloodstains to pieces of a broken automobile headlight. Evidence must be marked for identification, preserved, and protectively packaged for transportation to the laboratory. In addition, records (written or photographic) must be made of each piece of evidence and its exact location with respect to the crime. The law requires that the "chain of evidence" remain unbroken and makes police accountable for every item of evidence from the time of its discovery to its ultimate presentation in court.

Certain laboratory procedures and tests are carried out in the field when time is limited or when there is a possibility that the evidence might be disturbed. Among such procedures or tests are dusting for fingerprints, making casts of footprints and automobile tracks, making blood tests, and collecting specimens of organic and inorganic materials from the environment.

Hypnosis. In the 1970s police began experimenting with hypnosis as a means of interviewing witnesses to certain crimes. Hypnosis is reported to be particularly effective in helping witnesses give descriptions of criminals, relate the details of violent sexual attacks, and recall the scenarios immediately preceding certain accidents.

Drug and alcohol testing. Portable field chemical kits have been developed to help police detect narcotics and dangerous drugs. These kits are designed to provide a practical method for initial tests in the field and are not intended to replace the more thorough laboratory examinations.

Tests for determining if a suspect has taken a narcotic or dangerous drug must be performed in a forensic laboratory. These tests require the taking of a urine or blood sample from the suspect and require close medical supervision. Many new techniques for the detection of drugs, including specially trained police dogs, are used.

The breath analyzer

For many years the basic test for legally determining a driver's blood-alcohol content involved taking a blood sample and sending it to a laboratory for analysis. This method, although accurate, is not practical for law-enforcement purposes because many people object to having a blood sample taken and because test results are not immediately available. The development of the breath analyzer for the determination of blood-alcohol content has given the police a more practical tool. This instrument permits a precise amount of the suspect's breath to pass through a chemical solution (potassium dichromate and sulfuric acid), resulting in a colour density change that is proportional to the amount of alcohol in the air sample. The percentage of alcohol in a given air sample is determined, and from this the percentage of blood-alcohol content can be derived. In cases where the suspect is not conscious or requires serious medical attention, a blood test may be the only recourse.

Classes of evidence. Evidence can be categorized according to its potential value as proof. Certain types of evidence can be individualized, or associated with a unique source, whereas other types cannot be pinpointed but can aid in identification because they fall into particular classes. Types of evidence that can be individualized are handwriting, typewriting, fingerprints, footprints, tire marks, impressions or casts of nonstandard items, striated markings on bullets, tool marks, objects with random fractures or tears, and substances that have undergone an alteration that makes them unique among others of the same class. Among evidence that only identifies a certain class are blood and other body fluids, narcotics, toxicological materials, fibres, soils, inks, and many kinds of materials produced in batches.

Examinations can be either "unilateral" or comparative in nature. Blood-typing and establishing the distance from which a weapon was fired are unilateral examinations, whereas studies of handwriting or spent bullets involve comparisons with other examples.

Pathological examinations. Although pathological examinations are not ordinarily conducted by a crime laboratory, they are part of forensic science. Definitive statements relating to the cause of death and time of death are made by the coroner or medical examiner of the jurisdiction in which the death took place. Observations by crime laboratory personnel of the conditions at a crime scene, plus toxicological analysis by a crime laboratory, are used in concluding whether a death was by accident, suicide, or murder. In most jurisdictions an autopsy is conducted if a physician was not in attendance or if the cause of death is not obvious, as in cases involving poisoning, electrocution, asphyxiation, heart attack, or stroke. In cases involving identification of a decedent, crime laboratory personnel may work with a coroner or medical examiner to establish identity by comparison of fingerprints or other methods.

Fingerprints, palm prints, and footprints. Fingerprints found at the scene of a crime can be evidence connecting an individual with a crime. Fingerprints can be either visible or latent. Visible prints—formed by dirt or blood, for

example—or three-dimensional prints formed in soft matrices, can be photographed directly. Latent fingerprints, which are not ordinarily visible, can be brought out by dusting techniques when the surface is hard and by chemical techniques when the surface is porous. In dusting for fingerprints, a fine powder of contrasting colour is applied with a fine brush. The powder clings to the residual oils and fats in the print and the excess powder is removed with the brush. On porous surfaces such as paper, fuming iodine, silver nitrate, or ninhydrin solutions are used to develop the latent fingerprints. The most effective developer of latent fingerprints is ninhydrin, which can reveal prints that are several years old.

Fingerprint "points"

Fingerprints are identified on the basis of agreements in a significant number of individualities, commonly known as "points." These are the bifurcations, ending ridges, and dots in the fingerprint pattern. If sufficient points are found with spatial relationship to other points, a basis exists for identifying a fingerprint. It was formerly considered necessary to have 12 points to identify a fingerprint, but in current practice, a lesser number is often used. Palm prints and footprints are identified in the same manner as fingerprints.

Document examination. The work of the "questioned document examiner" encompasses the examination of handwriting, typewriting, ink, paper, and typography and covers such problems as the identification of handwriting and typewriting, the sequence of events involved in the preparation and handling of a document, the alteration of a document, and the age of a document.

Handwriting, like all other evidence, is identified on the basis of individuality. Children are taught specific ways of forming letters, but they rapidly begin to inject their own individuality into letter forms and writing execution. By the time a person reaches maturity, his writing has acquired peculiarities that serve to identify him.

A forged signature presents other problems. Simulated signatures based upon recollection contain a combination of the forger's own writing habits and his recollection of the victim's habits. In many cases, such simulations can be identified. On the other hand, when the perpetrator makes a careful drawing of the victim's signature or traces an authentic signature, the forgery can be exposed but cannot be identified with the handwriting of the perpetrator. Two individuals making careful tracings of the same signature can produce virtually identical drawings.

Typewriter identification

The make and model of typewriter used to prepare a document can be determined by examining typeface designs and comparing them with information contained in reference files and differentiation charts. The work of a specific typewriter can be identified if it contains sufficient identifying individuality, based on broken or tilting characters, badly aligned characters, characters that print more heavily on one side than on another, and rebounding characters. Mechanical conditions affecting the type basket or the escapement may also have identifying value. Since it is impossible to reinsert a document into a typewriter in its exact former position, it is possible, using a ruled screen, to determine if a document has been prepared at one sitting or has been reinserted.

Ink comparisons provide class evidence that is frequently of value, but it is not possible to individualize ink; *i.e.*, identify it as coming from a unique source. Chemical tests of various kinds are used for ink comparisons.

Papers can be differentiated on the basis of fibre, filler, and sizing constituents. Fibres can be identified by differential staining and microscopic examination. Fillers can be identified by X-ray diffraction because they are crystalline substances. Chemical tests are used for the identification of sizing constituents. Research with neutron activation analysis, a technique involving chemical analysis by radioactivity, indicates that it may even be possible to identify paper by batches.

Questions concerning the sequence of writing—*e.g.*, whether the ink signature is over or under the typewritten portion of a document—can be important in proving the authenticity of a document. In this and related problems, such as deciphering and restoration of eradicated or erased writing, examinations are carried out using stereomicroscopes, ultraviolet and infrared techniques, or chemical tests, depending on the specific problem.

Ballistics techniques. Firearms examinations involve such problems as ascertaining the type of weapon and the specific weapon used to fire a bullet, determining the distance between the weapon and the target, and establishing whether or not a particular individual fired a weapon.

The identification of a particular weapon is based upon the comparison of minute striations, or striae, left on bullets as they pass through the rifled barrel. Each barrel leaves a different combination of striae on the side of bullets. A bullet recovered from the scene of a crime can be compared with a test bullet fired from a suspect weapon. The striated markings on both bullets are brought into juxtaposition under a comparison microscope. If there is agreement in the width, depth, and spacing of a significant number of striae in both bullets, an identification can be made. (The same principle is used in the identification of tool marks.) The extractor and ejector markings on rifle and automatic-pistol cartridges can be identified in the same manner. In some cases the firing pins and breechblocks leave individual markings on the face of a cartridge.

The distance from which a weapon was fired can be determined, up to several feet, by examination of the form and size of the hole in a target, the extent of the burned area around the hole, the amount of embedded powder grains, and the presence of mercury or lead from primers. The effects vary with the weapon and the type of ammunition. Such information can be vital in determining whether or not a gunshot wound was self-inflicted. Infrared, X-ray, and chemical tests are used in determining the extent of powder residues.

The best means of determining whether or not an individual has recently fired a gun has been the subject of much controversy. In the dermal nitrate test, melted paraffin is used to pick up traces of nitrates and nitrites on the hand, after which a chemical test is made for such compounds. This method has been discredited because it can give incorrectly positive test results. Use of the neutron-activation technique to detect the presence on the hand of antimony and barium from cartridge primers is much more definitive.

Individualization of paint

Paint. Under certain circumstances, paint evidence can be individualized. If a jagged edge of a paint fleck found on a suspect can be matched with the edges of the matrix from which it came, a basis exists to connect a suspect with a crime scene. Multilayered paint chips can sometimes be individualized if the discrete layers possess sufficient individuality. Under other circumstances, paint must be considered class evidence. Examination of paint for class characteristics involves the use of chemical tests to determine constituency, as well as examination with instruments such as the infrared spectrophotometer, emission spectrograph, or electron microprobe.

Inorganic materials. Soil, safe insulation, and other inorganic evidence can also be compared to determine common origin. It is not possible to individualize the evidence, but the fact that the same type of mineral matter was found on a suspect and at a crime scene is usually considered significant. Inorganic materials can be compared in terms of metallic element constituents, density of constituents, and microscopic structure.

Blood, blood spots, and semen. Bloodstains are a natural consequence of many crimes of violence; examinations usually concern the position, extent, and nature of suspected bloodstains. In most cases, the blood is dry. The Teichmann and Takayama tests, both of which rely on a reaction with hemoglobin, are confirmatory tests for blood. Human blood can be differentiated from the blood of other species by the precipitin test, which involves the reaction between blood and antihuman rabbit serum.

Forensic scientists have worked to individualize blood samples. For many years blood analysis was limited to simple groupings (A, B, O, or AB) and a few traits like Rh, Lewis, and Kell factors. Advances in serology now permit the isolation and identification of additional blood factors, including histocompatibility antigens, blood enzymes, and serum proteins. These advances also point to the possibility of individualizing other body fluids.

The shape, position, and extent of bloodstains often give valuable information about the circumstances surrounding the commission of a crime. Blood falling from different heights forms stains of different shapes, which are further modified by the angle at which the blood drop hits the surface. Rough estimates can be made as to the age of bloodstains.

Ultraviolet light and the acid phosphatase test are used for the presumptive identification of semen. Positive identification for semen relies on identifying spermatozoa in the suspected sample.

Hairs and fibres. Hairs and fibres found at the scene of a crime can also be important evidence. It is not possible to trace hair to a specific individual, but narrow classes having probative value can be established. Human hair can be distinguished microscopically from other animal hairs. Hairs from different portions of the body vary in structure. Microscopic examination can reveal if hair has been recently cut, burned, or split. Chemical tests can demonstrate the presence of bleached or tinted hair. Neutron activation analysis has come to be used to identify specific trace elements in hair and to measure the quantities of such trace elements.

Human hair

The various fibres used in clothing, rope, and furniture are of vegetable, animal, or synthetic origin. They can be differentiated by microscopic study and by simple chemical tests for solubility and other properties. Colour comparisons can be conducted using a comparison microscope or by spectrophotometric evaluation. Because they are mass-produced, fibres cannot be individualized.

Toxicology. Toxicological examinations concern the identification of drugs and poisons, either in the form of pills and powders or as materials recovered from body organs. Among the drugs and narcotics of interest to crime laboratories are marijuana, morphine, heroin, and other narcotic alkaloids, cocaine, amphetamines, barbiturates, tranquillizers, and LSD. These substances must be differentiated from nonprescription drugs. In the case of pills, the exact shape, colour, and external marking may serve as aids in identification. Laboratory test methods for specific chemical substances include colour reaction tests, crystal tests, chromatography, and infrared spectrophotometry. Normally, several tests are performed for the sake of confirmation.

Examination of body organs for the presence of poisonous material requires an involved extraction procedure with a series of tests to confirm or eliminate the presence of toxic materials. In addition to the drugs previously mentioned, other poisonous substances include gaseous and liquid poisons, metallic and inorganic poisons, organic poisons, vegetable and animal poisons, and bacterial poisons.

Voice identification. The use of the sound spectrograph for voice identification is a relatively new development. The sound spectrograph is an instrument that graphically presents the time, frequency, and intensity of speech sound waves. The graphic forms for a sound, as spoken by one person, can be compared with those produced by the speech of a second person and thus differentiated. The accuracy of the technique in identifying individuals remains in doubt among speech scientists, even though voicegraphs, or voiceprints, have been used in court.

(W.B.McM./G.E.Mi.)

BIBLIOGRAPHY. Comparative works include JAMES CRAMER, *The World's Police* (1964), a comprehensive survey of modes of police administration; ROY D. INGLETON, *Police of the World* (1979); and JOHN M. ANDRADE, *World Police and Paramilitary Forces* (1985), later works; HAROLD K. BECKER, *Police Systems of Europe*, 2nd ed. (1980), a survey of police organizations in 10 European countries; RICHARD J. TERRILL, *World Criminal Justice Systems* (1984), a cross-cultural study of five countries, including Japan and the Soviet Union; DAVID H. BAILEY, *Forces of Order: Police Behavior in Japan and the United States* (1976), a comparative work; R.V.G. CLARKE and J.M. HOUGH (eds.), *The Effectiveness of Policing* (1980), a collection of research reports on Canadian, English, Dutch, and American police; and PETER K. MANNING, *Police Work: The Social Organization of Policing* (1977, reprinted 1979), a comparison of British and American systems. THE INTERNATIONAL POLICE ASSOCIATION has published a useful *International Bibliography of Selected Police Literature*, 2nd ed. (1968). Additional sources of information may be found in JACK E. WHITEHOUSE, *A Police Bibliography: Published and Unpublished Sources Through 1976* (1980). For further study, see the same author's *Research Guide for Law Enforcement and the Criminal Justice System: A Bibliography of Bibliographies, Journals, Research, and Reference Materials* (1982); and GEORGE T. FELKENES and HAROLD K. BECKER, *Law Enforcement*, 2nd ed. (1977), an extensive bibliography containing resource articles.

There are histories of the police forces of many countries: T.A. CRITCHLEY, *A History of the Police in England and Wales*, rev. ed. (1978); MARCEL LE CLÈRE, *Histoire de la police*, 4th ed. (1973); JAMES F. RICHARDSON, *Urban Police in the United States* (1974); PAUL RIEGE, *Kleine Polizei-Geschichte*, 3rd ed. (1966); ALAN WILLIAMS, *The Police of Paris, 1718–1789* (1979); ROBERT CONQUEST (ed.), *The Soviet Police System* (1968); WILLIAM KELLY and NORA KELLY, *Policing in Canada* (1976); and KERRY L. MILTE, *Police in Australia* (1977). SAMUEL WALKER, *Popular Justice: A History of American Criminal Justice* (1980), gives an analysis of the criminal justice process in the United States.

For popular accounts of police work and the problems of accountability, two British books deserve mention: BEN WHITAKER, *Police in Society* (1979); and PETER LAURIE, *Scotland Yard* (1970). MICHAEL BANTON, *The Policeman in the Community* (1964), is a sociological comparison of the police role in Scotland and the United States. JOHN ALDERSON, *Policing Freedom* (1979), discusses the responsibilities of the police in democratic societies; and HERMAN GOLDSTEIN, *Policing a Free Society* (1977), analyzes the problems of discretion, accountability, and the application of technology in American policing. JAMES Q. WILSON, *Varieties of Police Behavior* (1968; reprinted 1978), compares the enforcement policies of eight American police departments. ROBERT M. FOGELSON, *Big-City Police* (1977), concentrates on reform in the American police; THOMAS A. REPPETTO, *The Blue Parade* (1978), is a readable history of both federal and state forces; and MARK M. MOORE and GEORGE L. KELLING, "To Serve and Protect: Learning from Police History," *Public Interest*, 70:49–65 (Winter 1983), is an analysis of police strategy. JEROME H. SKOLNICK, *Justice Without Trial*, 2nd ed. (1975), discusses the detective's role in California and is sensitive to the legal as well as the sociological issues. LOUIS A. RADELET, *The Police and the Community*, 3rd ed. (1980), examines police attempts to incorporate management principles and technological advances into their work.

The problems of police operations and accountability are also featured in the reports of the UNITED STATES. PRESIDENT'S COMMISSION ON LAW ENFORCEMENT AND ADMINISTRATION OF JUSTICE, *The Challenge of Crime in a Free Society* (1967), and *Task Force Report: The Police* (1967); and the *Report of the National Advisory Commission on Civil Disorders*, 6 vol. (1968). See also UNITED STATES. PRESIDENT'S COMMISSION ON LAW ENFORCEMENT AND ADMINISTRATION OF JUSTICE, *Task Force Report: Science and Technology* (1967), a study of the application of science and technology to control crime and to solve related sociological problems; PAUL L. KIRK and LOWELL W. BRADFORD, *The Crime Laboratory* (1965), a basic guide to establishing a crime laboratory; EDWARD CLINTON EZELL, *Small Arms of the World*, 12th rev. ed. (1983), a compilation of photographs, technical descriptions, and specifications as well as operational comments; DONALD O. SCHULTZ and LORAN A. NORTON, *Police Operational Intelligence* (1968), a basic guide for the establishment of police intelligence systems; and GEORGE B. JOHNSON and HANS BERT LOCKHOVEN, *International Armament*, 2 vol. (1965), a source for information on small arms, their specifications, and manufacture. EDWARD J. IMWINKELRIED (ed.), *Scientific and Expert Evidence*, 2nd ed. (1981), provides a comprehensive review of forensic science and analyzes the utility and the legal issues of physical evidence. *Crime and Justice* (irregular) is a periodical collection of articles on research and developments in criminal justice. For discussion of the use of computers and automation in police work, see V.A. LEONARD, *The New Police Technology* (1980). The following references are significant and comprehensive texts pertaining to the establishment, management, and operation of a modern forensic laboratory: H.J. WALLS, *Forensic Science*, 2nd ed. (1974); WILSON R. HARRISON, *Suspect Documents, Their Scientific Examination* (1958, reprinted 1981); DAVID A. CROWN, *The Forensic Examination of Paints and Pigments* (1968); JAMES W. OSTERBURG, *The Crime Laboratory*, 2nd ed. (1982); ARNE SVENSSON, OTTO WENDEL, and BARRY A.J. FISHER, *Techniques of Crime Scene Investigation*, trans. from the Swedish, 3rd ed. (1981); LEWIS C. NICKOLLS, *The Scientific Investigation of Crime* (1956); P.L. KIRK, *Crime Investigation*, 2nd ed. (1974, reprinted 1985); CHARLES E. O'HARA and JAMES W. OSTERBURG, *An Introduction to Criminalistics* (1949, reissued 1972); CHARLES E. O'HARA and GREGORY L. O'HARA, *Fundamentals of Criminal Investigation*, 5th ed. (1980); RICHARD H. WARD, *Introduction to Criminal Investigation* (1975); and DONALD O. SCHULTZ, *Criminal Investigation Techniques* (1978).

(M.P.B./W.B.McM./G.E.Mi./G.L.Ke.)

Polish Literature

Poland acquired a literary language in Latin when it became a Christian land in the 10th century, and thereafter literature in the Polish language was slow to emerge. Its development as a national literature was restrained in part by Poland's remoteness from cultural centres and the difficulties that assailed the young state, frequently attacked by plundering invaders and weakened by division into small principalities, and in part by the domination of foreign priests (when Mieszko I, prince of Poland, accepted Christianity in 966, in effect he placed his lands in the hands of the pope). This article reviews the literary history of Poland from its origins to the present day.

The article is divided into the following sections:

The Middle Ages 973
Religious writings
Early secular literature
The Renaissance period 973
Kochanowski and his disciples
Achievements in prose writing
The Baroque period 974
Poetry
Other literary forms
The Enlightenment 974
Rise of the Polish drama
Didactic element in prose writing
Continued development of lyric poetry
The 19th century 975
Romanticism
Positivism
The 20th century 976
The "Young Poland" movement
Literature in restored Poland
Literature after 1945
Bibliography 978

THE MIDDLE AGES

Religious writings. As in other European countries, Latin was at first the only literary language of Poland, and early writings included saints' lives and annals and chronicles written by monks and priests. Two among these works are most important: the *Chronicon,* compiled in about 1115 by a Benedictine known only as Gallus Anonymous, and the *Annales seu cronicae inclyti Regni Poloniae,* written in about 1480 by Bishop Jan Długosz. The two works introduced Polish history and literature into the culture of Europe. Use of the vernacular was allowed by the church where Latin could not meet particular needs—in prayers, sermons, and songs; and the oldest surviving text of poetry in Polish is a song in honour of the Virgin Mary, "Bogurodzica," in which language and rhythm are used with high artistic craftsmanship. The earliest extant copy of the text dates from 1407, but its origins are much earlier. Preaching in Polish became established toward the end of the 13th century; the earliest known example of Polish prose, the *Kazania świętokrzyskie* ("Sermons of the Holy Cross"), dating from the end of the 13th or the beginning of the 14th century, was discovered in 1890. Part of a translation of the Bible, made in about 1455 for Queen Sofia, widow of Władysław II Jagiełło, has also survived.

Early secular literature. Secular works began to appear toward the second part of the 15th century. There was a poem criticizing the papacy (*c.* 1449) by Jędrzej Gałka, a follower of reformers John Wycliffe and Jan Hus, and a high literary standard was achieved in a morality verse dialogue, *Rozmowa mistrza Polikarpa ze Śmiercią* ("Dialogue Between Master Polycarp and Death"). The style of the medieval period lasted late in Poland. Marcin Bielski represented this late medievalism in his *Kronika wszystkiego świata* (1551; "Chronicle of the Whole World"), the first general history in Polish.

The best examples of Polish literature imply a continuous literary tradition. Although its themes were those of the common European heritage, Polish writing of this period could be intensely personal within its anonymous framework. The groundwork for a leap into the ranks of major literatures was laid during the reign of Casimir III the Great, who in 1364 founded the University of Kraków.

THE RENAISSANCE PERIOD

Although the Renaissance reached Poland comparatively late, it was the golden age of Polish literature. External security, constitutional consolidation, and the Reformation contributed to this flowering.

The first generation of writers that were influenced by the Italian humanists wrote in Latin. Included among this group were Jan Dantyszek (Johannes Dantiscus), author of incidental verse, love poetry, and panegyric; Andrzej Krzycki (Cricius), an archbishop who wrote witty epigrams, political verse, and religious poems; and Klemens Janicki (Janicius), a peasant who studied in Italy, won there the title of poet laureate, and was the most original Polish poet of the age.

Mikołaj Rej of Nagłowice was notable for combining medieval and Renaissance aspects. Self-educated, he was the first idiomatically Polish talent and widely read writer of his time, being known as "the father of Polish literature." He wrote satirical epigrams, but of more importance were his prose works, especially *Świętych słów a spraw Pańskich kronika albo Postilla* (1557; "Chronicle or Comments on the Holy Words and Matters of the Lord"), a collection of Calvinist sermons, and the *Żywot człowieka poczciwego* (1568; "Life of an Honest Man"), a description of an ideal nobleman.

Kochanowski and his disciples. The second generation of humanist poets, and indeed the whole Renaissance period, was dominated by Jan Kochanowski. The son of a country squire, he wrote in the vernacular and was the first Polish writer to attempt both satirical poetry and classical tragedy, but his lyrical works surpassed these experiments. His crowning achievement, the first Polish work to equal the great poems of western Europe, was *Treny* (1580; *Laments*), inspired by despair after the death in 1579 of his young daughter, to be succeeded by a final recovery of his spiritual harmony. In Kochanowski's poetry the archaisms still apparent in Rej's work had almost disappeared; language and idiom were modern. The flexibility and assurance of his poetic genius were immediately recognized as a sign that the literary language had attained its maturity.

The humanist poet Jan Kochanowski

The most notable of Kochanowski's followers was Szymon Szymonowic (Simonides). He introduced in his *Sielanki* (1614; "Idylls") a poetic genre that was to retain its vitality until the end of the 19th century. These pastoral poems exemplified the processes of imitation, adaptation, and assimilation by means of which writers of the Renaissance period brought foreign models into the native tradition.

The numerous poems, in Latin and Polish, of Sebastian Klonowic were of interest for their description of contemporary life. *Worek Judaszów* (1600; "Judas' Sack") was a satirical poem on the plebeian life of Lublin, of which he was mayor.

Achievements in prose writing. The prose of the 16th century ranked with its poetry in vitality and range. The most eminent writer in Latin was Andrzej Frycz-Modrzewski. In *Commentariorum de republica emendanda libri quinque* (1554; "Commentary on Reforming the Republic in Five Books"), he evolved a bold social and political system, based on the principle of equality before God and the

law. Another notable writer in Latin was Marcin Kromer, scholar, humanist, historian, and Catholic apologist. Of his controversial works, most interesting is the Polish *Rozmowy dworzanina z mnichen* (1551–54; "Dialogues of a Courtier and a Monk"), a defense of Catholicism. Many historical and polemical writings and translations of the Bible were also published during this period; the Catholic translation by Jakub Wujek is an outstanding literary work.

Polish literature had become a national literature, reflecting the country's position as a great power with far-flung boundaries, the evolution of the nobility as a ruling class, and the nation's economic prosperity. Its influence spread east, above all to Moscow, while to the west its culture was represented by men of such high repute as the scientist and astronomer Nicolaus Copernicus (Mikołaj Kopernik).

THE BAROQUE PERIOD

The Baroque period began very early in Poland. In 1564 Poland invited the Jesuits to settle in the country, and from about 1570 Protestant influence began to wane. The Baroque style and outlook were congenial to the Polish spirit; the period was one of considerable literary output, in spite of almost incessant wars. Indeed, perhaps it mirrored, in its stylistic tension, the external strife characteristic of the 17th century.

Poetry. A forerunner of Baroque poetry was Mikołaj Sęp Szarzyński, who wrote predominantly religious poetry akin to that of the English Metaphysical poets. In this period satire and pastoral were the most popular forms. Foremost among satirists was Krzysztof Opaliński. His *Satyry albo przestrogi do naprawy rządu i obyczajów w Polszcze należące* (1650; "Satires or Warnings on the Reform of the Government and Customs in Poland") is bitter, pessimistic, and wide-ranging. The pastoral was represented by Samuel Twardowski, author of *Daphnis drzewem bobkowym* (1638; "Daphne into Laurel Tree") and a romance *Nadobna Pasqualina* (1655; "Fair Pasqualina"), a tale of sacred and profane love, in which Polish Baroque achieves its most finely wrought splendour. The *Roxolanki* (1654), a collection of love songs by Szymon Zimorowic, and the *Sielanki nowe ruskie* (1663; "New Ruthenian Idylls"), written by his brother Józef Bartłomiej Zimorowic, introduced topical dramatic elements into the traditional pastoral lyric; images of war and death are superimposed upon the pastoral background, with macabre effect and typical Baroque incongruity.

A parallel but less formalized rustic genre produced much verse celebrating rural life. One of the more successful examples is the *Votum* by Zbigniew Morsztyn, whose finest achievement, however, was in religious poetry. In contrast was the work of his cousin, Jan Andrzej Morsztyn, whose language was marked by the extravagant style of 16th-century Italian. The formal complexity and skill of his verse were unsurpassed; and his translation of the French dramatist Pierre Corneille's *Le Cid* has remained the standard Polish version.

Polish epics

The age was characterized by ambition to write heroic epics—a preoccupation to be explained perhaps by historical events: wars with Sweden, Russia, and Turkey, internal revolts, and attempts to introduce constitutional reforms. The Italian poet Torquato Tasso's *Gerusalemme liberata* (1581; "Jerusalem Liberated"), brilliantly translated by Piotr Kochanowski, inspired attempts at epics on national themes, notably the vigorous *Wojna chocimska* (written 1670, published 1850; "The War of Chocim") by Wacław Potocki. Another epic, the *Psalmodia polska* (1695; "Polish Psalmody"), by Wespazjan Kochowski, was written under the impact of John III Sobieski's victory over the Turks at Vienna in 1683, at which Kochowski was present. The work was the first example of a theme developed by writers of the Romantic movement—the messianic interpretation of Poland's destiny.

Other literary forms. The prose of the period did not rise to the level of its poetry, though there was a wealth of diaries and memoirs. Outstanding were the memoirs of Jan Chryzostom Pasek, a country squire and soldier. The period was also notable for the emergence of the letter as a literary form. The letters of John III Sobieski to his wife are remarkable for their passion and tenderness and for their day-by-day account of his experiences of battles and diplomacy. Another interesting development was the rise of a popular anonymous literature, exemplified by the *komedia rybałtowska* ("ribald comedies"). These were generally popular satiric comedies and broad farces written mainly by playwrights of plebeian birth. One of the few whose names are known was Piotr Baryka, author of a carnival comedy, *Z chłopa król* (performed 1633; "From Peasant to King"), which developed the theme of the introduction to Shakespeare's *Taming of the Shrew*. Some 30 anonymous examples of this type of comedy survive, and they include realistic depictions of popular customs and grotesquely humorous situations, parodying the lofty themes of "official" literature and so expressing an indirect protest against social inequality.

The last stage of Baroque literature (*c.* 1675–*c.* 1750) displayed a long process of decline, marked only by the emergence of the first women writers and by the major figure of Stanisław Konarski, a reformer of education, literature, and the political system, who wrote poetry in Latin, a tragedy in Polish (*Tragedia Epaminondy* [1880]), and political treatises.

It was not until the mid-20th century that the literature of the Baroque period was fully appreciated. It may well be regarded as the most enduring of Polish styles, for many of its features have recurred in the Romantic period and in modern writing.

THE ENLIGHTENMENT

The literary developments of the Polish Enlightenment

Close contact with western Europe, especially France and England, characterized literature of the Enlightenment period in Poland, whose writers were imbued with a desire to save the national culture from effects of partition and foreign rule. Literary developments included the rise of drama; introduction of the periodical and the novel; publication of the first Polish dictionary; and, in poetry, the introduction of *dumy* (ballads).

Rise of the Polish drama. Drama was established late in Poland. The earliest significant event was the inauguration of a national theatre in Warsaw in 1765. There were three principal dramatists: Franciszek Bohomolec, who satirized the aristocracy in adaptations of Molière; Wojciech Bogusławski, who wrote a popular national comic opera, *Cud mniemany czyli Krakowiacy i górale* (1794; "The Pretended Miracle or Krakovians and Highlanders"); and Franciszek Zabłocki, important for *Fircyk w zalotach* (1781; "The Dandy's Courtship") and *Sarmatyzm* (1785; "Sarmatian Ways"). Aleksander Fredro's comedies appeared when the Romantic movement was under way, and in them the influences of Molière and Carlo Goldoni were assimilated, as *Zemsta* (1834; "Vengeance") illustrated. They were remarkable for brilliant "type" characterization, ingenious construction, and metrical facility.

Didactic element in prose writing. Didacticism permeated most of the period's prose writing. Modern periodicals appeared at this time (*e.g.*, *Monitor*, 1765–85), and a Polish dictionary was published between 1807 and 1814. Bishop Adam Naruszewicz wrote early in the Enlightenment. His poetic works, considered chronologically, reflected the transition from Baroque to Classical, and he also wrote a history of Poland in which modern methods of scholarship were used. The most important poet, Bishop Ignacy Krasicki, of European outlook and skeptical intellect, wrote two mock-heroic poems, *Myszeis* (1775; "The Idylls of the Mice") and *Monachomachia* (1778; "War of the Monks"), as well as *Satyry* (1779; "Satires") and *Bajki i przypowieści* (1779; "Fables and Moral Tales"), and was remarkable for concise expression, formal elegance, and wit. Krasicki also introduced the novel *Mikołaja Doświadczyńskiego przypadki* (1776; "The Adventures of 'Nicholas Try-all' "), written in diary form and showing the influence of Swift and Rousseau. Two other outstanding poets were Stanisław Trembecki, whose works had an important place in Polish classicism as models of stylistic fluency, and Kajetan Węgierski, a freethinker and admirer of Voltaire, best known for his notorious lampoons.

Continued development of lyric poetry. Lyrical poetry continued to develop. Franciszek Karpiński developed

features of the Baroque style in popular pastorals and religious songs, while Franciszek Dionizy Kniaźnin, a conscious experimenter, rewrote much of his earlier work to achieve stylistic uniformity; his verse, cool and intellectual in quality, anticipated Romantic themes of folk poetry, popular superstition, and gypsy life.

Julian Ursyn Niemcewicz's writings were inspired by patriotism and concern for reform. He knew English literature thoroughly and made early translations of English Romantic ballads, his original *dumy* being the first in Poland; he introduced the historical novel to Poland with *Jan z Tęczyna* (1825; "Jan of Tęczyn"), which showed Sir Walter Scott's influence. His comedy *Powrót posła* (1790; "The Return of the Deputy") was one of the best dramatic works of the period, and *Śpiewy historyczne* (1816; "Historical Songs") was widely read.

After loss of national independence, with the third partition of the country between Russia, Austria, and Prussia in 1795–96, the tradition of patriotic poetry was continued by émigré soldier-poets in the Polish legions of Napoleon's army, among them Józef Wybicki, whose popular patriotic song "Mazurek Dąbrowskiego" (1797; "Dąbrowski's Mazurka") was adopted as the national anthem in 1918.

THE 19TH CENTURY

A second distinct period of Polish classicism, called pseudoclassicism, occurred in the early 19th century. In general, pseudoclassicist writing lacked freshness. During the Enlightenment period, literature had reestablished contact with the West and become the voice of national consciousness. Although it was the literature of a community undergoing a severe political crisis, it exercised an influence on neighbouring countries.

Romanticism. The Romantic period began later in Poland than in England or Germany, and it lasted longer. It has been regarded as the greatest period in Polish literature. The rise of Romanticism coincided with the loss of independence, and great writers found in it an expression of their own mood. A need to interpret their country's destiny gave the work of the three great Romantic poets—Adam Mickiewicz, Juliusz Słowacki, and Zygmunt Krasiński—visionary power and moral authority. Writing in exile, they kept faith in the restoration of Polish independence alive, and their concern gave the literature of the Polish Romantic movement its strength and passion. Mickiewicz was the greatest Polish poet and leader of the Romantic period. His *Poezye* (2 vol., 1822–23; "Poetry") was the first major literary event of the period. In its second volume were included parts two and four of *Dziady* (*Forefathers' Eve*), in which he combined folklore and mystic atmosphere to create a new kind of Romantic drama. Mickiewicz' greatest works were written after 1824, when he was deported to Russia for revolutionary activities as a student; they included *Sonety krymskie* (1826; *Sonnets from the Crimea*); a visionary third part of *Dziady* (1832); a messianic interpretation of Poland's past and future destiny, *Księgi narodu polskiego i pielgrzymstwa polskiego* (1832; *The Books of the Polish Nation and of the Polish Pilgrimage*), written in biblical prose; and a great epic, *Pan Tadeusz* (1834; *Master Thaddeus*).

The poets in exile

The suppression of the insurrection of 1830–31 drove the cultural elite into exile in France; among poets whom Mickiewicz joined there were Słowacki, Krasiński, and Cyprian Kamil Norwid. Słowacki, a Romantic in the fullest sense, who in all his work showed genius, wrote verse narratives in the style of Byron and accomplished lyric poetry. He was inspired by patriotic themes: *Kordian* (1834) was a drama of conspiracy and problems of commitment. His subtle *W Szwajcarii* (1839; "In Switzerland") is probably the finest lyrical work in Polish. Much of Słowacki's work was in dramatic form, and although written for an imaginary stage rather than for an intended production, it laid the foundations of Polish tragic drama. His plays showed the influence of French Romantic drama, Shakespeare, classical tragedy, and Calderón. The last years of Słowacki's life were devoted to writing *Król-Duch* (1847; "The Spirit King"), an unfinished lyrical and symbolic epic describing the history of a people as a series of incarnations of the essential spirit of the nation.

Zygmunt Krasiński, when 23, published (anonymously, as with all his works) *Nieboska komedia* (1835; *The Undivine Comedy*), which presented, for the first time in Europe, a struggle between opposed worlds of aristocracy and disinherited masses. *Irydion* (1836; *Iridion*), his second play, was an allegory of Poland's fate. In *Przedświt* (1843; "Daybreak") he developed a messianic interpretation of Polish history, and this conception of Poland as "the Christ among the nations" was also expounded in *Psalmy przyszłości* (1845; "Psalms of the Future"). The introduction of fantastic or supernatural elements into a realistic setting was characteristic of many Polish Romantic works.

The genius of Cyprian Norwid was not fully recognized until the 20th century. During his lifetime he was misjudged and remained obscure, partly because he accepted some ideas of Romanticism while criticizing others but even more because of his ironic intellectual reserve. Among the most important works published in his lifetime was a verse dialogue on aesthetics, *Promethidion* (1851), which expounded a theory of the social and moral function of art anticipating that of John Ruskin. An authentic text of his most important lyrical collection, *Vade-mecum,* was first published in 1947. Norwid experimented with free verse and with the rhythms of speech, and, furthermore, he foreshadowed the French Symbolists in his analogical method of presenting the poetic concept.

The lesser talents of early Romanticism formed the Polish Ukrainian school, of whom Antoni Malczewski was outstanding as the author of a single poem, the Romantic verse narrative *Maria* (1825), a tale of love and treachery remarkable for original diction, dramatic tension, and unity of mood.

There were fewer prose writers than poets among the exiles. Zygmunt Miłkowski (pseudonym Teodor Tomasz Jeż) wrote on a wide range of subjects, including folklore and the history of the Balkan countries. The literary criticism of Maurycy Mochnacki, a passionate advocate of Romanticism and the first Polish critic to link literature with Poland's political progress, exercised a strong and not wholly beneficial influence on literary theory. The historical works of Joachim Lelewel, a great, many-sided scholar, were an impressive example of the prose of the period.

As a result of partition, Romantic poetry in Poland was limited to closed provincial circles. In Warsaw a group of young poets was formed, but its activities were restricted by political pressure. Its most fully developed talent was that of Teofil Lenartowicz. Ryszard Wincenty Berwiński, a poet of social radicalism, wrote *Poezje* (1844) and *Studia o literaturze ludowej* (1854; "Studies on Folk Literature"), which marked a step away from Romantic nationalist interpretations and stressed the international community of folk tradition.

Prose was more popular with writers in Poland than with those in exile. Henryk Rzewuski belonged spiritually to the 18th century: *Pamiątki J. Pana Seweryna Soplicy* (1839; "Memoirs of Mister Seweryn Soplica") evoked the atmosphere of the Baroque tradition. As the century progressed, signs of a Realistic tendency were discernible in Józef Korzeniowski's novels *Spekulant* (1846; "The Speculator") and *Kollokacja* (1847; "The Collocation"). A woman novelist, Narcyza Żmichowska (pseudonym Gabryella), produced *Poganka* (1846; "The Pagan"), a psychological allegory anticipating 20th-century sensibility in its subtle analysis of feeling. The dominant figure among prose writers was Józef Ignacy Kraszewski, whose output ran into hundreds of volumes of fiction, history, ethnography, criticism, and so on. His imaginative writings reflected a change of literary styles during his long career. Banished in 1863, he continued to influence Polish writers at home and in exile, maintaining the Polish cause by his manifold activities.

The influence of Polish Romanticism

Polish Romanticism, conscious of its role as the torch of national spirit, retained its force as a mode of thinking beyond the period of the political circumstances that fostered it. It produced works of highest artistic value, which excited the interest of foreign writers. Mickiewicz influenced Slavonic literatures and was compared by George Sand with Goethe and Byron. Słowacki's poetic technique

proved of fundamental importance to writers at the end of the 19th century, whereas Norwid's influence grew steadily stronger in the 20th century. The political implications of the Romantic movement led to the insurrection of 1863, which ended in Poland's becoming a province of Russia.

Positivism. The literary trend of the period following the 1863 uprising was called positivism and reflected a practical mood and reaction against Romanticism, the rise in Russian and Austrian Poland of an urban intelligentsia acting as a leaven in this ferment. Periodicals were of particular importance in disseminating new ideas, especially the *Tygodnik Ilustrowany* ("Illustrated Weekly"), founded in 1859. The natural consequence of a positivist outlook was a predominance of prose. With other writers of the Warsaw school, Aleksander Świętochowski voiced anti-clerical and anti-aristocratic views in his weekly *Prawda* ("Truth"). Bolesław Prus (Aleksander Głowacki), a journalist, ranked high among Polish novelists: *Lalka* (1890; *The Doll*) gave a complex picture of bourgeois life in Warsaw, and *Faraon* (1897; *The Pharaoh and the Priest*) was an ambitious evocation of ancient Egypt, masking political problems that could not be published in the form of a modern novel.

Eliza Orzeszkowa, a campaigner for social reform, wrote about women's emancipation, the ignorance of the peasants, and the Jewish problem. Her books showed psychological penetration and a fine sense of style.

In 1905 Henryk Sienkiewicz was a winner of the Nobel Prize for Literature; his early critical works propounded positivist aesthetic theories, and his early short stories showed an acute awareness of the contemporary situation and a stark Realism. His famous *Quo Vadis?* (1896; Eng. trans. 1898), a historical novel of Rome under Nero, has been widely translated.

A more flexible Naturalism gained ground toward the end of the period, as seen in the stories of Adolf Dygasiński, famous for portrayals of animal life—such as *Zajęc* (1900; "The Hare")—which could be compared with those of Rudyard Kipling. Gabriela Zapolska, a critic of social hypocrisy in Naturalistic novels and lively comedies, excelled in dialogue and dramatic situations, in such plays as *Moralność Pani Dulskiej* (1906; "The Morality of Mrs. Dulska").

The period produced two important positivist poets: Adam Asnyk, a reflective lyricist of formal dexterity; and Maria Konopnicka, who wrote of the plight of the oppressed.

THE 20TH CENTURY

The "Young Poland" movement. The "Young Poland" movement describes several different groups and tendencies united by opposition to positivism, and a desire to return to imagination in literature; hence its other name, Neoromanticism. Among its pioneers were Antoni Lange, a poet, and Zenon Przesmycki (pseudonym Miriam), editor of a Symbolist review, *Chimera.* Both made translations from a number of other languages and expressed aesthetic theories in critical essays. Przesmycki's most influential contribution to a development of modern literature was his discovery of Cyprian Norwid.

Kazimierz Tetmajer achieved popularity with his nostalgic *Poezje* (1891–1924), but his prose had greater vigour and precision of observation, *Na skalnym Podhalu* (1903–10; *Tales of the Tatras*) containing some effectively stylized folk material. An interesting poet, Jan Kasprowicz, was of peasant origin. His principal contribution to Polish literature lay in the structure of his longer lyrical poems; those in the volume *Ginącemu światu* (1901; "To a Dying World") employed a technique of association, quotation, musical repetition, and free metre that anticipated the early style of such poets as T.S. Eliot. Tadeusz Miciński, a forerunner of Expressionism and Surrealism, wrote philosophical and mystical poems and plays, notably a collection of poems, *W mroku gwiazd* (1902; "In the Twilight of the Stars"), and a play, *Kniaź Patiomkin* (1906; "Prince Potemkin"). The lyrical poet Leopold Staff, whose work shows great variety and technical dexterity, was at this period associated with the "Young Poland" movement, although some of his finest work was written later.

Stanisław Przybyszewski was a leading exponent of the movement's new aesthetic theories and edited a literary magazine, *Życie* ("Life"). Stefan Żeromski expressed passionate concern for social justice and national freedom in widely read works, but an excess of Realist documentation frequently vitiated the power of his later work. Władysław Stanisław Reymont, of peasant stock, adapted the Naturalistic technique to create a vision of peasant life in an epic novel cycle, *Chłopi* (4 vol., 1904–09; *The Peasants*), for which he received the Nobel Prize for Literature for 1924. One of the most effective novels of the period, *Żywot i myśli Zygmunta Podfilipskiego* (1898; "The Life and Thoughts of Zygmunt Podfilipski"), by Józef Weyssenhoff, presented an ironical portrait of the egoist in society. Wacław Berent's *Próchno* (1903; "Rotten Wood") portrayed with biting irony the late 19th-century decadence of life and art. His *Ozimina* (1911; "Winter Corn"), a Symbolist novel, foreshadowed the associative structure and narrative technique of James Joyce's *Ulysses* (1922). His *Żywe kamienie* (1918; "The Living Stones") stressed the unity of medieval culture and Poland's place within it. A bold experiment antedating by several years the psychoanalytical novel in western Europe was *Pałuba* (1903; "The Hag") by Karol Irzykowski. In it, motivation and behaviour were presented from different viewpoints, ingeniously cemented by the author's own analyses, as in a scientific study. Irzykowski was also a critic and, in *Dziesiąta muza* (1924; "The 10th Muse"), the first to give attention to the cinema as an art form.

Stanisław Brzozowski, an outstanding critic, insisted that a critic must represent the moral consciousness of his age; in *Legenda Młodej Polski* (1909; "The Legend of Young Poland") he analyzed the weakness of the period's literature and expounded his view of the unity of all work—physical, technical, intellectual, and artistic.

Stanisław Wyspiański was an artist and dramatist of genius. In his plays he reforged elements from classical tragedy and mythology, Polish Romantic drama, and national history in a complex whole. *Wesele* (1901; "The Wedding") was a visionary parable of Poland's past, present, and problematical future, cast in the form of the traditional puppet-theatre play. It was a masterpiece of evocative allusion, tragedy, and humour.

The literature of the period was characterized by close contact with western European literatures, but writers such as Wyspiański turned back to the Polish Romantics in a search for new poetic language.

Literature in restored Poland. The restoration of independence in 1918 decisively affected Polish literature. The period between 1918 and 1939 was characterized by richness and variety and increasing contact with other European literatures, especially through publication of many translations. For nearly a decade after 1918 lyrical poetry predominated. A periodical, *Zdrój* ("The Fountainhead"), showed affinities with German Expressionism. In Warsaw several poets formed a group called Skamander, from the name of their monthly publication; it was united by a desire to forge a poetic language, attuned to modern life. Julian Tuwim, a poet of emotional power and linguistic sensitivity, headed the group, and during World War II, in exile in Brazil and the United States, wrote a long, discursive, autobiographical poem, *Kwiaty polskie* (1949; "Polish Flowers"). Among Skamander members were Jan Lechoń, Kazimierz Wierzyński (both died abroad after World War II), Antoni Słonimski, and Jarosław Iwaszkiewicz, who was also a prolific prose writer; among sympathizers of the group were two eminent woman poets, Maria Pawlikowska-Jasnorzewska and Kazimiera Iłłakowiczówna, and Władysław Broniewski, a poet with strong left-wing sympathies, who became a master of the revolutionary lyric, expressing involvement in current social and ideological problems.

The Skamander poets

A writer of importance was Bolesław Leśmian, whose symbolic, Expressionist poetry was remarkable for the inventiveness of its vocabulary, sensuous imagery, and philosophic content, all anticipating Existentialism. He published only three notable collections—*Łąka* (1920; "The Meadow"), *Napój cienisty* (1936; "The Shadowy Drink"), and *Dziejba leśna* (posthumous, 1938; "Wood-

land Tale")—but has been recognized by his admirers as the most outstanding 20th-century Polish lyrical poet.

Other experimental movements, such as Futurism, followed revolutionary trends in poetry—particularly in Italy and Russia. More original was a group called Awangarda Krakowska ("Vanguard of Kraków"). Led by Tadeusz Peiper, it produced few works but had widespread influence in the regeneration of poetic technique. Two of its adherents, Julian Przyboś and Adam Ważyk, who was loosely connected with the movement, ranked among the outstanding poets of the post-World War II period. Also noteworthy was Józef Czechowicz, who assimilated traditional and regional elements to the new style.

Prose writing reached its ascendancy in the second decade of independence. Early stories by Zofia Nałkowska belonged to the "Young Poland" movement and aimed to expose the feminine psyche; later she turned to other themes, striving for narrative objectivity and technical simplicity. Two other women writers of distinction were Zofia Kossak-Szczucka, noted for historical novels, and Maria Kuncewiczowa, who wrote psychological novels. Juliusz Kaden-Bandrowski used experimental Realism in *Czarne skrzydła* (1928–29; "Black Wings") and *Mateusz Bigda* (1933), which treated social and political themes. Michał Choromański's *Zazdrość i medycyna* (1933; *Jealousy and Medicine*) used experimental methods of narrative sequence and was remarkable for its clinical analysis of character. A writer skilled in reflecting subtleties of perception was Bruno Schulz, author of *Sklepy cynamonowe* (1934; *Cinnamon Shops*), whose prose was reminiscent of that of Kafka.

Tadeusz Żeleński (pseudonym Boy), witty, irreverent, and widely read, was a leading literary critic and one of Poland's best interpreters of French literature. The essay form was represented by Jan Parandowski, whose main theme was the classical culture of Greece and Rome. A subversive attack on intellectual and social conventions was launched in *Ferdydurke* (1937; Eng. trans., *Ferdydurke*) by Witold Gombrowicz, who displayed in it a satirical talent reminiscent of Alfred Jarry. The taste for the cyclic novel was satisfied by Maria Dąbrowska with *Noce i dnie* (4 vol., 1932–34; "Nights and Days"), an outstanding modern Polish example of a chronicle novel in epic style, about the development of the Polish intelligentsia of upper-middle-class origin.

The drama was the weakest of the literary forms during this period and returned to the Young Poland's especially symbolic form (Karol Hubert Rostworowski, Jerzy Szaniawski). The experimental dramas of Stanisław Ignacy Witkiewicz were of interest chiefly for their expression of anti-Realist aesthetic theories; he developed many ideas of the Awangarda, applied the principles of "pure form" to painting and drama, and was the main exponent of a movement in Polish literature known as "catastrophism." Obsessed with the idea of a disintegration of European culture, endangered by totalitarian ideologies, and an attempt to impose the uniformity of a "mass society," he developed his ideas in plays combining elements of Surrealism, grotesque misrepresentation, and what later became known (in the plays of Eugène Ionesco, for example, whose work Witkiewicz to some extent foreshadowed) as the Theatre of the Absurd. His novels *Pożegnanie jesieni* (1927; "Farewell to Autumn") and *Nienasycenie* (1930; "Insatiability") expressed the same philosophy. After World War II his work attracted interest abroad and appeared in translation (*e.g., The Madman and the Nun and Other Plays,* 1968).

Witkiewicz and "catastrophism"

Literature after 1945. The impact of World War II, the experience of occupation, and the establishment in 1945 of the People's Republic decisively affected the character of literature in Poland and also produced a number of émigré writers who had become famous between World Wars I and II: lyrical poets of the Skamander group and former associates of the Awangarda movement. Among them was Czesław Miłosz, who emigrated to France in 1951 and moved to the United States a decade later. He was awarded the Prix Littéraire Européen for *Zdobycie władzy* (1955; first published in French, as *La prise du pouvoir,* 1953; *The Seizure of Power*), and he received the Nobel Prize for Literature for 1980. Many émigrés wrote of experiences in enemies' prisons and forced-labour camps. The most accomplished as literature were *Inny świat* (1953; *A World Apart*) by Gustaw Herling-Grudziński and *Sprawa Józefa Mosta* (1953; "The Affair of Józef Most") by Herminia Naglerowa. Jerzy Pietrkiewicz, whose early work was published in Poland during World War II, later wrote mainly in English (the novels *The Knotted Cord,* 1953; *Isolation,* 1959). Witold Gombrowicz, who died in France after a long stay in Argentina, also published his postwar work abroad and became famous with the novels *Trans-Atlantyk* (1953), *Pornografia* (1960; Eng. trans., *Pornografia*), and *Kosmos* (1965; *Cosmos*), which won him the 1967 International Prize for Literature. He also published abroad the plays *Ślub* (1953; *The Marriage*) and *Operetka* (1966; "Operetta"); three volumes of diaries (*Dziennik,* 1953–66); and a book of recollections and reflections, *Wspomnienia polskie* (1977; "Polish Reminiscences"). In all these works, especially the novels, Gombrowicz treated philosophical and psychological themes in a satirical narrative style in which, by emphasizing the grotesque and irrational elements in human nature, he presented an exposé of the conventions of modern life and culture.

In Poland the years immediately after the war saw publication of works written during the occupation or by writers who had been in concentration camp or prison. A frequent theme was the attempt to come to terms with Fascism and war. This was exemplified in the short stories of Tadeusz Borowski, a prisoner at Auschwitz, who, in collections published in 1948, *Pożegnanie z Marią* ("Farewell to Mary") and *Kamienny świat* ("The World of Stone," published in *This Way for the Gas, Ladies and Gentlemen, and Other Stories,* 1967), explored human depravity and degradation; and in those of Adolf Rudnicki, who treated moral and philosophical themes in lyrical prose and described, in *Szekspir* (1948; "Shakespeare") and *Ucieczka z Jasnej Polany* (1949; "Flight from Jasna Polana"), the wartime fate of the Jewish community in Poland. Another theme, shown, for example, by a novel *Popiół i diament* (1948; *Ashes and Diamonds*) by Jerzy Andrzejewski, was the examination of the moral controversies that accompanied political and social changes of the postwar period, especially the tragic situation of young conspirators involved in the struggle against the new regime.

During 1949–54 the literature of Socialist Realism gained ground, and writers attempted to emulate the great 19th-century Realists. A new type of hero was created—the ordinary man or woman actively engaged in "productive" work, and those elements in the social scene that served to present an idea of revolutionary progress were accentuated. The main writer was Leon Kruczkowski, a prewar exponent of the Marxist novel whose outstanding plays *Niemcy* (1949; "The Germans") and *Pierwszy dzień wolności* (1960; "The First Day of Freedom") were successfully performed abroad. Kazimierz Brandys, whose development typified postwar tendencies in Polish literature, published an epic novel cycle, *Między wojnami* (1948–52; "Between the Wars"), and a Socialist Realist novel, *Obywatele* (1954; "Citizens").

The literature of Socialist Realism and the reaction against it

Among writers of the period who stood aside from political involvement were Konstanty Ildefons Gałczyński, who combined lyricism with grotesque fantasy, and the reflective Mieczysław Jastrun, who in later work—for example, the essay collection *Mit śródziemnomorski* (1962; "The Mediterranean Myth")—moved toward Existentialism. A group of Catholic writers opposed to Socialist Realism included Antoni Gołubiew, author of an epic novel cycle, *Bolesław Chrobry* (1947–54), the prose writer and dramatist Jerzy Zawieyski, and a historical novelist, Hanna Malewska. Teodor Parnicki used a background of conflict between cultures for an analysis of contemporary problems in experimental and semi-historical novels set mainly in the early Christian period: *Koniec "Zgody Narodów"* (1955; "End of the Covenant of Nations"), *Nowa baśń* (6 vol., 1962–70; "A New Fairytale"), and others.

The weakness of the Socialist Realist movement—its attempt to impose a political pattern on creative writing, its avoidance of themes arising from contemporary

conflicts—resulted partly from the stranglehold of the Stalinist regime. In the period beginning in 1954–55, writers began to oppose these weaknesses. Andrzejewski, for example, presented contemporary ideas and problems in the novels *Ciemności kryją ziemię* (1957; *The Inquisitors*) and *Bramy raju* (1960; *The Gates of Paradise*), and Brandys criticized Stalinism in the novel *Matka Królów* (1957; "Mother of the Krols," Eng. trans., *Sons and Comrades*).

The political "thaw" that followed Nikita Khrushchev's 1953 rise to leadership of the Soviet Union, particularly after Khrushchev's denunciation of Stalinism in early 1956, allowed renewed contact with the West, bringing about a period of development and experiment marked by use of the essay for philosophical and intellectual discussion and by an increase in satirical literature. Tadeusz Breza published *Spiżowa brama* (1960; "The Bronze Gate"), a keen description of Vatican life. Still concerned with World War II were such novels as *Czarny potok* (1954; *Black Torrent*) by Leopold Buczkowski, *Kolumbowie-rocznik 20* (1957; "The Columbuses-Generation of 1920") by Roman Bratny, and *Tren* (1961; "Threnody") by Bohdan Czeszko. Tadeusz Konwicki, like others, wrote of the consequences of wartime experience, notably in his *Sennik współczesny* (1963; *A Dreambook for Our Time*). His later novels *Wniebowstąpienie* (1967; "Ascension") and *Nic albo nic* (1971; "Nothing or Nothing") project those themes on contemporary problems. In *Głosy w ciemności* (1956; "Voices in the Dark") and *Austeria* (1966; *The Inn*), Julian Stryjkowski gave universal application to the feeling of the Orthodox Jewish Polish community that the world has already ended.

Some prewar Polish writers continued to publish: the poets Leopold Staff and Kazimiera Iłłakowiczówna; Maria Dąbrowska, who enhanced her reputation with the short-story collection *Gwiazda zaranna* (1955; "Morning Star") and a series of critical essays on Joseph Conrad; the novelist Maria Kuncewiczowa; and the novelist, poet, and dramatist Jarosław Iwaszkiewicz, who published the epic novel *Sława i chwała* (1956–62; "Fame and Glory").

Many writers of the late 1950s and the 1960s wrote fiction that dealt with the contemporary scene, ranging from the political novels of Jerzy Putrament to the satirical novels of manners by Stanisław Dygat. Works by science-fiction writer Stanisław Lem were translated into many languages. Young writers such as Marek Hłasko and Marek Nowakowski, seeking a moral basis for life, often looked into the worlds of outcasts and misfits. Sławomir Mrożek, both in his plays—*Policja* (1958; "The Police"), *Na pełnym morzu* (1961; *Out at Sea*), *Striptease* (1962; Eng. trans., *Striptease*), and above all *Tango* (1964; Eng. trans., *Tango*), his most widely known work—and in stories collected in *Słoń* (1957; *The Elephant*), displayed an acute sense of satire and the grotesque. His comedy, belonging partly to the Theatre of the Absurd, was distinguished by subtle parody and highly stylized language.

New trends in poetry and drama. Poetry after 1956 was a vehicle for philosophical thought. The satirical poet Stanisław Jerzy Lec was noted for his skeptical aphorisms in *Myśli nieuczesane* (published in series from 1957; *Unkempt Thoughts*). Zbigniew Herbert, an outstanding 20th-century poet, distinguished himself with moralist and metaphysical poems (many of which appeared in English translation in two volumes entitled *Selected Poems*, 1968 and 1977). Poet Tadeusz Różewicz had a profound influence; from *Niepokój* (1947; *Faces of Anxiety*), his first collection, to *Głos anonima* (1961; "The Nameless Voice"), Różewicz's work was occupied with moral themes. He also wrote plays resembling those of Eugène Ionesco: *Świadkowie albo nasza mała stabilizacja* (1962; "The Witnesses, or Our Little Stabilization"; translated in *The Witnesses and Other Plays*), and one published with poems in *Kartoteka* (1961; *The Card Index and Other Plays*). The individual's entanglement in momentous 20th-century events dominated the intellectual, ironic poetry of Wisława Szymborska.

The work of lyric poets born about 1930 was characterized by a variety of aims and styles. Some poets, like Miron Białoszewski, experimented drastically; others, like Ernest Bryll, reasserted traditional poetic forms. Some—Tadeusz Nowak and Jerzy Harasymowicz, for example—turned for inspiration to peasant culture; others, among them Jarosław M. Rymkiewicz, an outstanding translator of English and American poetry, followed the example of T.S. Eliot and returned to Baroque and Neoclassical forms, writing erudite, allusive poetry. Most representative of these poets is perhaps Stanisław Grochowiak, who created a style based on unexpected juxtapositions and a deliberate emphasis on the grotesque.

Critics and essayists included Artur Sandauer; Jan Kott (who immigrated to the United States), whose *Szkice o Szekspirze* (1961; *Shakespeare, Our Contemporary*) was widely translated and discussed; Kazimierz Wyka; Jan Błoński; and Andrzej Kijowski. Leopold Tyrmand, a successful novelist, settled down in the United States, where he published a series of essays highly critical of communism.

Political events in 1968—student riots, anti-Semitic campaigns, harsher censorship—forced many writers to emigrate and publish abroad. Émigré centres such as those in Paris played an increasing role in supporting the opposition and promoting literature free of censorship. Particularly important was the Institut Littéraire in France, headed by Jerzy Giedroyc. The Institut was the publisher of the literary monthly *Kultura* (1947–2000). (After 1989, when the communist system was abolished in Poland, writers and books circulated freely, but the role of émigré publishers in promoting Polish literature remained quite visible.) Among those writers who stayed in Poland, many, including Paweł Jasienica and Stefan Kisielewski, were temporarily blacklisted for their political views. (Jasienica had published a series of historical studies emphasizing Poland's liberal traditions, while Kisielewski strongly criticized the political system in his magazine column.) In the 1970s and early 1980s, major civil unrest was expressed in particular in the workers' riots and strikes leading to the creation of the independent trade union Solidarity. Writers challenged the official aesthetic. Censorship still forced many of them to publish abroad, but independent publishing houses emerged in Poland. These published not only writers in Poland but hitherto proscribed foreign authors, such as George Orwell, and émigré authors, such as Miłosz and Gombrowicz. Among the most important books of the period were Konwicki's novels *Kompleks polski* (1977; *The Polish Complex*) and *Mała apokalipsa* (1979; *A Minor Apocalypse*); Andrzejewski's long-suppressed novel *Miazga* (1981; "Pulp"); and the ironic reflections recorded in a monthly diary called *Miesiące* (1980; *A Warsaw Diary 1978–1981*) by Kazimierz Brandys, who left Poland in 1981. Also important were volumes of poetry and essays by Ryszard Krynicki and Stanisław Barańczak, who moved to the United States; the diaries of Gombrowicz; and two additional volumes of Zbigniew Herbert's poems, *Pan Cogito* (1974; "Mr. Cogito"; partially translated in *Selected Poems*), the poetic meditations of a kind of intellectual Everyman, and *Raport z oblężonego miasta* (1983; *Report from the Besieged City*).

With the restoration of free Poland after 1989, literature's political influence diminished but its artistic values remained strong, as evinced by Wisława Szymborska's receipt of the Nobel Prize for Literature in 1996. Polish literature at the turn of the century accommodated various literary trends. While many young writers escaped into realms of fantasy and abstraction, a trend toward realism resulted in outstanding novels such as *Madame* (1998; Eng. trans. *Madame*), by Antoni Libera, and *Oksana* (1999), by Włodzimierz Odojewski.

BIBLIOGRAPHY. CZESŁAW MIŁOSZ, *The History of Polish Literature*, 2nd ed. (1983), is the best modern survey of Polish literature in English. JULIAN KRZYŻANOWSKI, *A History of Polish Literature* (1978; originally published in Polish, 1969), is also useful. MANFRED KRIDL, *A Survey of Polish Literature and Culture*, trans. from Polish (1956; reprinted 1967), is a slightly dated but still informative work. A useful reference work is S.J. CZERWINSKI (ed.), *Dictionary of Polish Literature* (1994). Other works of interest include WACŁAW LEDNICKI, *Life and Culture of Poland: As Reflected in Polish Literature* (1944); SAMUEL FISZMAN (ed.), *The Polish Renaissance in Its European Context* (1988); and KAZIMIERZ BRAUN, *A History of Polish Theater, 1939–1989: Spheres of Captivity and Freedom* (1996; originally published in Polish, 1994). (J.Kr./Ja.W./B.W.Ma./J.R.K./Ed.)

Political Parties and Interest Groups

A political party is a group organized to achieve and exercise power within a political system. An interest group (or special-interest group) is an aggregate of individuals who, bound by one or more concerns or wants, makes claims upon other groups or upon society in general in order to maintain or promote its position or objectives.

An interest group that attempts to influence government becomes a "pressure group." It is distinguished from a political party in that, whereas a party puts up candidates for election to public office and carries out activities to secure their election, a pressure group seeks to influence both the government and the parties. The Roman Catholic interest in Italy, for instance, is served by the Christian Democratic Party and also by the Catholic Action Society. The first is a party, and the second is a pressure group. Again the ethnic Swedes in Finland is served by the Swedish People's Party; this is clearly a special-interest group, yet it puts up candidates and is represented in the legislature because it thinks its members think their constituents' interest is best served by exerting pressure as part of a governmental coalition.

Thus, what begins as a pressure group may eventually become a political party. The British Labour Party, for example, had its origins in 1900, when trade unionists formed a Labour Representation Committee in order to elect working-class MPs to Parliament; later it broadened its membership and program to become the Labour Party. In France, Pierre Poujade's pressure group, the Union of Small Shopkeepers, was involved in the 1956 elections and won more than 40 seats, thus becoming a political party, before reverting to pressure-group status in 1958. The reverse process can also occur, as when a political party engenders pressure groups; thus, Communist parties set up interest groups for women, youth, workers, and the like. (S.E.F./Ed.)

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 541 and 543, and the *Index*.

The article is divided into the following sections:

Political parties 979
- Types of political party 979
  - Cadre parties
  - Mass-based parties
- Parties and political power 981
  - The struggle for power
  - Participation in power
  - Power and representation
- Party systems 983
  - Multiparty systems
  - Two-party systems
  - Single-party systems
- Future of political parties 986

Interest groups 986
- Types of interest group 986
  - Categories defined by purpose
  - Categories defined by structure
  - Categories defined by political activity
- Characteristics of pressure groups 987
  - Power elements
  - Strategy and tactics
- Role of interest groups in various political systems 988
  - The Anglo-American and Swedish patterns
  - The French and Italian patterns
  - Patterns in developing countries
  - Patterns in former Communist states

Bibliography 990

Political parties

Political parties originated in their modern form in Europe and the United States in the 19th century, along with the electoral and parliamentary systems, whose development reflects the evolution of parties. The term party has since come to be applied to all organized groups seeking political power, whether by democratic elections or by revolution.

In earlier, prerevolutionary, aristocratic and monarchical regimes, the political process unfolded within restricted circles in which cliques and factions, grouped around particular noblemen or influential personalities, were opposed to one another. The establishment of parliamentary regimes and the appearance of parties at first scarcely changed this situation. To cliques formed around princes, dukes, counts, or marquesses there were added cliques formed around bankers, merchants, industrialists, and businessmen. Regimes supported by nobles were succeeded by regimes supported by other elites. These narrowly based parties were later transformed to a greater or lesser extent, for in the 19th century in Europe and America there emerged parties depending on mass support.

Ubiquity of parties in the 20th century

The 20th century saw the spread of political parties throughout the entire world. In Africa large parties have sometimes been formed in which a modern organization has a more traditional ethnic or tribal basis; in such cases the party leadership is frequently made up of tribal chiefs. In certain areas of Asia, membership in modern political parties is often determined largely by religious factors or by affiliation with ritual brotherhoods. Many political parties in the developing countries are partly political, partly military. Certain Socialist and Communist parties in Europe earlier experienced the same tendencies.

These last-mentioned European parties have demonstrated an equal aptitude for functioning within multiparty democracies and as the sole political party in a dictatorship. Developing originally within the framework of liberal democracy in the 19th century, political parties have been used in the 20th century by dictatorships for entirely undemocratic purposes.

TYPES OF POLITICAL PARTY

A fundamental distinction can be made between cadre parties and mass-based parties. The two forms coexist in many countries, particularly in western Europe, where Communist and Socialist parties have emerged alongside the older conservative and liberal parties. Many parties do not fall exactly into either category but combine some characteristics of both.

Cadre parties. Cadre parties—*i.e.*, parties dominated by politically elite groups of activists—developed in Europe and America during the 19th century. Except in some of the states of the United States, France from 1848, and the German Empire from 1871, the suffrage was largely restricted to taxpayers and property owners, and, even when the right to vote was given to larger numbers of people, political influence was essentially limited to a very small segment of the population. The mass of people were limited to the role of spectators rather than that of active participants.

The cadre parties of the 19th century reflected a fundamental conflict between two classes: the aristocracy on the one hand and the bourgeoisie on the other. The former, composed of landowners, depended upon rural estates on which a generally unlettered peasantry was held back by a traditionalist clergy. The bourgeoisie, made up of industrialists, merchants, tradesmen, bankers, financiers, and professional people, depended upon the lower classes of clerks and industrial workers in the cities. Both aristocracy and bourgeoisie evolved its own ideology. Bourgeois liberal ideology developed first, originating at the time of the English revolution of the 17th century in the writings of John Locke, an English philosopher. It was then developed by French philosophers of the 18th century. In its clamouring for formal legal equality and acceptance of the inequities of circumstance, liberal ideology reflected the interests of the bourgeoisie, who wished to destroy the privileges of the aristocracy and eliminate the lingering economic restraints of feudalism and mercantilism. But, insofar as it set forth an egalitarian ideal and a demand for liberty, bourgeois classical liberalism expressed aspirations common to all men. Conservative ideology, on the other hand, never succeeded in defining themes that would prove as attractive, for it appeared to be more closely allied to the interests of the aristocracy. For a considerable period, however, conservative sentiment did maintain a considerable impact among the people, since it was presented as the expression of the will of God. In Catholic countries, in which religion was based upon a hierarchically structured and authoritarian clergy, the conservative parties were often the clerical parties, as in France, Italy, and Belgium.

Conservative and liberal cadre parties dominated European politics in the 19th century. Developing during a period of great social and economic upheaval, they exercised power largely through electoral and parliamentary activity. Once in power, their leaders used the power of the army or of the police; the party itself was not generally organized for violent activity. Its local units were charged with assuring moral and financial backing to candidates at election time, as well as with maintaining continual contact between elected officials and the electorate. The national organization endeavoured to unify the party members who had been elected to the assemblies. In general, the local committees maintained a basic autonomy and each legislator a large measure of independence. The party discipline in voting established by the British parties—which were older because of the fact that the British Parliament was long established—was imitated on the Continent hardly at all.

United States parties in the 19th century

The first United States political parties of the 19th century were not particularly different from European cadre parties, except that their confrontations were less violent and based less on ideology. The first United States form of the struggle between the aristocracy and the bourgeoisie, between conservative and liberal, was carried out in the form of the Revolutionary War, in which Great Britain embodied the power of the king and the nobility, and the insurgents that of bourgeoisie and liberalism. Such an interpretation is, of course, simplified. There were some aristocrats in the South and, in particular, an aristocratic spirit based on the institutions of slaveholding and paternalistic ownership of land. In this sense, the Civil War could be considered as a second phase of violent conflict between the conservatives and the liberals. Nevertheless, the United States was from the beginning an essentially bourgeois civilization, based on a deep sense of equality and of individual freedom. Federalists and Anti-Federalists, Republicans and Democrats—all belonged to the liberal family since all shared the same basic ideology and the same system of fundamental values and differed only in the means by which they would realize their beliefs.

In terms of party structure, United States parties in the beginning differed little from their European counterparts. Like them, the United States parties were composed of local notables. The ties of a local committee to a national organization were even weaker than in Europe. At the state level there was some effective coordination of local party organizations, but at the national level such coordination did not exist. A more original structure was developed after the Civil War—in the South to exploit the vote of the blacks and along the East Coast to control the votes of immigrants. The extreme decentralization in the United States enabled a party to establish a local quasi-dictatorship in a city or county by capturing all of the key posts in an election. Not only the position of mayor but also the police, finances, and the courts came under the control of the party machine, and the machine was thus a development of the original cadre parties. The local party committee came typically to be composed of adventurers or gangsters who wanted to control the distribution of wealth and to ensure the continuation of their control. These men were themselves controlled by the power of the boss, the political leader who controlled the machine at the city, county, or state levels. At the direction of the committee, each constituency was carefully divided, and every precinct was watched closely by an agent of the party, the captain, who was responsible for securing votes for the party. Various rewards were offered to voters in return for the promise of their votes. The machine could offer such inducements as union jobs, trader's licences, immunity from the police, and the like. Operating in this manner, a party could frequently guarantee a majority in an election to the candidates of its choosing, and, once it was in control of local government, of the police, the courts, and public finances, etc., the machine and its clients were assured of impunity in illicit activities such as prostitution and gambling rings and of the granting of public contracts to favoured businessmen.

The degeneration of the party mechanism was not without benefits. The European immigrant who arrived in the United States lost and isolated in a huge and different world might find work and lodging in return for his commitment to the party. In a system of almost pure capitalism and at a time when social services were practically nonexistent, machines and bosses took upon themselves responsibilities that were indispensable to community life. But the moral and material cost of such a system was very high, and the machine was often purely exploitative, performing no services to the community.

By the end of the 19th century the excesses of the machines and the bosses and the closed character of the parties led to the development of primary elections, in which party nominees for office were selected. The primary movement deprived party leaders of the right to dictate candidates for election. A majority of the states adopted the primary system in one form or another between 1900 and 1920. The aim of the system was to make the parties more democratic by opening them up to the general public in the hope of counterbalancing the influence of the party committees. In practice, the aim was not realized, for the committees retained the upper hand in the selection of candidates for the primaries.

The British Labour Party

In its original form the British Labour Party constituted a new type of cadre party, forming an intermediate link with the mass-based parties. It was formed with the support of trade unions and left-wing intellectuals. At the base, each local organization sent representatives to a district labour committee, which was in turn represented at the national congress.

The early (pre-1918) Labour Party was thus structured of many local and regional organizations. It was not possible to join the party directly; membership came only through an affiliated body, such as a trade union. It thus represented a new type of party, depending not upon highly political individuals brought together as a result of their desire to acquire and wield power but upon the organized representatives of a broader interest—the working class. Certain Christian Democrat parties—the Belgian Social Christian Party between the two world wars and the Austrian Popular Party, for example—had an analogous structure: a federation of unions, agricultural organizations, middle class movements, employers' associations, etc. After 1918, the Labour Party developed a policy of direct membership on the model of the continental Socialist parties, individual members being permitted to join local-constituency branches. The majority of its membership, however, continued to be affiliated rather than direct.

Mass-based parties. Cadre parties normally organize a relatively small number of party adherents. Mass-based parties, on the other hand, unite hundreds of thousands of followers, sometimes millions. But the number of members is not the only criterion of a mass-based party. The essential factor is that such a party attempts to base itself on an appeal to the masses. It attempts to organize not only those who are influential or well known or those who represent special interest groups but rather any citizen who is willing to join the party. If such a party succeeds in gathering only a few adherents, then it is mass based only in potential. It remains, nevertheless, different from the cadre-type parties.

Socialist mass-based parties

At the end of the 19th century the Socialist parties of continental Europe organized themselves on a mass basis in order to educate and to organize the growing population of labourers and wage earners, which industrialization was making increasingly large and which was becoming more important politically because of extensions of the suffrage, and to gather the money necessary for propaganda by mobilizing in a regular fashion the resources of those who, although poor, were numerous. Membership campaigns were conducted, and each member paid party dues. If its members became sufficiently numerous, the party emerged as a powerful organization, managing large funds and diffusing its ideas among an important segment of the population. Such was the case with the German Social Democratic Party, which by 1913 had more than 1,000,000 members.

Such organizations were necessarily rigidly structured. The party required an exact registration of membership, treasurers to collect dues, secretaries to call and lead local meetings, and a hierarchical framework for the coordination of the thousands of local sections. A tradition of collective action and group discipline, more developed among workers as a result of their participation in strikes and other union activity, favoured the development and centralization of party organization.

A complex party organization tends to give a great deal of influence to those who have responsibility at various levels in the hierarchy, resulting in certain oligarchical tendencies. The Socialist parties made an effort to control this tendency by developing democratic procedures in the choice of leaders. At every level those in responsible positions were elected by members of the party. Every local party group would elect delegates to regional and national congresses, at which party candidates and party leaders would be chosen and party policy decided.

The type of mass-based party described above was imitated by many non-Socialist parties. Some cadre-type parties in Europe, both conservative and liberal, attempted to transform themselves along similar lines. The Christian Democrat parties often developed organizations copied even more directly from the mass-based model. But non-Socialist parties were generally less successful in establishing rigid and disciplined organizations.

The first Communist parties were splinter groups of existing Socialist parties and at first adopted the organization of these parties. After 1924, as a result of a decision of the Comintern (the Third International, or federation of working class parties), all Communist parties were transformed along the lines of the Soviet model, becoming mass parties based on the membership of the largest possible number of citizens, although membership was and is limited to those who embraced and espoused the ideology of Marxism–Leninism.

Communist mass-based parties

The Communist parties developed a new structural organization: whereas the local committees of cadre and Socialist parties focussed their organizing efforts and drew their support from a particular geographical area, Communist groups formed their cells in the place of work. The workplace cell was the first original element in Communist party organization. It grouped together all party members who depended upon the same firm, workshop, or store or the same professional institution (school or university, for example). Party members thus tended to be tightly organized, their solidarity, resulting from a common occupation, being stronger than that based upon residence.

The work-place-cell system proved to be effective, and other parties tried to imitate it, generally without success. Such an organization leads each cell to concern itself with problems of a corporate and professional nature rather than with those of a more political nature. These basic groups, however, smaller and, therefore, more numerous than the Socialist sections, tend to go their separate way. It is necessary to have a very strong party structure and for party leaders to have extensive authority if the groups are to resist such centrifugal pressure.

This has resulted in a second distinctive characteristic of the Communist parties: a high degree of centralization. Although all mass-based parties tend to be centralized, Communist parties are more so than others. There is, in principle, free discussion, which is supposedly developed at every level before a decision is made, but afterward all must adhere to the decision that has been made by the central body. The splintering that has from time to time divided or paralyzed the Socialist parties is forbidden in Communist parties, which have generally succeeded in maintaining their unity. A further distinctive characteristic of Communist parties is the importance given to ideology. All parties have a doctrine or at least a platform. The European Socialist parties, which were doctrinaire before 1914 and between the two wars, later became more pragmatic, not to say opportunistic. But, in Communist parties, ideology occupies a much more fundamental place, a primary concern of the party being to indoctrinate its members with Marxism.

The Fascist model

The 1920s and '30s saw the emergence of Fascist parties that attempted, as do the Communist and Socialist parties, to organize the maximum number of members but that did not claim to represent the great masses of people. Their teaching was authoritarian and elitist. They thought that societies should be directed by the most talented and capable people—by an elite. The party leadership, grouped under the absolute authority of a supreme head, constituted such an elite. Party structure had as its goal the assurance of the obedience of the elite.

This structure resembled that of armies, which are also organized in such a way as to ensure, by means of rigorous discipline, the obedience of a large number of men to an elite leadership. The party structure, therefore, made use of a military-type organization, consisting of a pyramid made up of units that at the base were very tiny but that, when joined with other units, formed groups that got larger and larger. Uniforms, ranks, orders, salutes, marches, and unquestioning obedience were all aspects of Fascist parties. This similarity rests upon another factor; namely, that Fascist doctrine taught that power must be seized by organized minorities making use of force. The party thus made use of a militia intended to assure victory in the struggle for control over the unorganized masses.

Large parties built upon the Fascist model developed between the two wars in Italy and Germany, where they actually came into power. Fascist parties appeared also in most other countries of western Europe during this period but were unable to achieve power. The less-developed nations of eastern Europe and Latin America were equally infected by the movement. The victory of the Allies in 1945, as well as the revelation of the horrors of Nazism, stopped the growth of the Fascists and provoked their decline, but Fascist-type party organization and doctrine remain a potent means of exercising power.

PARTIES AND POLITICAL POWER

Whether they are conservative or revolutionary, whether they are a union of notables or an organization of the masses, whether they function in a pluralistic democracy or in a monolithic dictatorship, parties have one function in common: they all participate to some extent in the exercise of political power, whether by forming a government or by exercising the function of opposition, a function that is often of crucial importance in the determination of national policy.

The struggle for power. It is possible in theory to distinguish revolutionary parties, which attempt to gain power by violence (conspiracies, guerrilla warfare, etc.), from those parties working within the legal framework of elections. But the distinction is not always easy to make,

because the same parties may sometimes make use of both procedures, either simultaneously or successively, depending upon the circumstances. In the 1920s, for example, Communist parties sought power through elections at the same time that they were developing an underground activity of a revolutionary nature. In the 19th century, liberal parties were in the same situation, sometimes employing the techniques of conspiracy, as in Italy, Austria, Germany, Poland, and Russia, and sometimes confining their struggles to the ballot box, as in Great Britain and France.

Revolutionary methods vary greatly. Clandestine plots by which minority groups seize the centres of power presuppose monarchies or dictatorships in which the masses of people have little say in government. But terrorist and disruptive activity can serve to mobilize citizens and to demonstrate the powerlessness of any government. At the beginning of the 20th century leftist trade unionists extolled the revolutionary general strike, a total stoppage of all economic activity that would paralyze society completely and put the government at the revolutionaries' mercy. Rural guerrilla activity has often been used in countries with a predominantly agrarian society; urban guerrilla warfare was effective in the European revolutions of the 19th century, but the development of techniques of police and military control has made such activity more difficult.

Functions of parties within a democracy

Revolutionary parties are less numerous than parties that work within the law: the contest at election time is the means normally used in the struggle for power. Such activity corresponds, morever, to the original nature of political parties and involves three factors: the organization of propaganda, the selection of candidates, and the financing of campaigns. The first function is the most visible. The party first of all gives the candidate a label that serves to introduce him to the voters and to identify his position. Because of this party label the voters are better able to distinguish the candidates. The promises and declarations of individuals are seldom taken with too much seriousness, and it means more to indicate that one candidate is a Communist, another a Socialist, a third a Fascist, and a fourth a liberal. Finally, the party also furnishes the candidate with workers to raise funds, put up his posters, distribute his literature, organize his meetings, and canvass from door to door.

The function of selecting candidates is exercised in three ways. In cadre parties, candidates are selected by committees of the party activists who make up the party—the caucus system, as it is known in the United States. In general, local committees play essential roles in this regard. In some countries, however, the selection is centralized by a national caucus, as, for example, by the Conservative Party in Britain and the Union of Democrats in France. In mass-based parties, selection is made by members of the regional and national congresses according to apparently democratic procedures; in actual practice, the governing committees play an essential role, the local constituency members generally ratifying their choice. Thirdly, in the United States the mechanism of primary elections has established a system for selecting candidates by means of the votes of all party members or all voters within a particular electoral district.

The various processes of selecting candidates do not, however, differ significantly in their results, for it is almost always the party leaders who play the essential role. This introduces an oligarchical tendency into party politics, a tendency that has not been overcome by the congresses of the mass-based parties or the United States primaries, which provide only a partial limitation on the power of the governing committees.

An important aspect of the struggle for power between political parties is the financing of campaigns. Cadre parties always have in their committees some key figure having connections with businessmen who is responsible for collecting gifts from them. In mass-based parties, rather than looking for large sums of money from a few people, leaders gather smaller sums from a large number of people who usually give on a monthly or annual basis. This method has been viewed as one of the distinguishing characteristics of mass-based parties. Sometimes the law intervenes in the financing of elections and of parties. Laws often limit campaign expenses and attempt to restrict the resources of the parties, but they are generally inoperative because it is quite easy to circumvent them. In some countries the state contributes public funds to the parties. At first, such financial participation was limited to expenses for campaigns and was based on the uniform treatment of candidates (as in France), but in Sweden and Finland the state contributes to the general finances of parties.

Participation in power. Only the functions of parties in democratic regimes will be considered at this time. The role of the single party in a dictatorship will be analyzed separately.

Once a political party has achieved electoral victory, the question arises of how much influence the party is to have on the government. The influence of the party on members in elective office is frequently quite weak. It defines the general lines of their activity, but these lines can be quite hazy, and few decisions are taken in the periodic meetings between officeholders and their party. Each member of the legislature retains personal freedom of action in his participation in debates, in his participation in government, and, especially, in his voting. The party may, of course, attempt to enforce the party line, but parliamentary or congressional members cannot be compelled to vote the way the party wants them to. Such is the situation in the United States, within most of the liberal and conservative European parties, and within cadre parties in general.

Flexible and rigid parties

The question of how disciplined a party is, of the extent to which it will always present a united front, enables a distinction to be made between what may be termed rigid and flexible parties; that is, between those that attempt always to be united and disciplined, following what is most often an ideologically based party line, and those that, representing a broader range of interests and points of view, form legislatures that are assemblies of individuals rather than of parties.

Whether the parties operating within a particular system will be rigid or flexible depends largely on the constitutional provisions that determine the circumstances in which a government may continue in office. This is clearly illustrated by comparing the situation in the United States with that in Great Britain. In the United States the president and his government continue in office for the constitutionally defined period of four years, regardless of whether a majority in the legislature supports him or not. Since a united party is thus not crucial to the immediate survival of the government, both major parties are able to contain broad coalitions of interests, and votes on issues of major importance frequently split each party. In the United Kingdom the situation is quite different. There, government can continue in office only so long as it commands a majority in the legislature. A single adverse vote can result in the dissolution of Parliament and a general election. Party discipline and unity are thus of crucial importance, and this fact has far-reaching consequences for the composition, organization, and policies of each party. The consequences of party disunity within such a constitutional framework are well illustrated by the weakness and instability of the governments of the Third and Fourth French republics.

The distinction between flexible and rigid parties applies equally to parties in power and to those forming the opposition. Votes of censure or of lack of confidence, votes on proposed legislation or on the budget, questions put to ministers or challenges made to them—in short, all the functions of an opposition party—are worked out differently in flexible and rigid party systems.

In flexible party systems the absence of strong discipline is often of great consequence to the opposition party because only rigid parties can constitute an opposition force sufficiently strong to counterbalance the strength of the party in power. At the same time, party discipline permits the opposition to present the public with an alternative to the majority party; the logical consequence of such a situation is Britain's "shadow cabinet," which accustoms the electorate to the idea that a new group is ready to take over the reins of government.

Parties provide, moreover, a channel of communication

between opposition legislators and the public. The governing party performs a similar service for the government, although it is less necessary, since the government has at its disposal numerous means of communicating with the public. Opposition parties thus provide a means of expressing negative reaction to decisions of government and proposing alternatives. This role justifies the official recognition given to opposition parties, as is the case in Great Britain and Scandinavia.

Power and representation. It is difficult to envisage how representative democracy could function in a large industrialized society without political parties. In order for citizens to be able to make an intelligent choice of representative or president, it is necessary for them to know the real political orientation of each candidate. Party membership provides the clearest indication of this. The programs and promises of each individual candidate are not too significant or informative, because most candidates, in their attempt to gain the most votes, try to avoid difficult subjects; they all tend to speak the same language; that is, to camouflage their real opinions. The fact that one is a Socialist, another a conservative, a third a liberal, and a fourth a Communist provides a far better clue as to how the candidate will perform when in office. In the legislature the discipline of the party limits the possibility that elected representatives will change their minds and their politics, and thus the party label acts as a sort of guarantee that there will be at least some correspondence between promise and performance. Parties make possible the representation of varying shades of opinion by synthesizing different positions into a stance that each representative adopts to a greater or lesser extent.

Individual candidates and the party

But parties, like all organizations, tend to manipulate their members, to bring them under the control of an inner circle of leaders that often perpetuates itself by cooptation. In cadre parties, members are manipulated by powerful committees containing cliques of influential party leaders. In mass-based parties, leaders are chosen by the members, but incumbents are very often re-elected because they control the party apparatus, using it to ensure their continuation in power.

Democratic political systems, while performing the function of representation, thus rest more or less on the competition of rival oligarchies. But these oligarchies consist of political elites that are open to all with political ambition. No modern democracy could function without parties, the oligarchical tendencies of which are best regarded as a necessary evil.

PARTY SYSTEMS

Party systems may be broken down into three broad categories: two-party, multiparty, and single-party. Such a classification is based not merely on the number of parties operating within a particular country but on a variety of distinctive features that the three systems exhibit. Two-party and multiparty systems represent means of organizing political conflict within pluralistic societies and are thus part of the apparatus of democracy. Single parties usually operate in situations in which genuine political conflict is not tolerated. This broad statement is, however, subject to qualification, for, although single parties do not usually permit the expression of points of view that are fundamentally opposed to the party line or ideology, there may well be intense conflict within these limits over policy within the party itself. And even within a two-party or a multiparty system, debate may become so stymied and a particular coalition of interests so entrenched that the democratic process is seriously compromised.

The distinction between two-party and multiparty systems is not as easily made as it might appear. In any two-party system there are invariably some tiny parties in addition to the two major parties, and there is always the possibility that such small parties might prevent one of the two main parties from gaining a majority of seats in the legislature. This is the case with regard to the Liberal Party in Great Britain, for example. Other countries do not fall clearly into either category; thus, Austria and the Federal Republic of Germany only approximate the two-party system. It is not simply a question of the number of parties that determines the nature of the two-party system; many other elements are of importance, the extent of party discipline in particular.

Multiparty systems. In Anglo-Saxon countries there is a tendency to consider the two-party system as normal and the multiparty system as the exceptional case. But, in fact, the two-party system that operates in Great Britain, the United States, and New Zealand is much rarer than the multiparty system, which is found in almost all of western Europe.

In western Europe, three major categories of parties have developed since the beginning of the 19th century: conservative, liberal, and Socialist. Each reflects the interests of a particular social class and expounds a particular political ideology. After World War I other categories of parties developed that were partly the result of divisions or transformations of older parties. Communist parties began as splinter groups of Socialist parties, and Christian Democrat parties attempted to weld together moderate Socialists and conservatives and some liberals. Other distinctive types of party emerged in some countries. In Scandinavia, liberal rural parties developed in the 19th century, reflecting a long tradition of separate representation of the rural population. In many countries ethnic minorities formed the basis of nationalist parties, which then either joined existing parties or divided them.

The appearance of Socialism in the 19th century upset the earlier lines of battle between conservatives and liberals and tended to throw the latter two groups into a common defense of capitalism. Logically, this situation should have led to the fusion of conservatives and liberals into one bourgeois party that would have presented a united stand against Socialism. This is, in fact, what happened in Great Britain after World War I.

Influence of the electoral system on the number of parties

One of the most important factors determining the number of parties operating within a particular country is the electoral system. Proportional representation tends to favour the development of multiparty systems because it ensures representation in the legislature for even small parties. The majority, single-ballot system tends to produce a two-party system, because it excludes parties that may gain substantial numbers of votes but not the majority of votes necessary to elect a representative within a constituency. The majority system with a second ballot favours a multiparty system tempered by alliances between parties. Such a system is very rare, found only in the German Empire (1871–1914) and in the French Third (1870–1940) and Fifth (since 1958) republics. Voters choose between the parties that did best in a first ballot. This leaves small parties at a disadvantge but, nevertheless, gives them opportunity to strengthen their role during the second balloting as long as they are willing to enter into alliances with the leading parties.

Another factor producing multiparty systems is the intensity of political conflicts. If, within a given political movement, extremists are numerous, then it is difficult for the moderates in that party to join with them in a united front. Two rival parties are likely to be formed. Thus, the power of the Jacobins among 19th-century French liberals contributed to the inability of the moderates to form one great liberal party, as was successfully achieved in Great Britain. Likewise, the power of the extremists among the conservatives was an obstacle to the development of a strong conservative party.

The distinction between the multiparty system and the two-party system corresponds largely to a distinction between two types of Western political regime. In a two-party situation the administration has, in effect, an assurance of a majority in the legislature, deriving from the predominance of one party; it has, therefore, a guarantee of continuance and effectiveness. Such a system is often referred to as majority parliamentarianism. In a multiparty situation, on the other hand, it is quite rare for one party to have a majority in the legislature; governments must, therefore, be founded on coalitions, which are always more heterogeneous and more fragile than a single party. The result is less stability and less political power. Such systems may be referred to as nonmajority parliamentarianism.

In practice, majority and nonmajority parliamentary systems do not coincide exactly with two-party systems and multiparty systems. For, if each of the two parties is flexible and does not control the voting patterns of its members (as is the case in the United States), the numerical majority of one of the parties matters little. It can happen, moreover, that one party in a multiparty system will hold an absolute majority of seats in the legislature so that no coalition is required. Such a situation is unusual but did occur in West Germany, Italy, and Belgium at various times after 1945.

Ordinarily, however, a coalition will be the only means of attaining a parliamentary majority within the framework of the multiparty system. Coalitions are by nature more heterogeneous and more unstable than a grouping made up of one party, but their effectiveness varies greatly according to the discipline and organization of the parties involved. In the case of flexible parties that are undisciplined and that allow each legislator to vote on his own, the coalition will be weak and probably short-lived. The instability and weakness of governments is at its maximum in such situations, of which the Third French Republic provides a good example.

If, on the other hand, the parties involved in a coalition are rigid and disciplined, it is possible for a system quite similar to the two-party system to develop. This is often the case when two opposing alliances are formed, one on the left and one on the right, and when both are strong enough to endure through the legislative session. This type of coalition, referred to as bipolarized, introduces elements of the two-party system into a multiparty framework. A situation of this type has developed in Sweden, where conservative, liberal, and agrarian parties have been aligned against the Social Democrat Party, which eventually allied itself with the Communist Party (1970).

The system of bipolar alliances may be contrasted with the system of a centrist alliance. Rather than the parties on the right forming a centre–right coalition to oppose a centre–left coalition, there is the possibility that the centre–left and the centre–right will join forces and reject the extremes at both ends of the political spectrum. Such a situation occurred in Germany during the Weimar Republic, when the government rested on a majority formed of a coalition of Catholic Centrists and Social Democrats, with opposition coming from the Communists and the nationalists on the extreme left and right.

Benefits and disadvantages of coalition politics

Centrist coalitions all tend to give the average citizen a sense of political alienation. In rejecting both extremes, coalitions may well be isolating the radical, unstable elements, but the governing coalition may tend to be unresponsive to new ideas, uninspiringly pragmatic, and too ready to compromise. This situation gives rise to a more or less permanent breach between practical politics and political ideals. An advantage of bipolarization or of the two-party system is that the moderates of both sides must collaborate with those who are more extreme in their views, and the extremists must be willing to work with those who are more moderate; the pressure from the extremists prevents the moderates from getting bogged down, while collaboration with the moderates lends a touch of realism to the policies of the extremists.

Two-party systems. A fundamental distinction must be made between the two-party system as it is found in the United States and as it is found in Great Britain. Although two major parties dominate political life in the two countries, the system operates in quite different ways.

The American two-party system. The United States has always had a two-party system, first in the opposition between the Federalists and the Anti-Federalists, then in the competition between the Republicans and the Democrats. There have been frequent third-party movements in the history of the country, but they have always failed. Presidential elections seem to have played an important role in the formation of this type of two-party system. The mechanism of a national election in so large a country has necessitated very large political organizations and, at the same time, relatively simplified choices for the voter.

American parties are different from their counterparts in other Western countries. They are not tied in the same way to the great social and ideological movements that have so influenced the development of political life in Europe during the last two centuries. There have been Socialist parties at various times in the history of the United States, but they have never challenged the dominance of the two major parties. It can be argued that the main reason for the failure of Socialist parties in America has been the high degree of upward mobility permitted by a rich and continually expanding economy. The consequence of this mobility has been that class consciousness has never developed in the United States in a manner that would encourage the formation of large Socialist or Communist parties.

In comparison with European political movements, therefore, American parties have appeared as two varieties of one liberal party, and within each party can be found a wide range of opinion, going from the right to the left.

Flexible American party structure

The American parties have a flexible and decentralized structure, marked by the absence of discipline and rigid hierarchy. This was the structure of most of the cadre-type parties of the 19th century, a structure that most liberal parties have retained. Federalism and a concern for local autonomy accentuate the lack of rigid structure and the weakness of lines of authority in the parties. Organization may be relatively strong and homogeneous at the local level, but such control is much weaker on the state level and practically nonexistent on the national level. There is some truth to the observation that the United States has not two parties but 100—that is, two in each state. But it is also true that each party develops a certain degree of national unity for the presidential election and that the leadership of the president within his party gives the victorious party some cohesion.

In voting, Republicans and Democrats are usually found on both sides. An alliance between liberal Republicans and Democrats against conservative Republicans and Democrats tends to develop. But neither bloc is stable, and the alignment varies from one vote to another. As a consequence, despite the existence of a two-party system, no stable legislative majority is possible. In order to have his budget adopted and his legislation passed, the president of the United States must carefully try to gather the necessary votes on every question, bearing the wearisome task of constantly forming alliances. The American two-party system is thus a pseudo-two-party system, because each party provides only a loose framework within which shifting coalitions are formed.

The British two-party system. Another form of the two-party system is operative in Great Britain and in New Zealand. The situation in Australia is affected somewhat by the presence of a third party, the Country Party. A tight alliance between the Australian Liberal Party and the Country Party introduces, however, a rather rigid bipolarization with the Labour Party. The system thus tends to operate on a two-party basis. Canada also possesses what is essentially a two-party system, Liberals or Conservatives usually being able to form a working majority without the help of the small, regionally based parties.

Great Britain has had two successive two-party alignments: Conservative and Liberal prior to 1914 and Conservative and Labour since 1935. The period from 1920 to 1935 constituted an intermediate phase between the two. Britain's Conservative Party is actually a Conservative–Liberal Party, resulting from a fusion of the essential elements of the two great 19th-century parties. Despite the name Conservative, its ideology corresponds to political and economic liberalism. A similar observation could be made about the other major European conservative parties, such as the German Christian Democratic Party and the Belgian Social Christian Party.

Rigid party lines in the British system

The British two-party system depends on the existence of rigid parties; that is, parties in which there is effective discipline regarding parliamentary voting patterns. In every important vote, all party members are required to vote as a bloc and to follow to the letter the directives that they agreed upon collectively or that were decided for them by the party leaders. A relative flexibility may at times be tolerated, but only to the extent that such a policy does not compromise the action of the government. It may

be admissible for some party members to abstain from voting if their abstention does not alter the results of the vote. Thus, the leader of the majority party (who is at the same time the prime minister) is likely to remain in power throughout the session of Parliament, and the legislation he or she proposes will likely be adopted. There is no longer any real separation of power between the executive and legislative branches, for the government and its parliamentary majority form a homogeneous and solid bloc before which the opposition has no power other than to make its criticisms known. During the four or five years for which a Parliament meets, the majority in power is completely in control, and only internal difficulties within the majority party can limit its power.

Since each party is made up of a disciplined group with a recognized leader who becomes prime minister if his or her party wins the legislative elections, these elections perform the function of selecting both the legislature and the government. In voting to make one of the party leaders the head of the government, the British assure the leader of a disciplined parliamentary majority. The result is a political system that is at once stable, democratic, and strong; and many would argue that it is more stable, more democratic, and stronger than systems anywhere else.

This situation presupposes that both parties are in agreement with regard to the fundamental rules of a democracy. If a Fascist party and a Communist party were opposed to one another in Great Britain, the two-party system would not last very long. The winner would zealously suppress the opponent and rule alone.

The system, of course, does have its weak points, especially insofar as it tends to frustrate the innovative elements within both parties. But it is possible that this situation is preferable to what would happen if the more extreme elements within the parties were permitted to engage in unrealistic policies. The risk of immobility is in fact a problem for any party in a modern industrial society, and not just for those in a two-party situation. The problem is related to the difficulties involved in creating new organizations capable of being taken seriously by an important segment of the population and in revitalizing long-standing organizations encumbered by established practices and entrenched interests.

Single-party systems. There have been three historical forms of the single-party system: Communist, Fascist, and that found in the developing countries.

The Communist model. In Communist countries the party is considered to be the spearhead of the urban working class and of other workers united with it (peasants, intellectuals, etc.). Its role is to aid in the building of a Socialist regime during the transitory phase between capitalism and pure Socialism, called the dictatorship of the proletariat. An understanding of the exact role of the party requires an appreciation of the Marxist conception of the evolution of the state. In countries based on private ownership of the means of production, the power of the state, according to the Marxist point of view, is used to further the interests of the controlling capitalists. In the first stage of revolution the power of the state is broken. Power, however, still has to be wielded to prevent counter-revolution and to facilitate the transition to Communism, at which stage coercion will no longer be necessary. Thus, the party, in effect, assumes the coercive functions of the state during the dictatorship of the proletariat or, to be more accurate, during the dictatorship of the party in the name of the proletariat.

Role of the party in Communist states

In all Communist countries, the structure of the party has been determined largely by the need for it to govern firmly while at the same time maintaining its contact with the masses of the people. Party members are a part of the general public, of which they are the most active and most politically conscious members. They remain in contact with the masses by means of a network of party cells that are present everywhere. Party leaders are thus always "listening in on the masses," and the masses are always informed of decisions of party leaders, as long as the communication network is working in both directions.

The party is not only a permanent means of contact between the people and party leaders but also a propaganda instrument. Political indoctrination is essential to the survival of Communist parties, and many resources are devoted to it. Indoctrination is accomplished in training schools, by means of "education" campaigns, by censorship, and through the untiring efforts of militants, who play a role similar to that of the clergy in organized religion. The party is thus the guardian of orthodoxy and has the power to condemn and to excommunicate.

In the traditional Communist model, the party hierarchy, then, and not the official state hierarchy, has the real power. The first secretary of the party is the most important figure of the regime, and, whether the party leadership is in the hands of one man or several, the party remains the centre of political power.

Near the end of the 20th century, however, the Communist model began to change as the centre of power began shifting toward a popularly elected state hierarchy. A younger generation of Communist leaders, openly critical of the party's inefficient, unresponsive, and domineering management of the government—particularly the economy—sought a return to Lenin's original concepts of democratic centralism and socialism. In some countries, democratic concepts were emphasized, and constitutional amendments eliminated the party's official control, clearing the way for a multiparty system.

The Fascist model. Fascist parties in a single-party state have never played as important a role as Communist parties in an analogous situation. In Italy, the Fascist party was never the single most important element in the regime, and its influence was often secondary. In Spain the Falange never played a crucial role, and in Portugal the National Union was a very weak organization even at the height of dictator António Salazar's strength. Only in Germany did the National Socialist Party have a great influence on the state. But, in the end, Hitler's dictatorship was dependent on his private army, the SS (Schutzstaffel), which formed a separate element within the party and which was closed to outside influences, and on the Gestapo, which was a state organization and not an organization of the party. The Fascist party in the single-party state has a policing or military function rather than an ideological one.

Functions of Fascist parties

After their rise to power, the Fascist parties in both Germany and Italy gradually ceased to perform the function of maintaining contact between the people and the government, a function that is usually performed by the party in a single-party situation. It was possible to observe a tendency for the party to close in upon itself while suppressing its deviant members. The renewal of the party was then assured through recruitment from youth organizations, from which the most fanatical elements, the products of a gradual selection process starting at a very early age, entered the party. The party tended, therefore, to constitute a closed order.

The single party in the developing countries. Some of the Communist parties in power in developing countries do not differ significantly from their counterparts in industrialized countries. This is certainly true of the Communist Party of the Socialist Republic of Vietnam and the Workers' Party of North Korea. There have always been, however, countries in which the single party in power could not be characterized in terms of a traditional European counterpart. This observation applies to, for example, the former Arab Socialist Union in Egypt, the Neo-Destour Party in Tunisia (renamed the Destour Democratic Rally), and the National Liberation Front in Algeria, as well as many other parties in black Africa.

Most of these parties claimed to be more or less Socialist or at least progressive, while remaining far removed from Communism and, in some cases, ardent foes of Communism. President Nasser attempted to establish a moderate and nationalistic Socialism in Egypt. In Tunisia the Neo-Destour Party was more republican than Socialist and was inspired more by the example of the reforms in Turkey under Kemal Atatürk than by Nasserism. In black Africa, single parties have often claimed to be Socialist, but with few exceptions they rarely are in practice.

Single parties in developing countries are rarely as well organized as Communist parties. In Turkey the Republican People's Party was more a cadre party than a mass-

based party. In Egypt it has been necessary to organize a core of professional politicians within the framework of a pseudoparty of the masses. In sub-Saharan Africa the parties are most often genuinely mass based, but the membership appears to be motivated primarily by personal attachment to the leader or by tribal loyalties, and organization is not usually very strong. It is this weakness in organization that explains the secondary role played by such parties in government.

Some regimes, however, have endeavoured to develop the role of the party to the fullest extent possible. The politics of Atatürk in Turkey were an interesting case study in this regard. It was also Nasser's goal to increase the influence of the Arab Socialist Union, thereby making it the backbone of the regime. This process is significant in that it represents an attempt to move away from the traditional dictatorship, supported by the army or based on tribal traditions or on charismatic leadership, toward a modern dictatorship, supported by one political party. Single-party systems can institutionalize dictatorships by making them survive the life of one dominant figure.

FUTURE OF POLITICAL PARTIES

The necessity of parties in democratic systems

It has often been said in the West that political parties are in a state of decline. Actually, this has been a long-standing opinion in certain conservative circles, arising largely out of a latent hostility to parties, which are viewed as a divisive force among citizens, a threat to national unity, and an enticement to corruption and demagoguery. In certain European countries—France, for example—right-wing political organizations have even refused to call themselves parties, using instead such terms as movement, union, federation, and centre. And it cannot be denied that to some extent the major European and American parties of the late 20th century do appear old and rigid in comparison with their condition at the turn of the century or immediately following World War I. Even relatively new parties, such as the Christian Democratic parties of Germany and Italy (founded in 1945), seem somewhat lifeless.

In terms of size and number, however, political parties are not declining but growing. At the turn of the century they were confined mainly to Europe and North America; elsewhere they were quite weak or nonexistent. In the late 20th century, parties are found practically everywhere in the world. And in Europe and North America there are generally far more people holding membership in parties than prior to 1914. Parties of the late 20th century are larger, stronger, and better organized than those of the late 19th century. In the industrialized nations, especially in western Europe, parties have become less revolutionary and innovative, and this factor may explain the rigid and worn-out image that they sometimes present. But even this phenomenon is found only in a limited area and may, perhaps, pass.

The growth of parties into very large organizations may be responsible for the feelings of powerlessness on the part of many individuals who are involved in them. This is a problem experienced by people who find themselves part of any large organization, whether it be a political party, business enterprise, corporation, or union. The difficulties involved in reforming or changing political parties that have become large and institutionalized, coupled with the next-to-impossible task of creating new parties likely to reach sufficient strength to be taken seriously by the electorate, have resulted in much frustration and impatience with the party system. But it is difficult to imagine how democracy could function in a large industrialized country without political parties. In the modern world, democracy and political parties are two facets of the same reality, the inside and outside of the same fabric. (M.Du./Ed.)

Interest groups

TYPES OF INTEREST GROUP

Categories defined by purpose. Within the broad definition of interest groups, two polar types are recognizable: first, interest groups proper, such as trade unions, farmers' unions, and employers' associations, which have as their primary purpose the enhancement of the advantage of their members; and, second, promotional groups, such as the societies for the prevention of cruelty to children or various voluntary relief agencies, which exist primarily or entirely to enhance the advantage not of their own members but of the population, even perhaps to the discomfort or disadvantage of their own members. Some of these, such as churches or various evangelizing groups, exist to promulgate a distinctive set of values to be applied to society as a whole.

The distinction between interest groups proper and promotional groups is not sharp for two reasons. In the first place, most interest groups proper sincerely believe that in furthering their own material advantage they are also serving that of society as a whole—by promoting "free enterprise" or a healthy and wealthy body of farmers or a well-paid and enthusiastic corps of schoolteachers. But the propagation of such beliefs is certainly not the primary purpose of such organizations, thus distinguishing them from promotional groups. The second reason is that some groups—for instance, organized churches—fall between the two types; they can simultaneously pursue advantages for their own sect and seek to inculcate a distinctive set of values in a whole society.

Natural and associational groups

Categories defined by structure. In primitive or developing societies, the most prominent type of interest group is the natural (*i.e.*, primordial or communal) one—that is, one based on kinship, lineage, neighbourhood, or religious confession. In Western, industrialized societies, though such groups do sometimes retain influence (as do the nationalities in Latvia and Belgium), the most prominent interest group is the associational (*i.e.*, secondary or factitious) type, like the trade union or Campaign for Nuclear Disarmament, which is deliberately created to serve defined purposes. Often, associational (or factitious) groups are created for the sake of the specialized purposes of the first type; the Indian Workers Union, for instance, was created from the community of Indian persons in Britain.

Within the class of associational groups, further qualifications must be made:

1. Not all associational groups possess formal structure. "Wall Street" in the United States, or the "City" in London, though consisting of a loose network of persons or functions, may nevertheless exert powerful collective pressures.

2. Some associational groups are collectivities; but others are single, discrete organizations. Thus General Motors Corporation exerts its own strong influence vis-à-vis the U.S. government irrespective of the influence of the more general employers' organization to which it possibly belongs, like the National Association of Manufacturers.

3. Some associational groups are temporary or *ad hoc*. Such are the so-called anomic groups, such as enraged French farmers who come together briefly to put logs across roads in order to draw governmental attention to their grievances.

4. Some associational groups are "latent." There may be a common interest among certain members of the public even though these individuals may not have combined into a formal or informal organization. The individuals may be unaware that they have a common interest or, if aware, see no reason to defend or promote it. Or, even if the members consciously wish to defend or promote an interest, the laws may restrict or control their ability to associate for such a purpose.

Pressure groups

Categories defined by political activity. If or when any interest group tries to influence the government in the pursuit of its aims, it becomes a "pressure group." Not all interest groups proper try to exert influence on governments; many (such as businesses and trade unions) do so only as a part of their more general activities of promoting their own interests. Promotional groups, on the other hand, spend much or all their time precisely in trying to influence the government to favour their aims. Groups such as an anglers' association may, therefore, turn into pressure groups for a time—the time when they seek to influence the government for one of their purposes—and then revert to simple interest-group status. A pressure group is therefore definable as "any interest group that

is not a part of the government and does not itself seek to govern the country in its own name, but does seek to influence that government for its own purposes."

Difficulties of definition arise, however. Some groups are neither a governmental agency nor entirely a private group. In autocratic states, Communist or otherwise, for instance, trade unions are often controlled by the government or the governing party. In the Western liberal democracies, again, agencies like the Tennessee Valley Authority (in the U.S.) or the British Broadcasting Corporation, though subject to the overriding control of the government, enjoy substantial autonomy in certain broad areas.

Also, as noted at the beginning of this article, some pressure groups eventually turn into political parties, or vice versa.

CHARACTERISTICS OF PRESSURE GROUPS

Power elements. Even in the most pluralistic of the Western liberal democratic states, it is wrong to picture public policy as simply the result of a parallelogram of group forces. The groups are themselves restrained by institutions, procedures, and public beliefs. Among the *institutions,* the most important are the public bureaucracy, the political parties, and the independent branches of the government (the executive, legislature, and judiciary—or some combination of these, depending on the structure of government). To the extent that each or all of these institutions have strong traditions, no interest group supplants them but instead has to deal with them. The governmental procedures in force also affect groups' performance. To the extent that public issues are traditionally subject to publicity and wide discussion, groups are limited in the kind of activities that they can pursue. They are constrained by public *beliefs* as to what is in "the public interest," by what is a proper or improper procedure, or by what causes are respectable and what are not. If any of these three entities—institutions, procedures, or public beliefs—become feeble, interest groups can intrude into them. By allying with or gaining power over the parties, the legislature, or the bureaucracy, for instance, groups can tend to substitute themselves for such organs in the decision-making process.

Inside every country the power that pressure groups wield against the government and that each group wields relative to its rivals depends on a combination of the following variables, some of which are mutually dependent:

1. Density is the ratio of actual membership to potential membership. The higher the density, the more "representative" is the group of all the people whose interests it purports to represent and the greater will be the inclination for governments to recognize and consult it. On the other hand, too wide a membership brings with it internal cleavages, so that on certain issues the organization may not be able to tender clear-cut advice or, indeed, may have to refrain from offering advice altogether.

2. Wealth not only helps a group to develop a skilled management and bureaucracy but also enables it to propagandize and to finance political parties.

3. Prestige lends a group the ability to get a favourable initial hearing from government and public, even though its numbers may be small and its wealth meagre.

4. Organization consists notably of the ability to brief legislators and administrators well and quickly, to mobilize members and the public rapidly, and to receive advance intelligence of likely trends in policy.

5. Socioeconomic leverage is strong among some groups (such as trade unions) that can disrupt social life and low among others (such as consumers associations).

6. Militancy consists of making an effective nuisance of oneself, so that, hopefully, governments will be willing to "buy" time or peace.

7. Specialized information and skills add the weight of authority.

8. Electoral strength refers to the power of some groups that, though poorly organized or having a low density, nevertheless command wide support in the electorate and so find themselves courted by rival political parties.

Strategy and tactics. Although all groups seek out the most influential agencies of government, the most influential may not be accessible; and resort must be had to the next most influential organ that *is* accessible. Because the Communist trade unions in Italy, for instance, are not welcome at the Roman Catholic controlled Ministry of Labour, they try to penetrate the legislature through political election. Indeed, in any system the influence of the various organs of government varies. In France the president and bureaucracy are more influential than the majority party and Parliament; in the United States, the president and Congress balance each other, with control of the bureaucracy shared between the two. Strategy, for an interest group, consists of determining and going to the most influential organ that is accessible.

Tactics to affect the executive

The tactics of an interest group range from the constitutional to the unconstitutional (or "direct action"), the legal to the illegal—all subject, as noted above, to the restraints of institutions, public procedures, and public beliefs. When the target of the tactics is the executive and the governmental bureaucracy, the following are the prime methods: *Constitutional operations* include (1) advice via advisory bodies attached to the ministries (as in Britain and France), (2) official hearings outside the ministries (as in Royal Commissions in the U.K. and presidential commissions in the U.S.) or inside the ministries (as in Germany or Sweden), and (3) ad hoc consultations and cooperation. *Semiconstitutional operations* include (1) recruitment of state civil servants into the private bureaucracies of the interest groups, or vice versa, and (2) bribery and favours, which need not necessarily mean passing of money but can take the form of providing entertainment, gifts, or meals. *Semidirect operations* include such measures as refusing to provide information or to cooperate in administering legislation and thus generally withdrawing from advisory functions. *Direct action* includes such measures as the withdrawal of labour (the British Medical Association, for instance, advised its members to withdraw from the state medical service in 1965 unless its demands were met) or even violence such as the roadblocking activities of enraged French farmers in the 1960s).

Tactics to affect the legislative

When the target is the legislature, the following are the chief kinds of tactics used by interest groups: *Constitutional operations* include (1) testimony before legislative committee hearings (extremely important in the U.S. and somewhat important in Sweden, the German Federal Republic, Japan, and Italy); (2) direct representation in the legislature, which involves having individual businessmen, financiers, farmers, school teachers, journalists, and other professionals representing the larger interests of their profession or sponsoring group (in Great Britain in 1955 some 30 percent of Labour's MP's were trade-union sponsored); and (3) lobbying, which refers to the activities of salaried or nonsalaried persons who try to promote their interest groups' aims by seeking personal contacts with legislators, by sending communications or information to legislators or persuading others to do so, and by rendering campaign assistance to favoured legislators. (In the U.S., lobbying is conducted not only by private interest groups but also by certain members of the executive branch and the bureaucracy.) *Semiconstitutional operations* include subvention or payment of retainers as well as providing campaign contributions, secretarial help, office accommodations, and the like, which are restricted but not forbidden by law. *Semidirect action* includes mass lobbying, such as promoting mass demonstrations before legislators, ostensibly to relate grievances but in fact to impress public opinion. *Direct action* includes threatening legislators with a withdrawal of financial assistance, with physical violence, or with riots (as in Japan, where snake dances and riots become a means of protest).

When political parties are powerful, they may prove a highly preferred target. Tactics in such cases include the following: *Constitutional operations,* such as, (1) exacting election pledges, which is most important and fruitful in countries (such as the U.S.) where there is relatively little national-party discipline, where representatives owe a certain allegiance to local-party groups and constituents, and where, most important, representatives do possess legislative power; (2) financing the party, which in all systems is perhaps the most effective operation and which in most

Western parliaments—and in Western-style parliaments like Japan's—is systematically done by trade unions, employers associations, and individual firms; and (3) lobbying the party machinery, a method that is virtually nonexistent in the United States but is significant in such countries as Great Britain and Germany.

Tactics to affect political parties

A final target of interest groups is public opinion. The general opinion of the public at large can be of importance in issues where questions of "fairness" or "justice" arise (on which most people feel they can pronounce). These questions involve such matters as capital punishment, the use of nuclear weapons, or certain wars, about which large numbers of people feel very strongly.

ROLE OF INTEREST GROUPS IN VARIOUS POLITICAL SYSTEMS

The Anglo-American and Swedish patterns. In the United States, Great Britain, and Sweden, the general population is not irreparably divided on ideological or cultural lines. Although there arise from time to time some small extremist groups, the population at large is not split into irreconcilable blocks. Given this situation, the most influential and prominent groups are "secondary" associations that have highly specialized objectives, form freely without the requirement of prior governmental permission, and carry out their activities autonomously. With the exception of the affiliation of the trade-union movement to the Labour Party in Britain and in Sweden, the groups are politically neutral; and even the alignment of British and Swedish trade unions and cooperatives does not prevent them from cooperating with other parties nor hinder their free access to government departments. Latent groups are few, since freedom of speech and association and the right to petition the government permit most interests to organize, while good communication media facilitate such organization. Public participation in such groups is, by world standards, high. The pressure groups play highly specialized roles and are sharply differentiated from the political parties.

National differences

Although American and British means of influencing the executive are similar (by advisory committees, special inquiries, and day-to-day consultation and mutual assistance), these relations are more insitutionalized in Britain than in the U.S. British groups seek to influence the parties by affiliation or by securing their members' nomination as candidates, neither mode being employed in the U.S., and by financial assistance, which occurs in both states but in quite different ways owing to the very different legislation on this matter in the two countries. British groups seek to influence the legislature by securing direct representation through participation in party caucuses and by promoting amendments in the committee stage. In the U.S. the methods employed consist of lobbying influential congressmen and appearing at committee hearings, which are far more important in Congress than in the House of Commons. In influencing public opinion, American groups spend far more than their British counterparts—one reason being that purchase of time on radio and television networks is permissible in the U.S., whereas in Britain it is not. In addition, access to the presidency can prove decisive in the U.S.; and some groups such as the National Association for the Advancement of Colored People have scored notable successes by seeking court action.

In Sweden the tactics are somewhat different, owing to three factors: the existence of a multiparty system and the existence of coalition cabinets, the structure of the bureaucracy, and the much greater institutionalization of pressure groups. Sweden, Denmark, and Norway each possess a multiparty system; but when coalitions are formed, they are highly stable. Consensus is achieved in a three-tier operation—first at interparty (electoral) level, then at the interparty-coalition (legislative) level, and finally at the bureaucratic level. In Sweden groups seek to influence the parties by affiliation (the trade unions being affiliated to the Labour Party), by financing them (trade unions giving to the Labour Party, employers groups to the antisocialist parties), and by securing direct representation through the nomination of members as parliamentary candidates. The parties respond by trying to compose a balanced ticket of the spectrum of interests. Hence the groups can affect the legislature through their representatives inside each of the parties and may, on occasion, form a cross-party legislative pressure group in defense of their common interest; and they also seek access to the party caucus and influence the committee stage of legislation. In the latter case they are assisted by the standard Swedish procedure by which the "comments" of outside interests are attached to the government's bill, even if they contradict the bill's purposes. Groups have ready access to the legislature in the same ways that Anglo-American groups do, except that the procedure is highly institutionalized; the "Royal Commission" with representatives of the bureaucracy, the parties, and the interest groups seek to reach initial agreement on a bill to be presented to the legislature; and when this device is not used, "comments" appended to the bill serve to put the measure in the context of the avowed interests of various groups.

Not unlike the system in Britain, the U.S., and Sweden is the system in Japan, which, despite sharp ideological divisions, has a strong and stable two-party arrangement. This has led to the emergence of a pattern of highly institutionalized pressure groups of the "secondary" type, the most influential of these being business and organized labour. Business, on the whole, is aligned with the conservative Liberal-Democratic party, supplies the bulk of its funds, and has a large say in policy formation. Organized labour constitutes a powerful element within the Japanese Socialist Party. In an earlier day, rural farm groups carried considerable political weight—for the conservatives—but the increasing industrialization and urbanization of the country have tended to reduce its importance.

The French and Italian patterns. In France and Italy the political pattern is characterized by deep ideological cleavages. The Roman Catholic Church, for example, operates through both its own political party and its own interest groups, such as the Catholic Action Society and Catholic youth and women's group. By the same token, the highly ideological and sectarian Commmunist Party establishes its own pressure groups, such as trade unions, farmers associations, and youth movements. Several rival groups thus compete for the same clientele aligned or affiliated with rival political parties.

Under France's Fifth Republic, the executive branch currently controls the majority party in the legislature, and the role of the legislature has been reduced; whereas, under Italy's current system, feeble and fragile coalition cabinets have weakened executive leadership of the legislature, and the legislature itself is highly fragmented among numerous parties. Because contacts between interest groups on the one hand and the parties, civil service departments, and legislature on the other are poorly institutionalized, semidirect action or even direct action against the executive and the legislature are more common, and lawless activities of crowds and violent organizations have proved more widespread.

The means of influencing the chosen targets also vary. In France attention has shifted from the legislature (under the Fourth Republic) to the president and bureaucracy (under the Fifth); interest groups now act through the numerous advisory bodies, private contacts, and such official institutions as the Economic and Social Council. Interest groups secure contacts by putting former civil servants on their own payrolls. The fact that many civil servants and business executives come from the same schools facilitates private collusion.

In Italy interest groups direct their attention mainly to three targets. First, the bureaucracy, even though it is highly fragmented and provides few formal advisory channels of communication, is open to ad hoc and personal consultations (except to groups deemed ideologically distasteful, such as the Communists). Second, the parties are often twinned to interest groups, as noted earlier; and thus the groups seek and achieve direct representation in the legislature. Third, the legislature, because it is composed of interest groups, is heavily influenced in specialized legislative committees. In general, therefore, interest groups have tended to "intrude" into the organs of government, forming close alliances with certain departments and legislative commissions and inserting themselves into others

via their membership in political parties. In short, interest groups are not "contained" as they are in Anglo-American and Scandinavian countries—political institutions are fragmented and weak, procedures are irregular or personal and not formal, and public beliefs are polarized into uncompromising ideologies.

Patterns in developing countries. Although there is no one typical pattern for all developing countries, which of course embrace more than two-thirds of the world's sovereign states, there are a few common characteristics that can be suggested.

First, political culture in each of these countries is fragmented, sometimes because of ideological differences, sometimes because of the mutual hostility of "natural" groupings based on kinship, lineage, tribe, or language. Second, nascent interest groups representing new modernizing and industrializing forces tend to range against rural natural groupings. The modernizers have defined and specialized objectives, whereas the traditionalists usually represent a spectrum of attitudes, beliefs, values, and interests. Both, in any case, tend to sponsor or generate their own political parties. Third, during preindependence days in some countries, trade unions and other movements (such as the Somali Youth League) carried on activities that would otherwise have been carried on by political parties; when independence came, these groups continued such activities, thus blurring the functional distinctions between parties and interest groups. Fourth, associations in these countries tend to have low membership densities; and in most of these countries the principal "interest"—that is, the peasantry—is hardly organized at all. Attempts to develop widely represented peasant movements have had only limited success. Fifth, although organization and established procedures may be poor, latent interest is usually high. In a particularly divisive and administratively feeble country, this interest latency may reveal itself in lawless and violent group activity—such as strikes, riots, guerrilla movements, and assassinations.

In some countries "parties" in elections are not really parties but simply electoral coalitions of various interest groups—the most influential groups consisting of cliques of landlords; leaders of religious, ethnic, or linguistic groups; and the like. Legislatures tend to reflect these divisions and temporary arrangements in society and so enjoy only feeble authority.

In countries that have had single-party regimes, such as Kenya and Tanzania, the parties tended to lose their mass following and fairly well-defined organization in the years after independence. To enhance their authority, the party leaders tried to limit or abolish the autonomy of such interest groups as trade unions, cooperative movements, and youth and women's movements and tried to turn these groups into ancillaries of the party.

In some countries the military has taken over, either alone or with civilian elements, and imposed an authoritarian rule. In such cases, the most prominent groups are governmental—the military and the executive. Unless interest groups enjoy a long history of organization (as does the Roman Catholic hierarchy in Latin America), they are often turned into governmental ancillaries or suppressed.

Countries with one-party or military governments

In these countries ruled by the military or single parties, the political process turns into a highly informal medley of pressures and personal contacts between the government, the military forces, the various "natural" groups, and some of the more strategically placed groups (such as the church hierarchy or private industry). The "system" is constantly threatened with government coercion or violent outbreaks by sections of the population. (S.E.F./Ed.)

Patterns in former Communist states. From World War II to 1989, most of the Communist nations in eastern Europe were ruled by a single Communist party (as in Albania, Hungary, Romania, and the Soviet Union), or by a Communist party that dominated one or more satellite parties in a hegemonic multiparty system (as in Bulgaria, Poland, Czechoslovakia, and East Germany). Falling somewhere between these categories was Yugoslavia, which was governed by a League of Communists composed of Communist parties based in its several ethnic republics. Following the fall of the Berlin Wall in November 1989, these party systems were entirely transformed. By 1990, each country (Albania in 1991) quickly held relatively free elections that shattered the old regime. In most cases, power was transferred to those with little connection to the old leadership or who were dissident Communists. In most cases, the former Communist parties soon disbanded or reformed under different names to compete with new parties for votes and political influence. The parties and party systems in these countries are not yet institutionalized and will need time, perhaps decades, to achieve stability and acquire popular value. This slow process is common to countries seeking democratic government after authoritarian rule.

The first wave of elections (1989–90) tended to go heavily against Communist candidates and toward candidates backed by mass popular movements. In Poland, for example, Lech Wałęsa's labour-based Solidarity movement swept nearly all the offices it contested in 1989. In Czechoslovakia, Václav Havel's Civic Forum (and its Slovak counterpart) decisively defeated the Communist candidates in the 1990 assembly elections. Old-line Communist rulers were also ousted that year in Hungary and East Germany. Although Communist governments were reelected in Albania and Bulgaria, even these hard-line regimes were defeated by opposition forces in the second round of elections—Bulgaria in 1991 and Albania in 1992. In Russia only the Communist Party was allowed to participate as a party in the 1990 elections for the 1,068 seats in the Congress of People's Deputies. Nevertheless, many candidates were backed by popular fronts, interest groups, and political clubs that had arisen under glasnost (the Soviet policy of "openness" that began in the late 1980s). Democratic Russia, an organization of progressive forces, claimed 190 seats after the election. In Russia's historic popular election for president of the republic in 1991, Boris Yeltsin won 57 percent of the vote against five other candidates, some of whom were backed by the Communist Party of the Soviet Union. Only in Romania did voters keep former Communists in power through 1992, although old-line Communists also won power in most of the former Soviet republics. This was particularly true in the Asian republics, where elections were less free and marked by a high turnout of government-mobilized voters, but old-line Communists also won elections in Ukraine, the largest republic after Russia. However, three small Baltic republics (Estonia, Latvia, and Lithuania) ousted their former Communist leaders, most convincingly in Lithuania, where the mass popular movement Sąjūdis won about 65 percent of the parliamentary seats in early elections.

Despite the initial landslides toward mass-based democratic movements in some cases, the most characteristic feature of free elections in these former Communist nations was the proliferation of political parties, as political entrepreneurs sought to take advantage of an uprooted electorate. For example, Poland soon had more than 100 registered parties, Romania more than 80, and Bulgaria more than 50. A survey of parties in eastern Europe and the Soviet Union published in 1991 listed more than 500 different parties. Most of these were known as "couch" parties (the entire membership would fit on a sofa), and they had little structure or staff. The proliferation of ephemeral parties produced political confusion as voters faced a bewildering array of choices in an unfamiliar market. In Romania, for instance, citizens who were new to free elections could choose among the National Democrats, Romanian Democrats, Free Democrats, Social Democrats, Liberal Democrats, Constitutional Democrats, and Christian Democrats—to name a few. One consequence was disillusionment with the electoral process and low voter turnout. In Poland, for example, only 43 percent of the eligible electorate voted in the parliamentary elections of 1991, which saw 29 different parties elected to the lower house, including the Polish Party of the Friends of Beer (Beer Lovers' Party), which won 16 seats in the lower parliament in 1991.

Emergence of political parties

In general terms, the nascent parties that sprouted in the former Communist countries can be classified into seven types. First, there were the parties of mass demo-

cratic movements—Solidarity in Poland, Civic Forum in Czechoslovakia, and Sąjūdis in Lithuania, for instance—that were often instrumental in forcing the Communist authorities to schedule free elections. However, most parties of this type dramatically lost support in the second wave of elections. (Civic Forum had split into two wings by then.) Second were the remnants of the former Communist Party operating with names like the Socialist Party (Albania, Bulgaria, and Hungary) or the National Salvation Front (Romania). These parties may change their names as they develop. For instance, the Democratic National Salvation Front in Romania became the Party for Social Democracy in July 1993. In Lithuania, the former Communists—reorganized as the Democratic Labour Party—actually outpolled the Sąjūdis in the November 1992 parliamentary election and regained the government.

A third type consisted of parties that took up the mantles of pre-World War II parties, such as various farmers' and liberal parties. A fourth kind represented nationalist parties promoting ethnic interests, as, for example, the Hungarian Democratic Union in Romania. Fifth were religious parties, typically Christian Democrats. A sixth category consisted of parties modeled after Western political values, such as environmentalism, feminism, and capitalism. Finally, there were the frivolous parties, like the Beer Lovers' and Volcano parties in Poland.

As explained above, the nature of the electoral system affects the number of parties that win representation to parliament. Countries using proportional representation and having few electoral barriers to discourage minor parties sustained severely fragmented party systems. Poland, for example, did not require parties to achieve any minimum vote (threshold) to gain representation in 1991, and none of its 29 parliamentary parties had more than 13 percent of the vote. This fragmentation in the Polish parliament made it difficult to form a governing coalition. Hungary, on the other hand, required that parties win 4 percent of the national vote in 1990, and only six out of more than 65 registered parties entered parliament. Countries that did not use proportional representation, such as Russia, usually required that candidates win an absolute majority of the vote or face a runoff election. This two-ballot system, also used in France, favours party fragmentation by encouraging minor parties to form for the purpose of denying the leading candidate a majority on the first ballot and thus costing him the election. Minor parties can then bargain their support, in exchange for favours, on the second ballot, often held one week after the first. The alternative system, used in most Anglo-American democracies, requires only a simple plurality of the vote and tends to produce two-party rather than multiparty systems. The new democracies that have emerged from the former Communist countries are certain to experiment with different electoral systems as they seek to develop institutionalized parties and stable party systems. (K.F.J.)

BIBLIOGRAPHY

Political parties. ROBERT MICHELS, *Political Parties: A Sociological Study of the Oligarchical Tendencies of Modern Democracies* (1915, reissued 1978; originally published in German, 1911), provides the first modern theory of political parties, stemming from the doctrines of the "elite," or "ruling class," of Gaetano Mosca; the book is based largely on the example of German social democracy. MAURICE DUVERGER, *Political Parties: Their Organization and Activity in the Modern State*, 3rd ed., rev. (1964, reissued 1976; originally published in French, 1951), provides a comparative and systematic study of political parties, emphasizing the experience of the European countries. JOSEPH G. LAPALOMBARA and MYRON WEINER (eds.), *Political Parties and Political Development* (1966, reissued 1972), is a collection of contributions by various authors and is a good comparative explication of problems concerning political parties. DOUGLAS W. RAE, *The Political Consequences of Electoral Laws*, rev. ed. (1971), offers, on the basis of an examination of the balloting in numerous countries since 1945, an analysis of the relationship between the electoral system and political parties; it is a model of comparative study.

Reference works providing world or regional coverage of political parties include GEORGE E. DELURY (ed.), *World Encyclopedia of Political Systems & Parties*, 2nd ed. (1987); ALAN J. DAY (ed.), *Political Parties of the World*, 3rd ed. (1988); KLAUS VON BEYME, *Political Parties in Western Democracies* (1985; originally published in German, 1982); CHARLES HOBDAY (compiler), *Communist and Marxist Parties of the World* (1986); BOGDAN SZAJKOWSKI (ed.), *New Political Parties of Eastern Europe and the Soviet Union* (1991), describing more than 500 parties that have emerged in the former Communist countries; ROBERT J. ALEXANDER (ed.), *Political Parties of the Americas: Canada, Latin America, and the West Indies*, 2 vol. (1982), including coverage of organizations no longer functioning; CHARLES D. AMERINGER (ed.), *Political Parties of the Americas, 1980s to 1990s: Canada, Latin America, and the West Indies* (1992), with updated essays and bibliographies; JOHN COGGINS and D.S. LEWIS (eds.), *Political Parties of the Americas and the Caribbean* (1992), including coverage of outlawed groups; HARUHIRO FUKUI (ed.), *Political Parties of Asia and the Pacific*, 2 vol. (1985), with valuable historical information; and D.S. LEWIS and D.J. SAGAR (eds.), *Political Parties of Asia and the Pacific* (1992), with accounts up to mid-1992.

Works about American parties are numerous. Recommended are SAMUEL J. ELDERSVELD, *Political Parties: A Behavioral Analysis* (1964); WILLIAM NISBET CHAMBERS and WALTER DEAN BURNHAM (eds.), *The American Party Systems: Stages of Political Development*, 2nd ed. (1975, reprinted 1981); and XANDRA KAYDEN and EDDIE MAHE, JR., *The Party Goes On: The Persistence of the Two-Party System in the United States* (1985). Works that study western European political parties include PETER H. MERKL (ed.), *Western European Party Systems: Trends and Prospects* (1980), essays on 13 countries and a collection of essays on research topics; VINCENT E. MCHALE and SHARON SKOWRONSKI (eds.), *Political Parties of Europe*, 2 vol. (1983); and EVA KOLINSKY (ed.), *Opposition in Western Europe* (1987). On Great Britain, SAMUEL H. BEER, *Modern British Politics: A Study of Parties and Pressure Groups*, 2nd ed. (1969); and R.T. MCKENZIE, *British Political Parties: The Distribution of Power Within the Conservative and Labour Parties*, 2nd ed. (1963, reprinted 1992), are particularly useful. DAVID E. APTER, *The Politics of Modernization* (1965); THOMAS L. HODGKIN, *African Political Parties* (1961, reissued 1971); and A. MAHIOU, *L'Avènement du parti unique en Afrique noire* (1969), give informative historical accounts of the operations of political parties in developing countries. VICKY RANDALL (ed.), *Political Parties in the Third World* (1988), features case studies of Zambia, Ghana, Iraq, India, Mexico, Brazil, Jamaica, and Cuba.

Interest groups. GABRIEL A. ALMOND and JAMES S. COLEMAN (eds.), *The Politics of the Developing Areas* (1960), provides a synoptic view of government in Latin America, Asia, and Africa that places special-interest groups in the setting of the politics of developing countries. KLAUS VON BEYME, *Interessengruppen in der Demokratie*, rev. new ed. (1980), classifies the types, strategies, and tactics of groups in the Western democratic states. ARTHUR F. BENTLEY, *The Process of Government: A Study of Social Pressures* (1908, reissued 1967), is the major theoretical source for all subsequent Western democratic writing, advocating the theory by which all political activity is reduced to group activity. DONALD C. BLAISDELL, *American Democracy under Pressure* (1957), ranks among the best of numerous surveys of the role of groups in American society. S.E. FINER, *Anonymous Empire: A Study of the Lobby in Great Britain*, 2nd ed., rev. and enlarged (1966), briefly analyzes the types, strategies, and tactics of groups in Britain. A.G. JORDAN and J.J. RICHARDSON, *Government and Pressure Groups in Britain* (1987), surveys interest-group theory and the role of these groups in Britain. HENRY W. EHRMANN, *Organized Business in France* (1957, reprinted 1981), contains a highly detailed account of the typical methods of groups in France. HENRY W. EHRMANN (ed.), *Interest Groups on Four Continents* (1958, reprinted 1983), a collection of papers and summaries of the discussions based on them, provides an excellent country-by-country account of the operation of groups. JEAN MEYNAUD, *Les Groupes de pression en France* (1958), offers a detailed analysis of group activity in France. J.G. LAPALOMBARA, *Interest Groups in Italian Politics* (1964), is a detailed analysis, much of it based on interview and behavioral data. ALEC BARBROOK and CHRISTINE BOLT, *Power and Protest in American Life* (1980), studies organizations acting for women, ethnic minorities, and the public interest. (M.Du./S.E.F./K.F.J./Ed.)

The History of Western Political Philosophy

The central problem of political philosophy is how to deploy or limit public power so as to maintain the survival and enhance the quality of human life. Like all aspects of human experience, it is conditioned by environment and by the scope and limitations of mind; and the answers given by successive political philosophers to perennial problems reflect the knowledge and the assumptions of their times. Political philosophy, as distinct from the study of political and administrative organization, is more theoretical and normative than descriptive. It is inevitably related to general philosophy and is itself a subject of social anthropology, sociology, and the sociology of knowledge. As a normative discipline it is thus concerned with what ought, on various assumptions, to be and how this purpose can be promoted, rather than with a description of facts—although any realistic political theory is necessarily related to these facts. The political philosopher is thus not concerned so much, for example, with how pressure groups work or how, by various systems of voting, decisions are arrived at, as with what the aims of the whole political process should be in the light of a particular philosophy of life.

There is thus a distinction between political philosophy, which reflects the world outlook of successive theorists and which demands an appreciation of their historical settings, and modern political science proper, which, insofar as it can be called a science, is empirical and descriptive. Political philosophy, however, is not merely unpractical speculation, though it may give rise to highly impractical myths: it is a vitally important aspect of life, and one that, for good or evil, has had decisive results on political action; for the assumptions on which political life is conducted clearly must influence what actually happens. Political philosophy may thus be viewed as one of the most important intellectual disciplines, for it sets standards of judgment and defines constructive purposes for the use of public power. Such consideration of the purposes for which power should be used is in a sense more urgent today than it has been in earlier periods, for mankind has at its disposal the power either to create a world civilization in which modern technology can benefit the human race or to destroy itself in pursuit of political myths. The scope for political philosophy is thus great, the clarification of its purpose and limitations urgent—an aspect, indeed, of civilization's survival.

Despite this unique aspect of the contemporary situation and although ancient political philosophies were formulated under very different conditions, their study still illuminates questions vital today. Questions concerning the aims of government, the grounds of political obligation, the rights of individuals against the state, the basis of sovereignty, the relation of executive to legislative power, and the nature of political liberty and social justice have been asked and answered in many ways over the centuries. They are all fundamental to political philosophy and demand answers in terms of modern knowledge and opinion.

This article describes how these questions have been asked and answered by representative and influential political philosophers, from Greco-Roman antiquity through the Middle Ages, early modern times, and the 19th and early 20th centuries. During so long a time span the historical context of these formulations has changed profoundly, and an understanding of the political philosophers selected demands some account of their background. Because of limitations of space, only political philosophers of outstanding importance have been at all fully described, although many minor figures also are briefly discussed.

The article is divided into the following sections:

The history of political philosophy in the West to the end of the 19th century 991
- Antiquity 991
  - Plato
  - Aristotle
  - Cicero and the Stoics
  - St. Augustine
- The Middle Ages 994
  - John of Salisbury
  - Aquinas
  - Dante
- The 16th to the 18th centuries 994
  - Machiavelli
  - Hobbes
  - Spinoza
  - Richard Hooker's adapted Thomism
  - Locke
  - Burke
  - Vico
  - Montesquieu
  - Rousseau
- The 19th century 997
  - Utilitarianism
  - Tocqueville
  - T.H. Green
  - Liberal nationalism
  - American constitutionalism
  - Anarchism and utopianism
  - Saint-Simon and Comte
  - Hegel
  - Marx and Engels

Political philosophy in the 20th century 999
- Political pragmatism 999
- Religious and Existentialist approaches 1000
- Revolutionary doctrines 1001
  - Lenin
  - Other Marxist approaches

Conclusion 1002
Bibliography 1002

The history of political philosophy in the West to the end of the 19th century

ANTIQUITY

Although in antiquity great civilizations arose in Egypt and Mesopotamia, in the Indus Valley, and in China, there was little speculation about the problems of political philosophy as formulated in the West and since predominant. The laws of Hammurabi of Babylon (*c.* 1750 BC) are rules propounded by the monarch as a representative of God on Earth and are mainly concerned with order, trade, and irrigation; the *Admonitions* of the Egyptian vizier Ptahhotep (*c.* 2300 BC) are shrewd advice on how to prosper in a bureaucracy; and the *Arthaśāstra* of Kauṭilya, grand vizier to the Indian Candragupta Maurya in the late 4th century BC, are Machiavellian precepts on how to survive under an arbitrary power. To be sure, the Buddhist concept of *dharma* (social custom and duty), which inspired the Indian emperor Aśoka in the 3rd century BC, implies a moralization of public power, and the teachings of Confucius in the 6th century BC are a code of conduct designed to stabilize society; but there is not, outside Europe, much speculation about the basis of political obligation and the purpose of the state, with both of which Western political philosophy is mainly concerned. An authoritarian society is taken for granted, backed by religious sanctions, and a conservative and arbitrary power is generally accepted.

In contrast to this overwhelming conservatism, paralleled

by the rule of custom and tribal elders in most primitive societies, the political philosophers of ancient Greece question the basis and purpose of government; and, though they do not separate political speculation from shrewd observations that today would be regarded as empirical political science, they created the vocabulary of Western political thought.

Plato. The first elaborate work of European political philosophy is *The Republic* of Plato (*c.* 378 BC), a masterpiece of insight and feeling, superbly expressed in dialogue form and probably meant for recitation. Further development of Plato's ideas is undertaken in his *Statesman* and *Laws*, the latter prescribing the ruthless methods whereby they might be imposed. Plato grew up during the great war between Athens and Sparta in which Athens suffered defeat and, like many political philosophers, tried to find remedies for prevalent political injustice and decline. Indeed, *The Republic* is the first of the utopias, though not one of the more attractive; and it is the first classic attempt of a European philosopher to moralize political life.

Cast as a lively discussion between Socrates, whose wisdom Plato is recounting, and various leisured Athenians, Books V, VII–VIII, and IX of *The Republic* state the major themes of political philosophy with poetic power. Plato's work has been criticized as static and class bound, reflecting the moral and aesthetic assumptions of an elite in a slave-owning civilization and bound by the narrow limits of the city-state. The work is indeed a classic example of a philosopher's vivisection of society, imposing by relatively humane means the rule of a high-minded minority.

Epistemological assumptions in *The Republic*

The Republic is a criticism of current Hellenic politics—often an indictment. It is based upon a metaphysical act of faith, for Plato believes that a world of permanent Forms exists beyond the limitations of human experience and that morality and the good life, which the state should promote, are reflections of these ideal Forms (see PLATONISM). The point is best made in the famous simile of the cave, in which men are chained with their faces to the wall and their backs to the light, so that they see only the shadows of reality. So constrained, they shrink from what is truly "real" and permanent and need to be forced to face it. This idealistic doctrine, known misleadingly as Realism (in nontechnical language it is hardly realistic), pervades all Plato's philosophy: its opposite doctrine, Nominalism, declares that only particular and observed "named" data are accessible to the mind. On his Realist assumption, Plato, who was perhaps influenced by Indian thought, regards most ordinary life as illusion and the current evils of politics as the result of men pursuing brute instinct. It follows that

> unless philosophers bear kingly rule in cities or those who are now called kings and princes become genuine and adequate philosophers, and political power and philosophy are brought together . . . there will be no respite from evil for cities.

Only philosopher-statesmen can apprehend permanent and transcendant Forms and turn to "face the brightest blaze of being" outside the cave, and only philosophically minded men of action can be the saviours and helpers of the people.

Plato is thus indirectly the pioneer of modern beliefs that only a party organization, inspired by correct and "scientific" doctrines, formulated by the written word and interpreted by authority, can rightly guide the state. His rulers would form an elite, not responsible to the mass of the people. Thus, in spite of his high moral purpose, he has been called an enemy of the open society and the father of totalitarian lies. But he is also an anatomist of the evils of unbridled appetite and political corruption and insists on the need to use public power to moral ends.

Having described his utopia, Plato turns to analyze the existing types of government in human terms with great insight. Kingly government is the best but impracticable; in oligarchies the rule of the few and the pursuit of wealth divide societies—the rich become demoralized and the poor envious, and there is no harmony in the state. In democracy, in which the poor get the upper hand, demagogues distribute "a peculiar kind of equality to equals and unequals impartially," and the old flatter the young, fawning on their juniors to avoid the appearance of being sour or despotic. The leaders plunder the propertied classes and divide the spoils among themselves and the people until confusion and corruption lead to tyranny, a worse form of government. For the tyrant becomes a wolf instead of a man and "lops off" potential rivals and starts wars to distract the people from their discontent. "Then, by Zeus," Plato concludes, "the public learns what a monster they have begotten."

Plato's *Statesman* and *Laws*

In the *Statesman* Plato admits that, although there is a correct science of government, like geometry, it cannot be realized, and he stresses the need for the rule of law, since no man can be trusted with unbridled power. He then examines which of the current forms of government is the least difficult to live with, for the ruler, after all, is an artist who has to work within the limits of his medium. In the *Laws*, purporting to be a discussion of how best to found a polis in Crete, he presents a detailed program in which a state with some 5,000 citizens is ruled by 37 curators of laws and a council of 360. But the keystone of the arch is a sinister and secret Nocturnal Council to be "the sheet anchor of the state," established in its "central fortress as guardian." Poets and musicians will be discouraged and the young subjected to a rigid, austere, and exacting education. The stark consequence of Plato's political philosophy here becomes apparent. He had, nonetheless, stated, in the dawn of European political thought, the normative principle that the state should aim at promoting the good life and social harmony and that the rule of law, in the absence of the rule of philosopher-kings, is essential to this purpose.

Aristotle. Aristotle, who was a pupil in the Academy of Plato, remarks that "all the writings of Plato are original: they show ingenuity, novelty of view and a spirit of enquiry. But perfection in everything is perhaps a difficult thing." Aristotle was a scientist rather than a prophet, and his *Politics* (*c.* 335–322 BC), written while he was teaching at the Lyceum at Athens, is only part of an encyclopaedic account of nature and society, in which he analyzes society as if he were a doctor and prescribes remedies for its ills. Political behaviour is here regarded as a branch of biology, as well as of ethics; in contrast to Plato, Aristotle was an empirical political philosopher. He criticizes many of Plato's ideas as impracticable, but, like Plato, he admires balance and moderation and aims at a harmonious city under the rule of law. The book is composed of lecture notes and is arranged in a confusing way—a quarry of arguments and definitions of great value but hard to master. The first book, though probably the last written, is a general introduction; Books II, III, and VII–VIII, probably the earliest, deal with the ideal state; and Books IV–VII analyze actual states and politics. The treatise is thus, in modern terms, a mixture of political philosophy and political science. (See also ARISTOTELIANISM.)

Like Plato, Aristotle naturally thinks in terms of the city-state, which he regards as the natural form of civilized life, social and political, and the best medium in which men's capacities can be realized. Hence his famous definition of man as a "political animal," distinguished from the other animals by his gift of speech and power of moral judgment. "Man, when perfected," he writes,

> is the best of animals, but when separated from law and justice he is the worst of all, since armed injustice is the most dangerous, and he is equipped at birth with the arms of intelligence and wit, moral qualities which he may use for the worst ends.

Since all nature is pervaded by purpose and since men "aim at the good," the city-state, which is the highest form of human community, aims at the highest good. Like sailors with their separate functions, who yet have a common object in safety in navigation, citizens, too, have a common aim—in modern terms survival, security, and the enhancement of the quality of life. In the context of the city-state, this high quality of life can be realized only by a minority, and Aristotle, like Plato, excludes those who are not full citizens or who are slaves; indeed, he says that some men are "slaves by nature" and deserve their status. Plato and Aristotle aim at an aristocratic and exacting way of life, reflecting, in more sophisticated forms, the ideas of the warrior aristocracies depicted by Homer.

Aristotle's view of the role of law in the city-state

Having stated that the aim of the city-state is to promote the good life, Aristotle insists that it can be achieved only under the rule of law.

> The rule of law is preferable to that of a single citizen; if it be the better course to have individuals ruling, they should be made law guardians or ministers of the laws.

The rule of law is better than that even of the best men, for "he who bids law rule may be deemed to bid God and reason alone rule, but he who bids men rule adds the element of the beast; for desire is a wild beast, and passion perverts the minds of rulers, even if they are the best of men." This doctrine, which distinguishes between lawful government and tyranny, survived the Middle Ages and, by subjecting the ruler to law, became the theoretical sanction of modern constitutional government.

Aristotle also vindicates the rule of custom and justifies the obligations accepted by members of society: the solitary man, he writes, "is either a beast or a God." This outlook at once reflects the respect for custom and solidarity that have promoted survival in primitive tribal societies, even at the price of sacrificing individuals, and gives a theoretical justification for the acceptance of political obligation.

Like Plato, Aristotle analyzes the different kinds of city-states. While states are bound, like animals, to be different, he considers a balanced "mixed" constitution the best—it reflects the ideal of justice (*dikē*) and fair dealing, which gives every man his due in a conservative social order in which citizens of the middle condition preponderate. And he attacks oligarchy, democracy, and tyranny. Under democracy, he argues, demagogues attain power by bribing the electorate and waste accumulated wealth. But it is tyranny that Aristotle most detests; the arbitrary power of an individual above the law who is "responsible to no-one and who governs all alike with a view to his own advantage and not of his subjects, and therefore against their will. No free man can endure such a government."

The *Politics* contains not only a firm statement of these principles but also a penetrating analysis of how city-states are governed, as well as of the causes of revolutions, in which "inferiors revolt in order that they may be equal, and equals that they may be superior." The treatise concludes with an elaborate plan for educating the citizens to attain the "mean," the "possible," and the "becoming." The first implies a balanced development of body and mind, ability and imagination; the second, the recognition of the limits of mind and the range and limitations of talent; the third, an outcome of the other two, is the style and self-assurance that come from the resulting self-control and confidence.

While, therefore, Aristotle accepts a conservative and hierarchic social order, he states firmly that public power should aim at promoting the good life and that only through the rule of law and justice can the good life be attained. These principles were novel in the context of his time, when the great extra-European civilizations were ruled, justly or unjustly, by the arbitrary power of semidivine rulers and when other peoples, though respecting tribal custom and the authority of tribal elders, were increasingly organized under war leaders for depredation.

Cicero and the Stoics. Both Plato and Aristotle had thought in terms of the city-state. But Aristotle's pupil Alexander the Great swamped the cities of old Greece and brought them into a vast empire that included Egypt, Persia, and the Levant. Though the civilization of antiquity remained concentrated in city-states, they became part of an imperial power that broke up into kingdoms under Alexander's successors. This imperial power was reasserted on an even greater scale by Rome, whose empire at its greatest extent reached from central Scotland to the Euphrates and from Spain to eastern Anatolia. Civilization itself became identified with empire, and the development of eastern and western Europe was conditioned by it.

Emergence of universal philosophies

Since the city-state was no longer self-sufficient, universal philosophies developed that gave men something to live by in a wider world. Of these philosophies, Stoicism and Epicureanism were the most influential. The former inspired a rather grim self-sufficiency and sense of duty, as exemplified by the writings of the Roman emperor Marcus Aurelius; the latter, a prudent withdrawal from the world of affairs.

The setting for political philosophy thus became much wider, relating individuals to universal empire, thought of, as in China, as coterminous with the civilization itself. Its inspiration remained Hellenic; but derivative Roman philosophers reinterpreted it, and Roman legists enclosed the old concepts of political justice in a carapace of legal definitions, capable of surviving their civilization's decline.

Cicero lived in a time of political confusion during which the old institutions of the republic were breaking down before military dictators. His *De republica* and *Laws* are both dialogues and reflect the classical sense of purpose: "to make human life better by our thought and effort." Cicero defined the *res publica* (commonwealth) as an association held together by law; he further asserted, as Plato had maintained with his doctrine of Forms manifest in the just city, that government was sanctioned by a universal natural law that reflected the cosmic order. Cicero expresses the pre-Christian Stoic attempt to moralize public power, apparent in the exacting sense of public responsibility shown by Hadrian and Marcus Aurelius in the 2nd century AD.

St. Augustine. With the conversion of the emperor Constantine (AD 312), when Christianity, long influential, became the predominant creed of the empire, and, under Theodosius (379–395), the sole official religion, political philosophy changed profoundly. St. Augustine's *City of God* (413–426), written when the empire was under attack by barbarians within and without, sums up and defines a new division between church and state and a conflict between "matter" and "spirit" resulting from the Fall of man and original sin.

St. Augustine, whose *Confessiones* are a record of a new sort of introspection, combined a classical and Hebraic dualism. From the Stoics and Virgil he inherited an austere sense of duty, from Plato and the Neoplatonists a contempt for the illusions of appetite, and from the Pauline and patristic interpretation of Christianity a sense of the conflict between Light and Darkness that reflects Zoroastrian and Manichaean doctrines emanating from Iran. In this context worldly interests and government itself are dwarfed by the importance of attaining salvation and of escaping from an astrologically determined fate and from the demons who embody the darkness. Life becomes illuminated for the elect minority by the prospect of eternal salvation or, for those without grace, shrivels under the glare of eternal fires.

St. Augustine regarded salvation as predestinate and the cosmic process as designed to "gather" an elect to fill the places of the fallen angels and so "preserve and perhaps augment the number of the heavenly inhabitants." The role of government and indeed of society itself becomes subordinated into a "secular arm," part of an earthly city, as opposed to the "City of God." The function of government is to keep order in a world intrinsically evil.

Augustine's view of the role of government

Since Christianity had long played the main role in defense of the veneer of a precarious urban civilization in antiquity, this claim is not surprising. Constantine came of crude Balkan origins, a soldier putting to rights a breakdown in government that would continue in the West with the abdication of the last Western emperor in 476, though in the East the empire would carry on with great wealth and power, centred in the new capital of Constantinople (Byzantium).

St. Augustine thus no longer assumed, as did Plato and Aristotle, that a harmonious and self-sufficient good life could be achieved within a properly organized city-state; he projected his political philosophy into a cosmic and lurid drama working out to a predestinate end. The normal interests and amenities of life became insignificant or disgusting, and the Christian Church alone exercised a spiritual authority that could sanction government. This outlook, reinforced by other patristic writings, would long dominate medieval thought, for with the decline of civilization in the West the church became more completely the repository of learning and of the remnants of the old civilized life.

THE MIDDLE AGES

The decline of ancient civilization in the West was severe; not, indeed, in technology, for the horse collar, the stirrup, and the heavy plow now came in; but political philosophy, like other intellectual interests, became elementary. In the Byzantine Empire, on the other hand, Justinian's lawyers in 529–533 produced the Codex Constitutionum, the Digest, the Institutes, which defined and condensed Roman law, and the Novels. The Byzantine *basileus,* or autocrat, had moral responsibility for guarding and harmonizing an elaborate state, a "colony" of heaven in which reason and not mere will ought to rule. And this autocracy and the orthodox form of Christianity were inherited by the Christianized rulers of the Balkans, of Kievan Russia, and of Muscovy.

In the West, two essential principles of Hellenic and Christian political philosophy were transmitted, if only in elementary definitions, in rudimentary encyclopaedias. Isidore of Seville in his 7th-century *Etymologiae,* for example, asserts that kings rule only on condition of doing right and that the rule reflects a Ciceronic law of nature "common to all people and mankind everywhere by natural instinct." Further, the barbarians respected the civilization they took over and exploited. When converted, they revered the papacy, and in 800 the Frankish Charlemagne even revived the Western Empire as holy and Roman. The idea of Christian empire coterminous with civilization thus survived in Western as well as Eastern Christendom.

John of Salisbury. After Augustine, not until the 12th-century Renaissance did another full-length speculative work of political philosophy appear in the West. The *Policraticus* of John of Salisbury (*c.* 1159) is the work of a man experienced in politics who became bishop of Chartres. Based on wide classical reading, it centres on the ideal ruler, who represents a "public power." John admired Augustus and the Roman emperor Trajan, and, in a still predominantly feudal world, his book carried on the Roman tradition of centralized authority, though without its Byzantine autocracy. The prince, he insists, is he who rules in accordance with law, while a tyrant is one who oppresses the people by irresponsible power. This distinction, which derives from the Greeks, Cicero, and St. Augustine, is fundamental to Western concepts of liberty and the trusteeship of power.

John did not know Aristotle's *Politics,* but his learning is nevertheless remarkable, even if his political similes are unsophisticated. His favourite metaphor for the body politic is the human body: the place of the head is filled by the prince, who is subject only to God; the place of the heart by the senate; the eyes, ears, and tongue are the judges, provincial governors, and soldiers; and the officials are the hands. The tax gatherers are the intestines and ought not to retain their accumulations too long; and the farmers and peasants are the feet. John also compares a commonwealth to a hive and even to a centipede.

This vision of a centralized government, more appropriate to the memory of the Roman Empire than to a medieval monarchy operating in a still semifeudal world, is a landmark of the 12th-century revival of speculative thought and reflects the better organized monarchy that Henry II was then building up.

Aquinas. It is a far cry from this practical 12th-century treatise by a man of affairs to the elaborate justification of Christian kingship and natural law created by St. Thomas Aquinas in the 13th century, during the climax of medieval Western civilization. His political philosophy is only part of a metaphysical construction of Aristotelian range—for Aristotle had now been assimilated from Arabic sources and given a new Christian content, with the added universality of the Stoic and Augustinian world outlook. Aquinas' *Summa theologiae* purports to answer all the major questions of existence, including those of political philosophy. Like Aristotle, Aquinas thinks in terms of an ethical purpose. Natural law is discussed in the first part of the second book as part of the discussion of original sin and what would now be termed psychology, while war comes under the second part of the second book as an aspect of virtue and vice. Law is defined as "that which is regulation and measure." It is designed to promote the "felicity and beatitude" that are the ends of human life. Aquinas agrees with Aristotle that "the city is the perfection of community" and that the purpose of public power should be to promote the common good. The only legitimate power is from the community, which is the sole medium of man's well being. In his *De regimine* he compares society to a ship in need of a helmsman and repeats Aristotle's definition of man as a social and political animal. Again following Aristotle, he considers oligarchy unjust and democracy evil. Rulers should aim to make the "life of the multitude good in accordance with the purpose of life which is heavenly happiness." They should also create peace, conserve life, and preserve the state—a threefold responsibility. Here is a complete program for a hierarchical society within a cosmic order. It combines the Hellenic sense of purpose with Christian aims and asserts that, under God, power resides in the community, embodied in the ruler but only for so long as he does right. Hence the comment that "St. Thomas Aquinas was the first Whig"—a pioneer of the theory of constitutional government. The society he envisages, however, is medieval, static, hierarchical, conservative, and based on limited agriculture and even more limited technology. Nonetheless, Thomism remains the most complete and lasting political doctrine of the Catholic Church, since modified and adapted but not in principle superseded.

Aquinas' political views in the *Summa theologiae*

Dante. By the early 14th century the great European institutions, empire and papacy, were breaking down through mutual conflict and the emergence of national realms. But this conflict gave rise to the most complete political theory of universal and secular empire formulated in the medieval West. Dante's *De monarchia* (*c.* 1313), still in principle highly relevant, insists that only through universal peace can human faculties come to their full compass. But only "temporal Monarchy" can achieve this: "a unique princedom extending over all persons in time." The aim of civilization is to actualize human potentialities, and to achieve that "fullness of life which comes from the fulfillment of our being."

Monarchy, Dante argues, is necessary as a means to this end. The imperial authority of the Holy Roman emperor, moreover, comes direct from God and not through the pope. The empire is the direct heir of the Roman Empire, a legitimate authority, or Christ would not have chosen to be born under it. In subjecting the world to itself, the Roman Empire had contemplated the public good.

This high-flown argument, part of the political warfare between the partisans of the emperor and pope that was then affecting Italy, drives to essentials: that world peace can be secure only under a world authority. That Dante's argument was impractical did not concern this medieval genius, who was writing more the epitaph than the prospectus of the Holy Roman Empire; he was concerned, like St. Thomas, to create a political philosophy with a clear-cut aim and a universal view.

Out of the grand but impractical visions of the High Middle Ages in the 13th-century climax of Christian civilization there emerged by early modern times the idea of a well-governed realm, its authority derived from the community itself, with a program designed to ensure the solvency and administrative efficiency of a secular state. In spite of the decline of the civilization of antiquity in the West, the Greco-Roman sense of purpose, of the rule of law, and of the responsibility of power survived in Christian form.

THE 16TH TO THE 18TH CENTURIES

Machiavelli. In the thought of the Italian political philosopher Niccolò Machiavelli may be seen a complete secularization of political philosophy. Machiavelli was an experienced diplomat and administrator, and, since he stated flatly how the power struggle was conducted in Renaissance Italy, he won a shocking reputation. He was not, however, without idealism about the old Roman republic, and he admired the independent spirit of the German and Swiss cities. This idealism made him all the more disgusted with Italian politics, of which he makes a disillusioned and objective analysis. Writing in retirement after political disgrace, Machiavelli states firmly that,

Machiavelli's view of human nature

> Since this is to be asserted in general of men, that they are ungrateful, fickle, false, cowards, covetous, and as long as you succeed they are yours entirely: they will offer you their blood, property, life, and children . . . when the need is far distant; but when it approaches they turn against you.

And again,

> since the desires of men are insatiable, nature prompting them to desire all things and fortune permitting them to enjoy but few, there results a constant discontent in their minds, and a loathing of what they possess.

This view of human nature, already expressed by Plato and St. Augustine, is here unredeemed by Plato's doctrine of form and illusion or by St. Augustine's dogma of salvation through grace. Machiavelli accepts the facts and advises the ruler to act accordingly. The prince, he states, must combine the strength of the lion with the cunning of the fox: he must always be vigilant, ruthless, and prompt, striking down or neutralizing his adversaries without warning. And when he does an injury it must be total. For "men ought to be either well treated or crushed, because they can avenge themselves of lighter injuries, of more serious ones they cannot." Moreover, "irresolute princes who follow a neutral path are generally ruined." He advises that it is best to come down at the right moment on the winning side and that conquered cities ought to be either governed directly by the tyrant himself residing there or destroyed. Princes, furthermore, unlike private men, need not keep faith: since politics reflects the law of the jungle, the state is a law unto itself, and normal moral rules do not apply to it.

Machiavelli had stated with unblinking realism how, in fact, tyrants behave; and, far from criticizing their conduct or distinguishing between the just prince who rules by law and the tyrant whose laws are in his own breast, he considers that the successful ruler has to be beyond morality since the safety of and expansion of the state are the supreme objective. In this myopic view, the cosmic visions of Aquinas and Dante are disregarded, and politics becomes a fight for survival. Within his terms of reference, Machiavelli made a convincing case, although as an experienced diplomat he might have realized that dependability in fact pays and that systematic deceit, treachery, and violence usually bring about their own nemesis.

Hobbes. The 17th-century English political philosopher Thomas Hobbes, who spent his life as a tutor and companion to great noblemen, was a writer of genius with a greater power of phrase than any other English political philosopher. He was not, as he is sometimes misrepresented, a prophet of "bourgeois" individualism, advocating free competition in a capitalistic free market. On the contrary, he was writing in a preindustrial, if increasingly commercial, society and did not much admire wealth as such but rather "honours." He was socially conservative and anxious to give a new philosophical sanction to a hierarchical, if businesslike, commonwealth in which family authority was most important.

Philosophical influences on Hobbes

Philosophically, Hobbes was influenced by nominalist scholastic philosophy, which had discarded Thomist metaphysics and had accepted a strict limitation of mind. He therefore based his conclusions on the rudimentary mathematical physics and psychology of his day and aimed at practical objectives—order and stability. He believed that the fundamental physical law of life was motion and that the predominant human impulses were fear and, among those above the poverty level, pride and vanity. Men, Hobbes argued, are strictly conditioned and limited by these laws, and he tried to create a science of politics that would reflect them. "The skill of making, and maintaining Common-wealths," therefore, "consisteth in certain Rules, as doth Arithmetique and Geometry; not (as Tennis play) on Practise onely: which Rules, neither poor men have the leisure, nor men that have had the leisure, have hitherto had the curiosity, or the method to find out."

Hobbes ignores the classical and Thomist concepts of a transcendent law of nature, itself reflecting divine law, and of a "chain of being" whereby the universe is held harmoniously together and, following Descartes's practical method of investigation, states plainly that power creates law, not law power. For law is law only if it can be enforced, and the price of security is one supreme sovereign public power. For, without it, such is the competitive nature of men, that once more than subsistence has been achieved they are actuated by vanity and ambition, and there is a war of all against all. The true law of nature is self-preservation, he argues, which can be achieved only if the citizens make a compact among themselves to transfer their individual power to the "leviathan" (ruler), who alone can preserve them in security. Such a commonwealth has no intrinsic supernatural or moral sanction: it derives its original authority from the people and can command loyalty only so long as it succeeds in keeping the peace. He thus uses both the old concepts of natural law and contract, often invoked to justify resistance to authority, as a sanction for it.

Hobbes, like Machiavelli, starts from an assumption of basic human folly, competitiveness, and depravity, and contradicts Aristotle's assumption that man is by nature a "political animal." On the contrary, he is naturally antisocial; and, even when men meet for business and profit, only "a certain market-fellowship" is engendered. All society is only for gain or glory, and the only true equality among men is their power to kill each other. Hobbes sees and desires no other equality. Indeed, he specifically discouraged "men of low degree from a saucy behaviour towards their betters."

Significance of the *Leviathan*

The *Leviathan* horrified most of his contemporaries; Hobbes was accused of atheism and of "maligning the Human Nature." But, if his remedies were tactically impractical, in political philosophy he had gone very deep by providing the sovereign nation-state with a pragmatic justification and directing it to utilitarian ends.

Spinoza. The 17th-century Dutch philosopher Benedict de Spinoza also tried to make a scientific political theory, but it was more humane and more modern. Hobbes assumes a preindustrial and economically conservative society, but Spinoza, a Portuguese Jew born in Amsterdam, assumes a more urban setting. Like Hobbes, he is Cartesian, aiming at a scientific basis for political philosophy; but, whereas Hobbes was dogmatic and authoritarian, Spinoza desired toleration and intellectual liberty, by which alone human life achieves its highest quality. Spinoza, reacting against the ideological wars of religion and skeptical of both metaphysics and religious dogma, was a scientific humanist who justified political power solely by its usefulness. If state power breaks down and can no longer protect him or if it turns against him, frustrates, or ruins his life, then any man is justified in resisting it, since it no longer fulfills its purpose. It has no intrinsic divine or metaphysical authority.

In *Tractatus Theologico-Politicus* and the *Tractatus Politicus* Spinoza develops this theme. He intends, he writes, "not to laugh at men or weep over them or hate them, but to understand them." In contrast to St. Augustine, he glorifies life and holds that governments should not try to "change men from rational beings into beasts or puppets, but enable them to develop their minds and bodies in security and to employ their reason unshackled." The more life is enjoyed, he declares, the more the individual participates in the divine nature. God is immanent in the entire process of nature, in which all creatures follow the laws of their own being to the limit of their powers. All are bound by their own consciousness, and man creates his own values.

It seems that Spinoza thought good government approximated to that of the free burgesses of Amsterdam, a city in which religious toleration and relative political liberty had been realized. He is thus a pioneer of a scientific humanist view of government and of the neutrality of the state in matters of belief.

Richard Hooker's adapted Thomism. While out of the breakup of the medieval social order there emerged the humanist but sceptical outlook of Machiavelli, then the scientific humanist principles of Descartes, Hobbes, and Spinoza, from which the utilitarian and pragmatic outlook of modern times derives, another influential and politically important strain of political philosophy also emerged. During the Reformation and Counter-Reformation, Protestant and Catholic dogmatists denounced each other and even

attacked the authority of princes who, from interest or conviction, supported one side or the other. Political assassination became endemic, for both Protestant and Catholic divines declared that it was legitimate to kill an heretical ruler. Appeal was made to rival religious authority as well as to conscience. Men would resist authority and suffer execution rather than risk damnation, and in the resulting welter Hobbes and Spinoza advocated a sovereign state as the remedy. But other political philosophers salvaged the old Thomist concept of a divine cosmic order and of natural and human laws sanctioning the state. They also put forth the classical and medieval idea of the derivation of public power from the commonwealth as a whole and the responsibility of princes to the law. When Hobbes wrote that might makes right, he outraged such critics, who continued to assert that public power was responsible to God and the laws and that it was right to resist a tyrant who declared that the laws were in his own breast. This political theory was most influentially developed in England, where it inspired the constitutionalism that would also predominate in the United States.

Richard Hooker, an Anglican divine who wrote *Of the lawes of ecclesiasticall politie* (1593–1662), reconciled Thomist doctrines of transcendent and natural law, binding on all men, with the authority of the Elizabethan Anglican Church, which he defended against the Puritan appeal to conscience. Society, he argued, is itself the fulfillment of natural law, of which human and positive law are reflections, adapted to society. And public power is not something personal, for it derives from the community under law. Thus,

> The lawful power of making laws to command whole politie societies of men belongeth so properly unto the same entire societies, that for any prince . . . to exercise the same of himself . . . is no better than mere tyranny.

Such power can derive either directly from God or else from the people. The prince is responsible to God and the community; he is not, like Hobbes's ruler, a law unto himself. Law makes the king, not the king law.

Hooker, indeed, insisted that "the prince has a delegated power, from the Parliament of England, together with the convocation (of clergy) annexed thereto . . . whereupon the very essence of all government doth depend." This is the power of the crown in parliament in a balanced constitution. Hence an idea of harmonious government by consent. The Thomist medieval universal harmony had been adapted to the nation-state.

Locke. It was John Locke, politically the most influential English philosopher, who further developed this doctrine. His *Two Treatises of Government* (1690) were written to justify the Glorious Revolution of 1688–89, and his *Letter Concerning Toleration* (1689) was written with a plain and easy urbanity, in contrast to the baroque eloquence of Hobbes. Locke was a scholar, physician, and man of affairs, well-experienced in politics and business. As a philosopher he accepted strict limitations for mind, and his political philosophy is moderate and sensible, aimed at a balance among executive, judicial, and legislative powers, although with a bias toward the last.

His first *Treatise* was devoted to confuting the Royalist doctrine of patriarchal divine right by descent from Adam, an argument then taken very seriously and reflecting the idea of government as an aspect of a divinely ordained chain of being. If this order were broken, chaos would come about. The argument was part of the contemporary conflict of the ancients and the moderns.

Locke's view of the role and limits of public power

Locke tried to provide an answer by defining a limited purpose for political power, which purpose he considered to be "a right of making laws with penalties of death, and consequently all less penalties, for the regulating and preserving of property, and of employing the force of the community in execution of such laws, and in the defense of the commonwealth from foreign injury, and all this only for the public good." The authority of government derives from a contract between the rulers and the people, and the contract binds both parties. It is thus a limited power, proceeding according to established laws and "directed to no other end but the peace, safety, and public good of the people."

Whatever its form, government, to be legitimate, must govern by "declared and reasoned laws," and, since every man has a "property" in his own person and has "mixed his labour" with what he owns, government has no right to take it from him without his consent. It was the threat of attack on the laws, property, and the Protestant religion that had roused resistance to James II. Locke is expressing the concerns and interests of the landed and moneyed men by whose consent James's successor, William III, came to the throne, and his commonwealth is strictly conservative, limiting the franchise and the preponderant power to the propertied classes. Locke was thus no democrat in the modern sense and was much concerned to make the poor work harder. Like Hooker, he assumes a conservative social hierarchy with a relatively weak executive power and defends the propertied classes both against a ruler by divine right and against radicals. In advocating toleration in religion he was more liberal: freedom of conscience, like property, he argued, is a natural right of all men. Within the possibilities of the time, Locke thus advocated a constitutional mixed government, limited by parliamentary control of the armed forces and of supply. Designed mainly to protect the rights of property, it was deprived of the right of arbitrary taxation or imprisonment without trial and was in theory responsible to all the people through the politically conscious minority who were thought to represent them.

Though he was socially conservative, Locke's writings are very important in the rise of liberal political philosophy. He vindicates the responsibility of government to the governed, the rule of law through impartial judges, and the toleration of religious and speculative opinion. He is an enemy of the totalitarian state, drawing on medieval arguments and deploying them in practical, modern terms.

Burke. The Irishman Edmund Burke, while elaborating Whig constitutional doctrine expressed with such common sense by Locke, wrote with more emotion and took more account of time and tradition. While reiterating that government is responsible to the governed and distinguishing between a political society and a mere mob, he thought that governments were trustees for previous generations and for posterity. He made the predominant political philosophy of the 18th-century establishment appear more attractive and moral, but he wrote no great single work of political philosophy, expressing himself instead in numerous pamphlets and speeches.

In his early *Vindication of Natural Society* Burke is critical of the sufferings imposed by government, but his "Thoughts on the Cause of the Present Discontents" defines and defends the principles of the Whig establishment. He invoked a transcendent morality to sanction a constitutional commonwealth, but he detested abstract political theories in whose name men are likely to vivisect society. He set great store by ordered liberty and denounced the arbitrary power of the Jacobins who had captured the French Revolution. In his *Reflections on the Revolution in France* (1790) and *An Appeal from the New to the Old Whigs* (1791), he discerned in the doctrine of sovereignty of the people, in whose name the revolutionaries were destroying the old order, another and worse form of arbitrary power. No one generation has the right to destroy the agreed and inherited fabric of society, and "Neither the few nor the many have the right to govern by their will." A country is not a mere physical locality, he argued, but a community in time into which men are born, and only within the existing constitution and by the consent of its representatives can changes legitimately be made. Once the frame of society is smashed and its law violated, the people become a "mere multitude told by the head," at the mercy of any dictator who can seize power. He was realistic in predicting the consequences of violent revolution, which usually ends up in some kind of dictatorship. Burke, in sophisticated accents, spoke for the ancient and worldwide rule of custom and conservatism and supplied a needed romanticism to the calculating good sense of Locke.

Burke's concept of an obligation to posterity

Vico. The political philosophies hitherto surveyed contained little idea of progress. In antiquity the idea of cyclic recurrence predominated, and even 18th-century Chris-

tians believed that the world had been created in 4004 BC and would end in the Second Coming of Christ and a judgment. The 14th-century Arab philosopher of history Ibn Khaldūn of Tunis, in the *Muqaddimah* to his *Kitāb al-ʿibar,* had pioneered a vast sociological view of the historical process; but in western Europe it was a neglected Neapolitan philosopher, Giambattista Vico (1668–1744), who first interpreted the past in terms of the changing consciousness of mankind. His *Scienza nuova* (1725; revised edition 1744) interpreted history as an organic process involving language, literature, and religion and attempted to reveal the mentality or ethos of earlier ages: the age of the gods, the heroic age, and the human age, its climax and decadence. These ages recur, and each is distinguished by mythology, heroic poetry, and rational speculation respectively. In contrast to the legalistic, contractual, and static political philosophies then prevalent, Vico had discerned new horizons.

Montesquieu. This sort of vision was developed and elegantly popularized by the cosmopolitan French savant Montesquieu, whose work *The Spirit of Laws* (Eng. trans. 1750) won immense influence. It was an ambitious treatise on human institutions and a pioneer work of anthropology and sociology. Believing in an ordered universe—for "how could blind fate have produced intelligent beings?"—Montesquieu examined the varieties of natural law, varying customs, laws, and civilizations in different environments. He made the pedestrian good sense of Locke seem provincial, although he admired him and the British constitution. Unfortunately, he overemphasized the separation of executive, judicial, and legislative powers, considerable in Locke's day but by his own time tending to be concentrated in the sovereignty of Parliament. This doctrine much influenced the founders of the United States and the early French Revolutionaries.

Rousseau. The revolutionary romanticism of the Swiss-French philosopher Jean-Jacques Rousseau may be interpreted in part as a reaction to the analytic rationalism of the Enlightenment. He was trying to escape the aridity of a purely empirical and utilitarian outlook and attempting to create a substitute for revealed religion. Rousseau's *Émile* (1762) and *Du contrat social* (1762) proved revolutionary documents, and his posthumous *Considérations sur le gouvernement de Pologne* ("Considerations on the Government of Poland") contains desultory but often valuable reflections on specific problems.

There had been radical political slogans coined in medieval peasant revolts and in the 17th century, as in the Putney debates (1647) in the Cromwellian army, when a Puritan officer declared that "the poorest hee that is in England hath a life to live as the greatest hee," but the inspiration of these movements had been religion. Now Rousseau proclaimed a secular egalitarianism and a romantic cult of the common man. His famous sentence, "man is born free, but he is everywhere in chains," called into question the traditional social hierarchy: hitherto, political philosophers had thought in terms of elites, but now the mass of the people had found a champion and were becoming politically conscious.

Rousseau's idea of the general will

Rousseau was a romantic, given to weeping under the willows on Lake Geneva, and the *Social Contract* and *Discourses* are hypnotically readable, flaming protests by one who found the hard rationality of the 18th century too exacting. But man is not, as Rousseau claims, born free. Man is born into society, which imposes restraints on him. Casting about to reconcile his artificial antithesis between man's purported natural state of freedom and his condition in society, Rousseau utilizes the old theories of contract and transforms them into the concept of the "general will." This general will, a moral will that aims at the common good and in which all participate directly, reconciles the individual and the community by representing the will of the community as deriving from the will of moral individuals, so that to obey the laws of such a community is in a sense to follow one's own will, assuming that one is a moral individual.

Similar ideas to that of the general will became accepted as a basis for both the social-democratic welfare state and for totalitarian dictatorships. And, since the idea was misapplied from small village or civic communities to great sovereign nation-states, Rousseau was also a prophet of a nationalism that he never advocated. Rousseau himself wanted a federal Europe. He never wrote the proposed sequel to the *Du contrat social* in which he meant to deal with international politics, but he declared that existing governments lived in a state of nature, that their obsession with conquest was imbecilic, and that "if we could realize a European republic for one day, it would be enough to make it last for ever" (*Political Writings* I, pp. 365–388). But, with a flash of realism, he thinks the project impracticable, owing to the folly of men.

The incursion of this revolutionary romantic into political philosophy changed the climate of political opinion, for it coincided with the breakdown of the old dynastic order and the emergence first of the middle classes and then of the masses to political consciousness and power.

That the concept of general will was vague only increased its adaptability and prestige: it would both make constitutionalism more liberal and dynamic and give demagogues and dictators the excuse for "forcing people to be free" (that is, forcing people to follow the general will, as interpreted by the ruling forces). Rousseau could inspire liberals, such as the 19th-century English philosopher T.H. Green, to a creative view of a state helping people to make the best of their potential through a variety of free institutions. It could also play into the hands of demagogues claiming to represent the general will and bent on molding society according to their own abstractions.

THE 19TH CENTURY

Utilitarianism. A major force in the political and social thought of the 19th century was Utilitarianism, the doctrine that the actions of governments should be judged simply by the extent to which they promoted the "greatest happiness of the greatest number." The founder of the Utilitarian school was Jeremy Bentham, an eccentric Englishman trained in the law.

Bentham judged all laws and institutions by their utility thus defined. "The Fabric of Felicity," he wrote, "must be reared by the hands of reason and Law."

Bentham's *Fragment, on Government* (1776) and *Introduction to the Principles of Morals and Legislation* (1789) elaborated a Utilitarian political philosophy. Bentham was an atheist and an exponent of the new laissez-faire economics of Adam Smith and David Ricardo, but he inspired the spate of legislation that, after the Reform Bill of 1832, had tackled the worst consequences of 18th-century inefficiency and of the Industrial Revolution. His influence, moreover, spread widely abroad. At first a simple reformer of law, Bentham attacked notions of contract and natural law as superfluous, "The indestructible prerogatives of mankind," he wrote, "have no need to be supported upon the sandy foundation of a fiction." The justification of government is pragmatic, its aim improvement and to release the free choice of individuals and the play of market forces that will create prosperity. Bentham thought men far more reasonable and calculating than they are and brushed aside all the Christian and humanist ideas rationalizing instinctive loyalty and awe. He thought society could advance by calculation of pleasure and pain, and his *Introduction* even tries to work out "the value of a lot of pleasure and pain, how now to be measured." He compared the relative gratifications of health, wealth, power, friendship, and benevolence, as well as those of "irascible appetite" and "antipathy." He also thought of punishment purely as a deterrent, not as retribution, and graded offenses on the harm they did to happiness, not on how much they offended God or tradition.

Bentham's calculation of pain and pleasure

If Bentham's psychology was naïve, that of his disciple James Mill was philistine. Mill postulated an economic man whose decisions, if freely taken, would always be in his own interest, and he believed that universal suffrage, along with Utilitarian legislation by a sovereign parliament, would produce the kind of happiness and well-being that Bentham desired. In his *Essay on Government* (1828) Mill thus shows a doctrinaire faith in a literate electorate as the means to good government and in laissez-faire economics as a means to social harmony.

This Utilitarian tradition was humanized by James Mill's son, John Stuart Mill, one of the most influential of mid-Victorian liberals.

Whereas James Mill had been entirely pragmatic, his son tried to enhance more sophisticated values. He thought that civilization depended on a tiny minority of creative minds and on the free play of speculative intelligence. He detested conventional public opinion and feared that complete democracy, far from emancipating opinion, would make it more restrictive. Amid the dogmatic and strident voices of mid-19th-century nationalists, utopians, and revolutionaries, the quiet, if sometimes priggish, voice of mid-Victorian liberalism proved extremely influential in the ruling circles of Victorian England.

Accepting democracy as inevitable, J.S. Mill expressed the still optimistic and progressive views of an intellectual elite. Without complete liberty of opinion, he insisted, civilizations ossify. The quality of progress results not merely from the blind forces of economic competition but from the free play of mind. The worth of the state in the long run is only the worth of the individuals composing it, and without men of genius society would become a "stagnant pool." This militant humanist, unlike his father, was aware of the dangers of even benevolent bureaucratic power and declared that a state that "dwarfs its men" is culturally insignificant.

Mill also advocated the legal and social emancipation of women, holding that ability was wasted by mid-Victorian conventions. He believed that the masses could be educated into accepting the values of liberal civilization, but he defended private property and was as wary of rapid extensions of the franchise as of bureaucratic power.

Tocqueville. Mill's friend Alexis de Tocqueville, whose *De la démocratie en Amérique* (*Democracy in America*) appeared in 1835–40, was a French civil servant also concerned with maintaining the standards and creativeness of civilization in face of the rising tide of mass democracy. Since the United States was then the only large-scale democracy extant, Tocqueville decided to go there, and as a result of his visit wrote a classic account of early 19th-century American civilization. "We cannot," he wrote, "prevent the conditions of men from becoming equal, but it depends upon ourselves whether the principle of equality will lead them to servitude or freedom, to knowledge or barbarism, to prosperity or wretchedness." He feared the possible abuse of power by centralized government, unrestrained by the power of the old privileged classes, and thought it essential to "educate democracy" so that, although it would never have the "wild virtues" of the old regimes, it would have its own dignity, good sense, and even benevolence. Tocqueville greatly admired American representative institutions and made a penetrating analysis of the new power of the press. He realized, as few people then did, that the United States and Russia would become world powers, and he contrasted the freedom of the one and the despotism of the other. He also foresaw that under democracy education would be respected more as a ladder to success than for its intrinsic content and might thus become mediocre. He was alive to the dangers of uniform mediocrity but believed, like Mill, that democracy could be permeated by creative ideas.

T.H. Green. This kind of humanism was given a more elaborate philosophical content by the English philosopher T.H. Green, whose *Lectures on the Principles of Political Obligation* (1895; reprinted from *Philosophical Works*, vol. 2, 1885) greatly influenced the Liberals in the British governments of the period 1906–15. Green, like J.S. Mill and Tocqueville, wished to extend the minority culture to the people and even to use state power to "hinder hindrances to the good life." He had absorbed from Aristotle, Spinoza, Rousseau, and Hegel an organic theory of the state. The latter, by promoting the free play of spontaneous institutions, ought to help individuals both to "secure the common good of society [and] enable them to make the best of themselves."

While hostile to the abuse of landed property, Green was not a Socialist. He accepted the idea that property should be private and unequally distributed and thought the operation of the free market the best way to benefit the whole society; for free trade would, he thought, diminish the inequalities of wealth in a common prosperity. But Green would have extended the power of the state over education, health, housing and town planning, and the relief of unemployment—a new departure in Liberal thought. These recommendations are embedded in the most elaborate and close-knit intellectual construction made by any modern British political philosopher, and they laid the foundation of the British welfare state.

Green's writing on social reform

Liberal nationalism. Whereas Green shirked the extension of liberal and constitutional principles into international affairs, the Italian patriot and revolutionary prophet Giuseppe Mazzini made it his vision and became the most influential prophet of liberal nationalism. In his *The Duties of Man* and *Essays* he envisaged a harmony of free peoples—a "sisterhood of nations," in which the rule of military empires would be thrown off, the destruction of clerical and feudal privileges accomplished, and in which the emancipated peoples would be regenerated by means of education and universal suffrage. This vision inspired the more idealistic aspects of the Italian Risorgimento (national revival or resurrection) and of nationalistic revolts in Europe and beyond. Though, in fact, fervid nationalism often proved destructive, Mazzini advocated a united Europe of free peoples, in which national singularities would be transcended in a pan-European harmony. This sort of liberal democratic idealism was catching, and even if it frequently inspired Machiavellian policies, it also inspired President Woodrow Wilson of the United States, who, had he not been thwarted by domestic opposition, might well have made a Mazzinian-type League of Nations a success. Moreover, the Europe of the Common Market owes much to the apparently impractical liberal idealism of Mazzini.

American constitutionalism. The United States was founded by men deeply influenced by republicanism, by Locke, and by the optimism of the French Enlightenment. George Washington, John Adams, and Thomas Jefferson all concurred that laws, rather than men, should be the final sanction and that government should be responsible to the governed. But the influence of Locke and the Enlightenment was not entirely happy. John Adams, who followed Washington as president, prescribed a constitution with a balance of executive and legislative power checked by an independent judiciary. The federal constitution, moreover, could be amended only by a unanimous vote of the states. Anxious to safeguard state liberties and the rights of property, the founding fathers gave the federal government insufficient revenues and coercive powers, as a result of which the constitution was stigmatized as being "no more than a Treaty of Alliance." Yet the federal union was preserved. The civil power controlled the military, and there was religious toleration and freedom of the press and of economic enterprise. Most significantly, the concept of natural rights had found expression in the Declaration of Independence and was to influence markedly political and legal developments in the ensuing decades, as well as inspire the French Declaration of the Rights of Man.

Natural rights in American constitutional thought

Anarchism and utopianism. While a liberal political philosophy within a framework of capitalistic free trade and constitutional self-government dominated the greatest Western powers, mounting criticism developed against centralized government itself. Radical utopian and anarchist views, previously expounded mainly by religious sects, became secularized in such works as William Godwin's *Political Justice* (1793), Robert Owen's *New View of Society* (1813), and Pierre-Joseph Proudhon's voluminous and anticlerical writings.

The English philosopher William Godwin, an extreme individualist, shared Bentham's confidence in the reasonableness of mankind. He denounced the wars accepted by most political philosophers and all centralized coercive states. The tyranny of demagogues and of "multitudes drunk with power" he regarded as being as bad as that of kings and oligarchs. The remedy, he thought, was not violent revolution, which produces tyranny, but education and freedom, including sexual freedom. His was a program of high-minded, atheistic anarchy.

The English Socialist Robert Owen, a cotton spinner who had made a fortune, also insisted that bad institu-

tions, not original sin or intrinsic folly, caused the evils of society, and he sought to remedy them by changing the economic and educational system. He thus devised a scheme of model cooperative communities that would increase production, permit humane education, and release the naturally benevolent qualities of mankind.

The French moralist and advocate of social reform Pierre-Joseph Proudhon attacked the "tentacular" nation-state and aimed at a classless society in which major capitalism would be abolished. Self-governing producers, no longer slaves of bureaucrats and capitalists, would permit the realization of an intrinsic human dignity, and federation would replace the accepted condition of war between sovereign states. Proudhon tried to transform society by rousing the mass of the people to cooperative humanitarian consciousness.

Saint-Simon and Comte. Another revolt against the prevalent establishment, national and international, was made by the French social philosopher Henri de Saint-Simon. Saint-Simon wanted to develop the Industrial Revolution so as to ameliorate the condition of the poorest class. This would be achieved not through political revolution, but through a government of bankers and administrators who would supersede kings, aristocrats, and politicians. If France were suddenly deprived of three thousand leading scientists, engineers, bankers, painters, poets, and writers, he argued, the result would be catastrophic; but if all the courtiers and bishops and 10,000 landowners vanished, the loss, though deplorable, would be much less severe. Saint-Simon also demanded a united Europe, superseding the warring nation-states, with a European parliament and a joint development of industry and communication. He also invented a synthetic religion appropriate to a scientific phase of history, with a cult of Newton and the great men of science.

Comte's "religion of humanity"

Saint-Simon's disciple Auguste Comte went further. His *Course of Positive Philosophy* (1830–42) and *System of Positive Polity* (1851–54) elaborated a "religion of humanity," with ritual, calendar, and a priesthood of scientists, and secular saints, including Julius Caesar, Dante, and Joan of Arc. Society would be ruled by bankers and technocrats and Europe united into a Western republic. This doctrine, backed by pioneering sociology, won much influence among intellectuals. Comte, like Saint-Simon, tackled the essential questions: how to deploy the power of modern technology for the benefit of all mankind; how to avoid wars between sovereign states; and how to fill the void left by the waning of Christian beliefs.

Hegel. Whereas the utopian reformers had discarded metaphysical arguments, the German philosopher G.W.F. Hegel claimed to apprehend the totality of the cosmos by speculative cognition. Like Vico, he saw the past in terms of changing consciousness, but he viewed the historical process as one of "becoming" rather than as one of eternal recurrence. Hegel had no adequate historical data for his intuitions, since the whole of world history was even less known then than it is today, but his novel sweep and range of theory proved an intoxicating substitute for religion. He divided world history into four epochs: the patriarchal Eastern empire, the brilliant Greek boyhood, the severe manhood of Rome, and the Germanic phase after the Reformation. The "Absolute," like a conductor, summons each people to their finest hour, and neither individuals nor states have any rights against them during their historically determined period of supremacy. Many felt some sense of anticlimax, however, when he claimed that the Prussian state embodied the hitherto highest self-realization of the "Absolute" (see HEGELIANISM). Not since St. Augustine had so compelling a drama been adumbrated. Hegel's drama, moreover, culminates in this world, for "the state is the divine idea as it exists on earth."

Marx and Engels. Hegel was a conservative, but his influence on the revolutionaries Karl Marx and his collaborator Friedrich Engels was profound. They inherited the Hegelian claim to understand the "totality" of history and life as it progressed through a dialectic of thesis, antithesis, and synthesis. But, whereas Hegel envisaged a conflict of nation-states, Marx and Engels thought that the dynamism of history was generated by inevitable class conflict economically determined. This was an idea even more dynamic than Hegel's and more relevant to the social upheavals that were a consequence of the Industrial Revolution. Marx was a formidable prophet whose writings lead up to an apocalypse and redemption. A deeply learned humanist, his ideal was the fullest development of the human personality. But, whereas Plato was concerned with an elite, Marx cared passionately for the elevation of whole peoples.

The Marxist credo was all the more effective as it expressed with eloquent ferocity the grievances of the poor, while prophesying retribution and a happy ending. For the state, once captured by the class-conscious vanguard of the proletariat, would take over the means of production from the capitalists, and a brief "dictatorship of the proletariat" would establish a truly communist society. The state would then wither away and man at last become "fully human" in a classless society.

The powerful slogans of Marx and Engels were a natural result of the unbridled capitalism of laissez-faire, but politically they were naïve. In classical, medieval, and humanistic political philosophy the essential problem is the control of power, and to imagine that a dictatorship, once established, will wither away is utopian. As even Marx's fellow revolutionary the Russian anarchist M.A. Bakunin observed,

Marx's naïveté regarding the dictatorship of the proletariat

> The revolutionary dictatorship of the doctrinaires who put science before life would differ from the established state only in external trappings. The substance of both are a tyranny of the minority over the majority—in the name of the many and the supreme wisdom of the few.

The revolutionaries would vivisect society in the name of dogmas and "destroy the present order, only to erect their own rigid dictatorship among its ruins." For a full account of Marxist philosophy, see MARXISM.

Political philosophy in the 20th century

Nineteenth-century European civilization had been the first to dominate and pervade the whole world and to create a new self-sustaining productivity in which all eventually might share. But, as Saint-Simon had pointed out, this civilization had a fatal flaw. The rule of law, accepted within the politically advanced states, had never been achieved among them. Heavily armed nations and empires remained in a Hobbesian "posture of war," and classical and medieval ideals of world order had long been discarded. Within states, also, laissez-faire capitalism had exacerbated class conflicts, while the decline of religious belief had undermined traditional solidarity. And in 1914, when a general European war broke out, the peoples, contrary to the hopes of cosmopolitan revolutionaries, rallied behind their national governments. When the victorious powers failed to promote world order through the League of Nations, a second global conflict followed, during which were developed weapons so destructive as to threaten life everywhere.

In the aftermath of these catastrophes and the worldwide revulsion they occasioned, not least against the European colonial powers, three mainstreams of mid-20th-century political philosophy may be discerned.

In liberal-constitutional states, with modified, managerial capitalism and various degrees of public welfare, a political pragmatism has emerged, still maintaining the Aristotelian distinction between the rule of law and government by consent, on the one hand, and tyranny on the other. Second, there has been a reaffirmation of religious or quasi-religious values appealing to conscience and the inner man, expressed persuasively in Existentialist writings. Third, revolutionary ideas have also developed, most of them along Marxist lines. Other revolutionary doctrines appeal to anarchist traditions and are elaborated with neo-Marxist and neo-Freudian insights. Within these categories many shades of opinion are expressed, and only a sampling of representative views is presented here.

POLITICAL PRAGMATISM

The first, pragmatist approach probably has been most powerfully asserted in the United States and Great Britain.

The American writer Lewis Mumford, for example, has advocated a militant humanism, defending people against the alienations of megalopolitan life and attacking mechanization and materialism. Like the Greek philosophers and like Tocqueville, whom he admires, Mumford declares, "In the end, all our contrivances have but one object; the continued growth of human personalities and the cultivation of the best life possible." The American philosopher and educationist John Dewey, on the other hand, sought to counteract the dehumanization of industrial mass society by a freer form of education, liberating the personality.

Both these writers criticize the existing structure of society and its modified capitalism, but try to work within it. Another humanist, the English philosopher Bertrand Russell, was more radical. Russell carried into political philosophy an aristocratic individualism, campaigning for toleration, sexual freedom, compassion, and common sense. He broadcast elite values to a mass society and attacked materialism, crass bureaucracy, and war. He twice went to prison in pacifist protest and was obsessed with the universal menace of nuclear weapons. He denounced warlike political theories: "Remember your humanity," he said, "and forget the rest." On political tactics often inept, Russell won wide influence as a man of principle, concerned to adapt archaic institutions to the changed environment of mankind.

The Austrian-born British philosopher Sir Karl Popper has demonstrated the pretensions of the 19th-century determinist philosophies such as those of Hegel and Marx, while an English historian and philosopher, Sir Isaiah Berlin, has ridiculed the idea of a supposedly objective march of histoy. Berlin also rejects the Marxist belief that all values are conditioned by the place men occupy on the "moving stair of time." Marx, he points out, was as romantic as Hegel in envisaging a "world which moves from explosion to explosion in order to fulfil the great cosmic design." Moral values, he insists, are not just a "subjective gloss unworthy of consideration on the great hard edifice of historical construction." No single formula can be found, Berlin argues, whereby the various objectives of men can be harmoniously realized. There are many human goals, which may well be in conflict with one another.

This empirical, pluralist, and liberal political philosophy has much in common with the approach of the Frenchman Émile Durkheim and the Englishman Graham Wallas, both founding fathers of modern sociology. Statesmen and political philosophers, they contend, should not play the part of prophets but rather confine themselves to investigating social patterns and the ideas that are part of them. Ways might thus be found of promoting the survival and vitality of a given society in its particular setting.

Scientific humanism of Graham Wallas

Graham Wallas was concerned to adapt constitutional societies by consent. He wanted to nationalize many essential means of production, including transport and communications, and through increased taxation strengthen social democracy by greater economic and social equality. He was not a revolutionary but a reformer, who understood the precariousness of civilization and the dangers of nationalism, which could only bring, he prophesied, centuries of warfare and regression. He advocated a worldwide and constitutionalist scientific humanism, inspired by the idea of the solidarity of the whole species, for "the master task of civilized mankind is to promote the conditions leading to the good life."

Other political sociologists who accepted the established order did not expect to improve it. The Italian Vilfredo Pareto, and Gaetano Mosca, a Sicilian-born lawyer, set themselves not to state what they wanted but to record what occurs in society. Pareto's *Mind and Society* (1916) is an elaborate, quasi-mathematical classification of nonlogical political myths. Its form is daunting, but its insights are penetrating, especially a hilarious dissection of Rousseau's *General Will*, of which, Pareto concludes, "the intrinsic logico-experimental value . . . is zero." Ranging sardonically over history, Pareto insists that elites will always manipulate society, power merely shifting from one set of rulers to another.

Mosca, in *The Ruling Class* (1939), analyzed how political myths are exploited. He also concluded that elites everywhere are bound to rule and that the least bad government occurs when abuse of power is checked by legal means; that is, by the rule of law. Mosca admired the liberal constitutionalism of the 19th century, although he was aware of its precariousness and limitations. He argued that there is no total explanation of history, which has always been the unpredictable outcome of competing and interacting interests. One thing is certain, nevertheless: in various forms there will always be a struggle for predominance. Mosca's views, more clearly set out than Pareto's, have a salutary realism.

The American philosopher and critic James Burnham also analyzed shifts of power. In *The Managerial Revolution* (1941) he propounded a theory of bureaucratic revolution: the rulers of the new society, the class with power and privilege, will be the bureaucratic managers of "super states." In *The Machiavellians, Defenders of Freedom* (1943), he reinterprets Machiavelli and cites Mosca as a modern Machiavellian. Following Pareto's idea of the "circulation of elites," he asserts that, when a ruling class becomes inadequate, frivolous, or bored, loses confidence in itself and its myths, and becomes irresolute in deploying necessary force, new elites are bound to take over—as in the managerial revolution of the 20th century.

RELIGIOUS AND EXISTENTIALIST APPROACHES

In the second religious and quasi-religious group of political philosophies, the Catholic hierarchy has reiterated its ancient neo-Thomist doctrine of original sin and redemption. Pope Leo XIII, in the encyclicals *Inscrutabili Dei Consilio* (1878), *Immortale Dei* (1885), and *Rerum Novarum* (1891), dismissed all anthropocentric political philosophies as new versions of old heresies. The world, "through an insatiable craving for things perishable," was "rushing wildly upon the straight road to destruction." Society is intelligible only in the light of the Christian revelation and a future life:

> exclude the idea of futurity and forthwith the very notion of what is good and right would perish, nay, the whole scheme of the universe would become a dark and unfathomable mystery.

Such is the human condition that visionary innovations are fruitless, and "venomous" teachings can only bring "death-bearing fruit." Society, as St. Augustine had declared, if organized without God, can only be a present hell. Hierarchy, authority, and censorship can alone "control the excesses of the unbridled intellect, which unfailingly end in the oppression of the untutored multitude." Property is essential to the family, on which the social order depends, and inequality is inherent in all human societies. Only a harmonious Christian commonwealth can assuage the consequences of sin, and within that social order the state should therefore encourage Christian trade unions and promote the welfare of the poor. Thus, with these views the papacy, maintaining its monopoly of revelation, tried to come to terms with the demands of industrial civilization.

During the rise of 19th-century nationalism and of Communist, Fascist, and Nazi dictatorships, and in face of the increasing dominance of governments and large-scale industry in all mass societies, the importance of individual responsibility with regard to moral issues was emphasized by a divergent group of thinkers who have come to be described as Existentialists. Søren Kierkegaard (died 1855), a Danish philosopher, declared that "truth is subjectivity" and that only by means of inward revelation can man know God. Jean-Paul Sartre, a brilliant French Existentialist, tried to come to terms with dialectical Materialism. His *Existentialism and Humanism* (1948) comprises an affirmation of human dignity. "If," he writes, "I have excluded God the Father, there must be someone to invent values." Man, who has abandoned God, "must liberate himself by some practical commitment," for only then can he become fully human. Sartre's elaborate *L'Etre et le néant* (1943; Eng. trans., *Being and Nothingness*, 1956) is at once Cartesian and laborious, complete with a "key" to its special and pedantic terms. It investigates the loneliness of the human condition, attitudes to others, love,

Sartre's view of human dignity

masochism, indifference, desire, and hate. This intense introspection is even more vividly expressed in his fiction and drama. The Algerian-born Albert Camus in *The Myth of Sisyphus* (1942) and *The Rebel* (1951) also agonizes brilliantly over the current human condition, and in *Man in Revolt* he discards hope of pragmatic improvement.

REVOLUTIONARY DOCTRINES

The third stream of contemporary political philosophy is Marxist–Leninist totalitarian and neo-Marxist anarchist. Many of Marx's original insights into the socio-economic process and its effect on ideas are now generally accepted. His prophecies, on the other hand, have not been fulfilled. The proletarian revolution, for example, came not in an economically advanced country but in one of the most backward; and the state, far from withering away or being diminished by inexorable economic trends, has in fact become more powerful both in Communist and in social-democratic countries. Those who have accepted the total Marxist revelation as superseding all else have had thus to adapt and revise it. Hence, much tortuous and artificial debate has ensued. All orthodox Marxists accept the Hegelian position that one can get beyond empirical knowledge and perceive the historically revealed installments of a total explanation. They also start from Marx's 19th-century belief that economically determined conflicts among feudal, bourgeois, and proletarian classes are the dynamic of history and that the rule of law is not a safeguard for the whole society against arbitrary power but merely the expression of class interest.

Doctrinal assumptions of latter-day Marxists

Lenin. The first and by far the most significant interpretation of Marx's doctrine as realized in the Soviet Union was made by Lenin and developed by Stalin and is entirely authoritarian. According to Marx and Engels, the revolution could occur only after the bourgeois phase of production had "contradicted" the tsarist order, but Lenin determined to take advantage of the opportunities provided by World War I and settle accounts directly with the "accursed heritage of serfdom, of Asiatic barbarism . . . an insult to mankind," and in 1917 he engineered a coup that secured the support of the peasantry and the industrial workers. He also adopted the revolutionary theorist Leon Trotsky's idea of a "permanent revolution" from above by a small revolutionary elite.

Already in *What Is To Be Done?* (1902), Lenin had argued that an educated elite must direct the proletarian revolution, and when he came to power he dissolved the constituent assembly and ruled through a "revolutionary and democratic dictatorship supported by the state power of the armed workers." In asserting the need for an elite of professional revolutionaries to seize power, Lenin reverted to Marx's program in *The Communist Manifesto* rather than conforming to the fated pattern of economic development worked out in *Das Kapital.*

In 1921 he further adapted theory to the times. His new economic policy sanctioned the development of a class of prosperous "kulak" peasantry to keep the economy viable. For Lenin always thought in terms of world revolution; and, in spite of the failure of the Marxists in central Europe and the defeat of the Red armies in Poland, he died in the expectation of a global sequel. Thus, in *Imperialism, the Highest Stage of Capitalism* (1917), he had extended the class war into an inevitable conflict between European imperialism and the colonial peoples involved. He had been influenced by the English historian J.A. Hobson's *Imperialism, a Study* (1902), which alleged that decadent capitalism was bound to turn from glutted markets at home to exploit the toil of "reluctant and unassimilated peoples."

But, as observed by classical, medieval, and modern constitutionalist political philosophers, authoritarian regimes suffer the tensions of all autocracies. Marx himself might have thought that such planned autocracies had made the worst of his revelation.

Other Marxist approaches. Many Marxist revisionists tend toward anarchism, stressing the Hegelian and utopian elements of his theory. The Hungarian György Lukács, for example, and the German Herbert Marcuse, who fled from the Nazis to the United States, have won some following among those in revolt against both authoritarian "peoples' democracies" and the diffused capitalism and meritocracy of the managerial welfare state. Lukács' *Geschichte und Klassenbewusstsein* (1923; Eng. trans., *History and Class Consciousness,* 1971), a neo-Hegelian work, claims that only the intuition of the proletariat can properly apprehend the totality of history. But world revolution is contingent, not inevitable, and Marxism is an instrument, not a prediction. Lukács renounced this heresy after residence in the Soviet Union under Stalin, but he maintained influence through literary and dramatic criticism. After Khrushchev's denunciation of Stalin, Lukács advocated peaceful coexistence and intellectual rather than political subversion. In *The Meaning of Contemporary Realism* (trans. 1963), he again relates Marx to Hegel and even to Aristotle, against the Stalinist claim that Marx made a radically new departure. Lukács' neo-Marxist literary criticism can be tendentious, but his neo-Hegelian insights, strikingly expressed, have appealed to those anxious to salvage the more humane aspects of Marxism and to promote revolution, even against a modified capitalism and social democracy, by intellectual rather than by political means.

Marcuse's assessment of revolutionary prospects

Marcuse also reached back to the more utopian Marx. Now that most of the proletariat has been absorbed into a conformist managerial capitalism or has been regimented into bureaucratic peoples' democracies, freedom, argues Marcuse, is in retreat. In Western affluent societies most employers and workers are equally philistine, dominated by the commercialized mass media, or "cogs in a culture machine." The former Soviet Union had reverted to an even more philistine monolithic repression, distorting art and literature. This enslavement of man by his own industrial productivity had been clinched by the colossal power of governments, which rendered the old brief and brisk class warfare a romantic, impracticable idea. Marcuse attacked all establishments and transferred the redeeming mission of the proletariat to a fringe of alienated minorities—radical students and the exponents of the "hippie" way of life—as well as to Viet Cong guerrillas and Black Power militants. Such groups, he declared, could apparently form liberating elites and destroy the managerial society. Thus reappeared the old Marxist–Hebraic pattern of redemption through struggle by a chosen people.

The Italian Communist Antonio Gramsci deployed a vivid rhetorical talent in attacking existing society. Like Marcuse, Gramsci was alarmed that the proletariat was being assimilated by the capitalist order. He took his stand on the already obsolescent Marxist doctrine of irreconcilable class war between bourgeois and proletariat. He aimed to unmask the bourgeois idea of liberty and to replace parliaments by an "implacable machine" of workers' councils, which would destroy the current social order through a dictatorship of the proletariat. "Democracy," he wrote, "is our worst enemy. We must be ready to fight it because it blurs the clear separation of classes."

Not only would parliamentary democracy and established law be unmasked, but culture, too, would be transformed. A workers' civilization, with its great industry, large cities, and "tumultuous and intense life," would create a new civilization with new poetry, art, drama, fashions, and language. Gramsci insisted that the old culture should be destroyed and that education should be wrenched from the grip of the ruling classes and the church.

But this militant revolutionary was also a utopian. He turned bitterly hostile to Stalin's regime, for he believed, like Engels, that the dictatorship of the workers' state would wither away. "We do not wish," he wrote, "to freeze the dictatorship." Following world revolution, a classless society would emerge, and mankind would be free to master nature instead of being involved in a class war.

Since World War II, Gramsci's notions have enjoyed a minor revival. They appeal to the fringe of revolutionaries who admire Marcuse and detest the embourgeoisement of an idealized proletariat. But, in a civilization in which, if total war can be avoided, material prospects are good, the destruction of the old culture out of rage, envy, and naïve idealism appears to be a pointless program. Like Marcuse's doctrine, it is a cry of pain, typical of the 1920s in Italy.

Conclusion

The history of political philosophy from Plato until the present day makes plain that modern political philosophy is still faced with the basic problems defined by the Greeks. The need to redeploy public power in order to maintain the survival and enhance the quality of human life, for example, has never been so essential. And, if the opportunities for promoting well-being are now far greater, the penalties for the abuse of power are nothing less than the destruction or gross degradation of all life on the planet.

In these circumstances it is of no great importance that some analytical philosophers have declared themselves neutral; they have at least often discredited pretentious metaphysical myths. On the empirical evidence, constitutionalism and the rule of law, with the ancient classic, medieval, and humanist traditions behind them, have proved themselves a more successful response to the environment than tyranny and repression. In the current and more sophisticated view, there are no shortcuts to the millennium. As Mosca points out, utopian ideas become

> dangerous when they succeed in bringing a large mass of intellectual and moral energies to bear upon an end that can never be achieved, and that in the day of purported achievement can mean nothing more than the triumph of the worst people and distress and disappointment for the good.

There will perhaps always be a struggle for preeminence in any society, and public laws are necessary to regulate it. Too much cannot be hoped of government, and the best society is that in which tyranny and caprice of power are prevented and in which men are free to create diverse and spontaneous institutions within the framework of law. Only within such a framework of a tolerably well-organized constitutionalism, gradually extended to relations between states, can the swiftly mounting opportunities provided by applied science be taken and the pattern of social life adjusted, so that the human species, instead of being thwarted and deformed by its institutions, can realize its full potentialities. (J.E.B.)

BIBLIOGRAPHY

General histories of political philosophy. GEORGE HOLLAND SABINE, *A History of Political Theory,* 4th ed. rev. by THOMAS LANDON THORSON (1973), provides a comprehensive survey. WILLIAM ARCHIBALD DUNNING, *A History of Political Theories,* 3 vol. (1902–20, reissued 1936–38), is still valuable. Also of interest is K.R. POPPER, *The Open Society and Its Enemies,* 5th ed., rev., 2 vol. (1966). Additional surveys that will be useful include LEO STRAUSS and JOSEPH CROPSEY (eds.), *History of Political Philosophy,* 2nd ed. (1972, reprinted 1981); LEO RAUCH, *The Political Animal: Studies in Political Philosophy from Machiavelli to Marx* (1981); and ANTHONY PAGDEN (ed.), *The Languages of Political Theory in Early-Modern Europe* (1987). ERNEST BARKER, *Principles of Social & Political Theory* (1951, reissued 1980), analyzes essential problems. Other important works are JOHN BOWLE, *Politics and Opinion in the Nineteenth Century* (1954, reissued 1966); WILLIAM EBENSTEIN, *Modern Political Thought: The Great Issues,* 2nd ed. (1960); HAROLD D. LASSWELL, *The Future of Political Science* (1962, reissued 1974); and JOSEPH CROPSEY, *Political Philosophy and the Issues of Politics* (1977, reissued 1980).

Antiquity. C.M. BOWRA, *The Greek Experience* (1957, reissued 1985), is concerned with the social background of Greek thought. PETER BROWN, *The World of Late Antiquity* (1971, reissued 1989), is also of interest.

Plato: PLATO, *The Republic of Plato,* trans. by A.D. LINDSAY, 2nd ed. (1908, reissued 1992), and another edition with the same title trans. by ALLAN BLOOM, 2nd ed. (1991), *The Statesman,* trans. by HAROLD N. FOWLER (1925, reissued 1990), the *Loeb Classical Library* edition, and *The Laws of Plato,* trans. by A.E. TAYLOR (1934, reissued 1969); JULIA ANNAS, *An Introduction to Plato's Republic* (1981); DAVID GRENE, *Greek Political Theory: The Image of Man in Thucydides and Plato* (1965), providing further discussion of Greek political thought; for more advanced students, LEO STRAUSS, *Studies in Platonic Political Philosophy* (1983).

Aristotle: ARISTOTLE, *The Politics of Aristotle,* trans. by ERNEST BARKER (1946, reissued 1972); CARNES LORD and DAVID K. O'CONNOR (eds.), *Essays on the Foundations of Aristotelian Political Science* (1991), a useful survey of contemporary interpretations.

Cicero and the Stoics: MARCUS TULLIUS CICERO, *On the Commonwealth,* trans. by GEORGE HOLLAND SABINE and STANLEY BARNEY SMITH (1929, reprinted 1976), and *De legibus,* trans. by CLINTON WALKER KEYES (1928, reissued 1966).

St. Augustine: ST. AUGUSTINE, *The City of God,* trans. by JOHN HEALEY and ed. by R.V.G. TASKER, 2 vol. (1945, reissued 1972); CHARLES NORRIS COCHRANE, *Christianity and Classical Culture: A Study of Thought and Action from Augustus to Augustine* (1940, reissued 1974), worth consulting.

The Middle Ages. Broad treatments include R.W. CARLYLE and A.J. CARLYLE, *A History of Mediæval Political Theory in the West,* 6 vol. (1903–36), a general survey; A.L. SMITH, *Church and State in the Middle Ages* (1913, reprinted 1964); HENRY OSBORN TAYLOR, *The Mediæval Mind,* 4th ed., vol. 2 (1925, reissued 1971), still valuable for patristic and medieval thought; CHARLES HOWARD MCILWAIN, *The Growth of Political Thought in the West, from the Greeks to the End of the Middle Ages* (1932, reissued 1968); RALPH LERNER and MUHSIN MAHDI (eds.), *Medieval Political Philosophy* (1963, reissued 1972); and J.H. BURNS (ed.), *The Cambridge History of Medieval Political Thought c. 350–c. 1450* (1988).

John of Salisbury: JOHN OF SALISBURY, *The Statesman's Book of John of Salisbury,* trans. by JOHN DICKINSON (1927).

Aquinas: AQUINAS, *Selected Political Writings,* ed. by A.P. D'ENTRÈVES (1948, reissued 1984).

Dante: A.G. FERRERS HOWELL and PHILIP H. WICKSTEED (eds.), *A Translation of the Latin Works of Dante Alighieri* (1904), including *De Monarchia;* as does DONALD NICHOLL, *Monarchy* (1954); concerning Dante's philosophy, A.P. D'ENTRÈVES, *Dante As a Political Thinker* (1952, reprinted 1965).

The 16th to the 18th centuries. Essays by several political thinkers of this period can be found in ERNEST BARKER (ed.), *Social Contract: Essays by Locke, Hume, and Rousseau* (1947, reissued 1980), an excellent translation of these essays. Two of the best histories of political thought during this period are QUENTIN SKINNER, *The Foundations of Modern Political Thought,* 2 vol. (1978); and J.H. BURNS and MARK GOLDIE (eds.), *The Cambridge History of Political Thought, 1450–1700* (1991). A solid survey of social contract theory is presented in PATRICK RILEY, *Will and Political Legitimacy: A Critical Exposition of Social Contract Theory in Hobbes, Locke, Rousseau, Kant, and Hegel* (1982). J.B. BURY, *The Idea of Progress* (1920, reissued 1987); and CARL L. BECKER, *The Heavenly City of the Eighteenth Century Philosophers* (1932, reissued 1991), both deal with the French Enlightenment generally. ALFRED NORTH WHITEHEAD, *Science and the Modern World* (1925, reissued 1967), provides background on the 18th and 19th centuries. The political thought of the American founding fathers can be found in ALEXANDER HAMILTON, JAMES MADISON, and JOHN JAY, *The Federalist* (1788); one of the most accessible editions was ed. by JACOB E. COOKE (1961, reprinted 1989).

Machiavelli: NICCOLÒ MACHIAVELLI, *The Prince, and The Discourses,* ed. by MAX LERNER (1940, reissued 1950), and *The Literary Works of Machiavelli,* ed. and trans. by J.R. HALE (1961, reprinted 1979); ALLAN H. GILBERT, *Machiavelli's Prince and Its Forerunners* (1938, reissued 1968); H. BUTTERFIELD, *The Statecraft of Machiavelli* (1940, reissued 1962); FELIX RAAB, *The English Face of Machiavelli* (1964).

Hobbes: THOMAS HOBBES, *Leviathan,* ed. by A.D. LINDSAY (1914, reissued 1970), and another edition with the same title ed. by C.B. MACPHERSON (1968, reissued 1987), and *De Cive: The English Version ...* and *De Cive: The Latin Version ...,* both ed. by HOWARD WARRENDER (1983), the best editions of the English and Latin versions of Hobbes's *De Cive;* K.V. THOMAS, "The Social Origins of Hobbes's Political Thought," in K.C. BROWN (ed.), *Hobbes: Studies* (1965), an excellent account; LEO STRAUSS, *The Political Philosophy of Hobbes: Its Basis and Its Genesis* (1936, reissued 1984), an enlightening study; JOHN BOWLE, *Hobbes and His Critics* (1951, reissued 1969), describing Hobbes's political impact and contemporary reaction to it.

Spinoza: BENEDICTUS DE SPINOZA, *The Political Works: The Tractatus theologico-politicus in Part, and the Tractatus politicus in Full,* ed. and trans. by A.G. WERNHAM (1958, reissued 1965); STUART HAMPSHIRE, *Spinoza* (1951, reprinted with revisions, 1988).

Richard Hooker's adapted Thomism: C.J. SISSON, *The Judicious Marriage of Mr. Hooker and the Birth of The Laws of Ecclesiastical Polity* (1940, reprinted 1974).

Locke: JOHN LOCKE, *Two Treatises of Government,* ed. by PETER LASLETT (1960, reissued 1988), and *The Second Treatise of Government (An Essay Concerning the True Original, Extent, and End of Civil Government), and, A Letter Concerning Toleration,* ed. by J.W. GOUGH, 3rd ed. (1966, reprinted 1976); JOHN DUNN, *The Political Thought of John Locke* (1969, reissued 1982), a historical account of Locke's arguments.

Burke: EDMUND BURKE, *The Works of the Right Honorable Edmund Burke,* 8 vol. (1854–58), and *Burke's Politics: Selected Writings and Speeches on Reform, Revolution, and War,* ed. by

ROSS J.S. HOFFMAN and PAUL LEVACK (1949, reissued 1967); ALFRED COBBAN, *Edmund Burke and the Revolt Against the Eighteenth Century,* 2nd ed. (1960).

Vico: GIAMBATTISTA VICO, *The New Science of Giambattista Vico,* trans. by THOMAS GODDARD BERGIN and MAX HAROLD FISCH, rev. ed. (1968); BENEDETTO CROCE, *The Philosophy of Giambattista Vico* (1913, reissued 1964; originally published in Italian, 1911); H.P. ADAMS, *The Life and Writings of Giambattista Vico* (1935, reprinted 1970).

Montesquieu: CHARLES DE SECONDAT, BARON DE MONTESQUIEU, *The Spirit of Laws,* trans. by THOMAS NUGENT, new ed., rev. by J.V. PRICHARD, 2 vol. (1914, reprinted 1991).

Rousseau: JEAN JACQUES ROUSSEAU, *The Political Writings of Jean Jacques Rousseau,* ed. by C.E. VAUGHAN, 2 vol. (1915, reprinted 1971); WILLIAM H. BLANCHARD, *Rousseau and the Spirit of Revolt* (1967); and J. MCMANNERS, *The Social Contract and Rousseau's Revolt Against Society* (1968), both quite illuminating.

The 19th century. BASIL WILLEY, *Nineteenth Century Studies* (1949, reissued 1980), treats a variety of topics in addition to 19th-century philosophy. A selection of the works of Nietzsche can be found in R.J. HOLLINGDALE (compiler and trans.), *A Nietzsche Reader* (1977).

Utilitarianism: JEREMY BENTHAM, *A Fragment on Government, and An Introduction to the Principles of Morals and Legislation,* ed. by WILFRID HARRISON (1948, reprinted 1967); J.S. MILL, *On Liberty, and, Considerations on Representative Government,* ed. by R.B. MCCALLUM (1946); JOHN PLAMENATZ, *The English Utilitarians,* 2nd rev. ed. (1958, reprinted 1966), particularly for Mill.

Tocqueville: ALEXIS DE TOCQUEVILLE, *Democracy in America,* ed. by PHILLIPS BRADLEY, trans. from French, 2 vol. (1945, reissued 1990), an excellent edition, also available in an abridged version, *Democracy in America,* ed. by HENRY STEELE COMMAGER (1946), and *De Tocqueville's L'Ancien Régime,* ed. by G.W. HEADLAM (1904, reissued as *L'Ancien Régime,* 1969).

Anarchism and utopianism: JAMES JOLL, *The Anarchists,* 2nd ed. (1979), tracing the history of anarchism; ROBERT OWEN, *A New View of Society and Other Writings* (1927, reissued 1991), the *Everyman's Library* edition; PIERRE-JOSEPH PROUDHON, *What Is Property? An Enquiry into the Principle of Right and of Government,* trans. by BENJ. R. TUCKER (1876, reissued 1970; originally published in French, new ed., 1867); D.W. BROGAN, *Proudhon* (1934).

Comte: AUGUSTE COMTE, *System of Positive Polity,* trans. from French, 4 vol. (1875–77, reprinted 1973); J.S. MILL, *Auguste Comte and Positivism* (1865, reprinted 1993); EDWARD CAIRD, *The Social Philosophy and Religion of Comte* (1885, reissued 1968).

Hegel: GEORG WILHELM FRIEDRICH HEGEL, *Hegel's Philosophy of Right,* trans. by T.M. KNOX (1942, reissued 1967; originally published in German, 1821), and *The Philosophy of History,* trans. by J. SIBREE, rev. ed. (1899, reissued 1991; originally published in German, 1837); HERBERT MARCUSE, *Reason and Revolution: Hegel and the Rise of Social Theory,* 2nd ed. (1954, reprinted 1989), a new interpretation.

Marx and Engels: ROBERT C. TUCKER (ed.), *The Marx-Engels Reader,* 2nd ed. (1978), an accessible survey of the works of Marx and Engels for students of any level; JOHN PLAMENATZ, *German Marxism and Russian Communism* (1954, reprinted 1975), an illuminating study.

The 20th century. JOHN RAWLS, *A Theory of Justice* (1971, reissued 1986), presenting an alternative theory to Utilitarianism, will be of interest to students of government as well as of political philosophy. Also useful is GEORGE FRIEDMAN, *The Political Philosophy of the Frankfurt School* (1981).

Political pragmatism: ISAIAH BERLIN, *Historical Inevitability* (1954), and *Two Concepts of Liberty* (1958); ÉMILE DURKHEIM, *The Rules of Sociological Method,* 8th ed., trans. by SARAH A. SOLOVAY and JOHN H. MUELLER, ed. by GEORGE E.G. CATLIN (1938; originally published in French, 1895); VILFREDO PARETO, *The Mind and Society,* ed. by ARTHUR LIVINGSTON, trans. by ANDREW BONGIORNO and ARTHUR LIVINGSTON, 4 vol. (1935, reprinted 1983; originally published in Italian, 2 vol., 1916).

Religious and Existentialist approaches: JEAN-PAUL SARTRE, *Being and Nothingness: An Essay on Phenomenological Ontology,* trans. by HAZEL E. BARNES (1956, reissued 1992; originally published in French, 1943), and *Existentialism and Humanism,* trans. by PHILIP MAIRET (1948, reissued 1970; originally published in French, 1946); MARK BLITZ, *Heidegger's Being and Time and the Possibility of Political Philosophy* (1981).

Revolutionary doctrines: GEORGES SOREL, *Reflections on Violence,* trans. by T.E. HULME (1912, reissued 1975; originally published in French, 1908), and *The Illusions of Progress* (1969; originally published in French, 1908); ROBERT C. TUCKER (ed.), *The Lenin Anthology* (1975), with interpretive comments; GEORGE LICHTHEIM, *Marxism: An Historical and Critical Study,* 2nd ed. rev. (1964, reprinted 1982), a good short account; GEORG LUKÁCS (GYÖRGY LUKÁCS), *Goethe and His Age,* trans. by ROBERT ANCHOR (1968, reissued 1978; originally published in German, 1947); HERBERT MARCUSE, *Eros and Civilization* (1955, reissued 1987), *Soviet Marxism: A Critical Analysis* (1958, reissued 1985), and *One-Dimensional Man* (1964, reissued 1991); ALASDAIR MACINTYRE, *Marcuse* (1970), dealing with Marcuse's philosophy; ANTONIO GRAMSCI, *The Modern Prince, and Other Writings* (1957, reissued 1968); JOHN M. CAMMETT, *Antonio Gramsci and the Origins of Italian Communism* (1967); A. POZZOLINI, *Antonio Gramsci: An Introduction to His Thought* (1970; originally published in Italian, 1968). (J.E.B./Ed.)

Political Systems

The term political system may be used narrowly or broadly. Narrowly defined, it is the set of formal legal institutions that constitute a "government" or a "state." This is the definition adopted by many studies of the legal or constitutional arrangements of advanced political orders. More broadly defined, it comprehends actual as well as prescribed forms of political behaviour, not only the legal organization of the state but also the reality of how the state functions. Still more broadly defined, the political system is seen as a set of "processes of interaction" or as a subsystem of the social system interacting with other nonpolitical subsystems, such as the economic system. This points to the importance of informal sociopolitical processes and emphasizes the study of political development.

Traditional legal or constitutional analysis, using the first definition, has produced a huge body of literature on governmental structures, many of the specialized terms that are a part of the traditional vocabulary of political science, and several instructive classifying schemes. Similarly, empirical analysis of political processes and the effort to identify the underlying realities of governmental forms have yielded a rich store of data and an important body of comparative theory. The third definition has inspired much scholarly work that employs new kinds of data, new terms, and some new concepts and categories of analysis. The discussion that follows draws on all three approaches to the study of political systems.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, sections 541, 542, 543, and 544, and the *Index*.

The article is divided into the following sections:

Typologies of government 1004
  Supranational political systems 1004
    Empires
    Leagues
    Confederations
    Federations
    The United Nations organization
  National political systems 1006
    Unitary nation-states
    Federal systems
  Subnational political systems 1008
    Tribal communities
    Rural communities
    Cities
    Regions
  Issues of classification 1010
  Governments classified by mode of succession 1011
    Hereditary succession
    Succession by constitutional prescription
    Succession by election
    Succession by force
    Autocratic versus nonautocratic rule
  Governments classified by stage of development 1013
The structure of government 1013
  Contemporary forms of government 1014
    Monarchy
    Dictatorship
    Oligarchy
    Constitutional government
  Contemporary levels of government 1016
    National government
    Regional and state government
    City and local government
  Contemporary divisions of government 1018
    The legislature
    The executive
    The judiciary
The functions of government 1020
  The tasks 1020
    Self-preservation
    Supervision and resolution of conflicts
    Regulation of the economy
    Protection of political and social rights
    Provision of goods and services
  Public administration 1022
Development and change in political systems 1022
  Causes of stability and instability 1023
    Unstable political systems
    Stable political systems
  Types of political change 1024
    Radical revolution
    Structural revision
    Change of leaders
    Change of policies
  Electoral processes 1025
    Functions of elections
    Systems of counting votes
    Types of elections
    Constituencies: districting and apportionment
    Voting practices
    Participation in elections
    Influences on the direction of voting
Bibliography 1031

Typologies of government

The most important type of political system in the modern world is the nation-state. The world today is divided territorially into more than 175 states, in each of which a national government claims to exercise sovereignty—or the power of final authority—and seeks to compel obedience to its will by its citizens. This fact of the world's political organization suggests the distinction employed in the following section among supranational, national, and subnational political systems.

SUPRANATIONAL POLITICAL SYSTEMS

The formation of supranational relationships is a principal result of the division of the world into a number of separate national entities, or states, that have contact with one another, share goals or needs, and face common threats. In some cases, as in many alliances, these relationships are short-lived and fail to result in significant institutional development. In other cases, they lead to interstate organizations and supranational systems. The discussion below examines several types of supranational political systems, together with historical and contemporary examples of each.

Empires. Because they are composed of peoples of different cultures and ethnic backgrounds, all empires are ultimately held together by coercion and the threat of forcible reconquest. Imposing their rule on diverse political structures, they are characterized by the centralization of power and the absence of effective representation of their component parts. Although force is thus the primary instrument of imperial rule, it is also true that history records many cases of multiethnic empires that were governed peaceably for considerable periods and were often quite successful in maintaining order within their boundaries. The history of the ancient world is the history of great empires—Egypt, China, Persia, and imperial Rome—whose autocratic regimes provided relatively stable government for many subject peoples in immense territories over many centuries. Based on military force and religious belief, the ancient despotisms were legitimized also by their achievements in building great bureaucratic and legal structures,

Ancient and modern empires

in developing vast irrigation and road systems, and in providing the conditions for the support of high civilizations. Enhancing and transcending all other political structures in their sphere, they could claim to function as effective schemes of universal order.

In contrast to the empires of the ancient world, the colonial empires of recent times fell far short of universal status. In part, these modern European empires were made up of "colonies" in the original Greek sense; peopled by immigrants from the mother country, the colonies usually established political structures similar to those of the metropolitan centre and were often able to exercise a substantial measure of self-government. In part, also, the European empires were composed of territories inhabited by native populations and administered by imperial bureaucracies. The government of these territories was generally more coercive than in the European colonies and more concerned with protection and supervision of the commercial, industrial, and other exploitative interests of the imperial power. The disintegration of these empires occurred with astonishing speed. The two world wars of the 20th century sapped the power of the metropolitan centres, while their own doctrines of democracy, equality, and self-determination undermined the principle of imperial rule. Powers such as Britain and France found it increasingly difficult to resist claims to independence couched in terms of the representative concepts on which their home governments were based, and they lacked the military and economic strength to continue their rule over restive native populations. In the two decades after 1945, nearly all the major colonial territories won their independence; the great colonial empires that had once ruled more than half the world were finally dismembered.

Leagues. One of the commonest forms of supranational organization in history is that of leagues, generally composed of states seeking to resist some common military or economic threat by combining their forces. This was the case with the early city leagues, such as the Achaean and Actolian leagues in ancient Greece and the Hanseatic and the Swabian leagues in Europe; and to a great extent it was the case with the League of Nations. Other common features of leagues include the existence of some form of charter or agreement among the member states, an assembly of representatives of the constituent members, an executive organ for the implementation of the decisions of the assembly of representatives, and an arbitral or judicial body for adjudicating disputes.

The League of Nations

The League of Nations was one of the great experiments in supranational organization of the 20th century and the predecessor in several important respects of the United Nations. The Covenant of the League was drafted by a special commission of the Peace Conference after World War I, with Pres. Woodrow Wilson of the United States as its leading advocate, and approved by a plenary conference of the victorious powers in 1919. The initial membership of the League consisted of 20 states. The United States failed to take membership in the League, but by 1928 the organization had a total membership of 54. The machinery of the League consisted of an Assembly of all the member nations, acting through agents of their governments; a council on which the great powers were permanently represented and to which the other member powers were elected by the Assembly for three-year terms; a Secretariat to administer the internal affairs of the League; and a number of specialized agencies, such as the International Labour Organisation, that were responsible for implementing various economic and humanitarian programs on an international basis. The Covenant required that international disputes be submitted to peaceful settlement with a provision for adjudication or arbitration by the Permanent Court of International Justice or for intervention by the Council of the League. The Covenant also provided for the use of financial and economic penalties, such as embargoes, to enforce the decisions of the League and for joint military action against convicted aggressors. In practice, however, the League failed its most important tests and was unable to master the crises that led to World War II and its own collapse.

Confederations. Confederations are voluntary associations of independent states that, to secure some common purpose, agree to certain limitations on their freedom of action and establish some joint machinery of consultation or deliberation. The limitations on the freedom of action of the member states may be as trivial as an acknowledgment of their duty to consult with each other before taking some independent action or as significant as the obligation to be bound by majority decisions of the member states. Confederations usually fail to provide for an effective executive authority and lack viable central governments; their member states typically retain their separate military establishments and separate diplomatic representation; and members are generally accorded equal status with an acknowledged right of secession from the confederation. Historically, confederations have often proved to be a first or second step toward the establishment of a national state, usually as a federal union. Thus, the federal union of modern Switzerland was preceded by a confederation of the Swiss cantons; Germany's modern federal arrangements may be traced to the German Confederation of the 19th century (the Deutsche Bund); and the federal constitution of the United States is the successor to the government of the Articles of Confederation. In some other cases, confederations have replaced more centralized arrangements, as, for example, when empires disintegrate and are replaced by voluntary associations of their former colonies. The British Commonwealth, or Commonwealth of Nations, and the French Community are cases of this type.

An example of confederal arrangements that gave birth to a federal union is the Articles of Confederation (1781–89) that preceded the Constitution of the United States. The Articles established a Congress of the confederation as a unicameral assembly of ambassadors from the 13 states, each possessing a single vote. The Congress was authorized to appoint an executive committee of states

The Articles of Confederation

> to execute, in the recess of Congress, such of the powers of Congress as the United States, in Congress assembled, by the consent of nine States, shall from time to time think expedient to vest them with;

in turn, the committee of states could appoint a presiding officer or president for a term of one year. The Congress could also appoint such other committees and "civil officers as may be necessary for managing the general affairs of the United States" and was given the authority to serve as "the last resort or appeal in all disputes and differences, now subsisting or that hereafter may arise between two or more states." Although the Congress was given authority in important areas such as the regulation of foreign affairs, the establishment of coinage and weights and measures, the appointment of officers in the confederation's land and naval forces, and the issuance of bills of credit, all its powers were in fact dependent for their enforcement upon the states. The Congress lacked both an independent source of revenue and the executive machinery to enforce its will directly upon individuals. As the language of the Articles summarized the situation,

> each State retains its sovereignty, freedom and independence, and every power, jurisdiction and right, which is not by this Confederation expressly delegated to the United States in Congress assembled.

The Commonwealth of Nations

The Commonwealth of Nations is an example of a confederation born as the result of the decentralization and eventual disintegration of an empire. The original members were the United Kingdom, Australia, Canada, the Irish Free State, Newfoundland, New Zealand, and the Union of South Africa. In 1949 Newfoundland became a province of Canada, and the Irish Free State became an independent republic; South Africa became an independent republic in 1961. New commonwealth members in the latter half of the 20th century were newly independent former British colonies, such as Malaysia (1957), Cyprus (1961), and Kiribati (1979), and numbered well over 40.

The Statute of Westminster (1931) established that all members were equal in status, although all recognized the British monarch as head of the commonwealth. Commonwealth governments were represented in the capitals of other commonwealth countries by high commissioners equal in status to ambassadors. In 1965 the common-

wealth established its Secretariat to organize meetings, keep the membership informed, and implement its collective decisions. The Commonwealth Fund for Technical Cooperation, which is financed by all member states, was implemented to provide technical assistance to less-developed states. The nations of the commonwealth rarely acted in concert on the international scene, and, despite fairly regular meetings of the commonwealth prime ministers, there were at times severe strains in the relations among several of the member states. The fairly general use of the English language and of English common law, together with some common symbols and remaining cultural affinities, appeared to be the major ties binding together this loose association.

Federations. The term federation is used to refer to groupings of states, often on a regional basis, that establish central executive machinery to implement policies or to supervise joint activities. In some cases such groupings are motivated primarily by political or economic concerns; in others, military objectives are paramount. Examples of the former include the European Communities (EC), actually a combination of three main structures—the European Coal and Steel Community, established in 1952; the European Economic Community (Common Market), established in 1958; and the European Atomic Energy Community (Euratom), established in 1958. The Communities quickly developed executive machinery exercising significant regulatory and directive authority over the governments and private business firms of the member countries. Although each of the member governments retains a substantial measure of sovereignty, and a systematic effort by one or more of the governments to resist the authority of the Communities' agencies could endanger the whole fabric of cooperative effort, the Communities have developed significant supranational features. These include the staffing of executive organs with persons other than governmental representatives, the making of binding decisions on important matters by majority vote, and the capability of the Communities' agencies to deal directly and authoritatively with individuals and companies within the member states. For example, the high authority of the Coal and Steel Community acts by majority vote of its members, without instruction from any of the governments, to "assure the achievement of the purposes stated in the Treaty"; and in pursuing this function it involves itself deeply in the economies of each of the member nations.

The European Communities

The North Atlantic Treaty Organization (NATO), established in April 1949, is an example of a modern military alliance endowed with complex and permanent executive machinery, employing multilateral procedures, and involving the continuous elaboration of plans for the conduct of joint military action by its member states (Belgium, Canada, Denmark, France, Germany, Greece, Iceland, Italy, Luxembourg, The Netherlands, Norway, Portugal, Spain, Turkey, the United Kingdom, and the United States). As stated in its treaty, the purpose of NATO is to maintain the security of the North Atlantic area by exercise of the right of collective security recognized in the Charter of the United Nations. An impressive array of institutional mechanisms was established, including a secretary-general and a permanent staff, a council, a military command structure, and liaison staffs; and an ongoing system of collaboration in planning and joint military exercises was brought into being. With the continued development of its organization, NATO gradually added a number of economic and cultural activities to its functions until it came to possess several of the features of a multipurpose supranational organization.

The United Nations organization. A supranational political system that does not fit precisely any of the conventional classifications of such systems is the United Nations, a voluntary association of most of the world's nation-states. Its membership had grown from an original 51 states to more than 175 by the late 20th century. (The government of the People's Republic of China was admitted in place of the government of Taiwan in 1971.) The United Nations was founded in 1945 at a conference in San Francisco that was attended by representatives of all the nations that had declared war on Germany or Japan. The purposes of the organization are declared in its Charter to be the maintenance of international peace and security, the development of friendly relations among states, and international cooperation in solving the political, economic, social, cultural, and humanitarian problems of the world. Its organizational structure consists of a Security Council of five permanent members (China, France, Russia, the United Kingdom, and the United States) and 10 nonpermanent members elected for two-year terms, a General Assembly, a secretary-general and a Secretariat, an Economic and Social Council, a Trusteeship Council, and the International Court of Justice. Attached to the United Nations are a number of specialized agencies, including the Food and Agriculture Organization, the International Atomic Energy Agency, the International Civil Aviation Organization, the International Labour Organisation, the International Monetary Fund, the International Telecommunications Union, the Universal Postal Union, the United Nations Educational, Scientific and Cultural Organization, the World Health Organization, and the International Bank for Reconstruction and Development (World Bank).

Agencies of the United Nations

Aside from the rather generally stated and decidedly elusive aims of the Charter, the member states of the United Nations cannot be said to have any common goal, and they have often failed to unify in the face of common external threats to security. There also has been difficulty in reaching and implementing decisions. Two different formulas are employed for voting in the two principal organs, the General Assembly and the Security Council. In the General Assembly a two-thirds majority decides on important matters, but, since the Assembly's decisions are not binding and are merely recommendations, this qualified majority principle must be viewed as of little significance. Although, on the other hand, the decisions of the Security Council may be binding, a unanimous vote of all five of the permanent members joined by the votes of at least four of the nonpermanent members is required; whenever important questions of peace and security are at stake, it has rarely been possible to achieve agreement among the five great powers of the council. Although these difficulties might be fatal to the survival of many supranational organizations, they are not in fact totally debilitating for the United Nations. The United Nations continues to serve as a very important forum for international debate and negotiation, and its specialized agencies play an important role in what is sometimes referred to as "the functional approach to peace."

NATIONAL POLITICAL SYSTEMS

The term nation-state is used so commonly and yet defined so variously that it will be necessary to indicate its usage in this article with some precision and to give historical and contemporary examples of nation-states. To begin with, there is no single basis upon which such systems are established. Many states were formed at a point in time when a people sharing a common history, culture, and language discovered a sense of identity. This was true in the cases of England and France, for example, which were the first nation-states to emerge in the modern period, and of Italy and Germany, which were established as nation-states in the 19th century. In contrast, however, other states, such as India, the Soviet Union, and Switzerland, came into existence without a common basis in race, culture, or language. It must also be emphasized that contemporary nation-states are creations of different historical periods and of varied circumstances. Before the close of the 19th century, the effective mobilization of governmental powers on a national basis had occurred only in Europe, the United States, and Japan. It was not until the 20th century and the collapse of the Ottoman, Habsburg, French, and British empires that the world could be fully organized on a national basis. This transformation was completed in 1991 with the dissolution of the U.S.S.R. In 1920 the League of Nations recognized seven nation-states as "Great Powers" (Britain, France, the United States, Germany, Italy, Japan, and Russia) and eventually admitted more than 40 other states to membership; the United Nations had more than 175 member

The nation-state defined

states in the late 20th century. States in the post-Cold War world include the United States as the preeminent power; the established powers of Britain, France, China, Japan, Germany, and Russia; emerging powers such as Ukraine and Brazil; and a host of old and new states such as Denmark, Namibia, Kazakhstan, Switzerland, Egypt, Turkey, Malaysia, and Chile.

The characteristics that qualify these variously composed and historically differing entities as nation-states and distinguish them from other forms of social and political organization amount in sum to the independent power to compel obedience from the populations within their territories. The state is, in other words, a territorial association that may range in size from Russia to Singapore, in population from China to Luxembourg, and that claims supremacy over all other associations within its boundaries. As an association, the state is peculiar in several respects: membership is compulsory for its citizens; it claims a monopoly of the use of armed force within its borders; and its officers, who are the government of the state, claim the right to act in the name of the land and its people.

A definition of the state in terms only of its powers over its members is not wholly satisfactory, however. Although all states make a claim to supremacy within their boundaries, they differ widely in their ability to make good their claims. States are, in fact, often challenged by competing associations within their boundaries; their supremacy is often more formal than real; and they are sometimes unable to maintain their existence. Moreover, a definition in terms of power alone ignores the fact that there are great differences among states in the structures they employ for the exercise of power, in the ways they use power, and in the ends to which they turn their power. Some of these differences are explored in the discussion that follows of two general categories of nation-states: the unitary state and the federal state. Partly from administrative necessity and partly because of the pressures of territorial interests, nearly all modern states provide for some distribution of governmental authority on a territorial basis. Systems in which power is delegated from the central government to subnational units and in which the grant of power may be rescinded at the will of the central government are termed unitary systems. Systems in which a balance is established between two autonomous sets of governments, one national and the other provincial, are termed federal. In federal systems, the provincial units are usually empowered to grant and take away the authority of their own subunits in the same manner as national governments in unitary systems. Thus, although the United States is federally organized at the national level, each of the 50 states is in a unitary relationship to the cities and local governments within its own territory.

Unitary nation-states. A great majority of all the world's nation-states are unitary systems, including Belgium, Bulgaria, France, Great Britain, The Netherlands, Japan, Poland, Romania, the Scandinavian countries, Spain, and many of the Latin-American and African countries. There are great differences among these unitary states, however, specifically in the institutions and procedures through which their central governments interact with their territorial subunits.

Decentralization of power in unitary states

In one type of unitary system, decentralization of power among subnational governments goes so far that in practice, although not in constitutional principle, they resemble federal arrangements. In Great Britain, for example, there are important elements of regional autonomy in the relationship between Northern Ireland, Wales, and Scotland and the national government in London; and the complex system of elected local governments, although in constitutional theory subject to abrogation by Parliament, is in practice a fixed and fairly formidable part of the apparatus of British government. In other unitary systems of this type, decentralization on a territorial basis is actually provided for constitutionally, and the powers of locally elected officials are prescribed in detail. Thus, the Japanese constitution, for example, specifies certain autonomous functions to be performed by local administrative authorities.

A second type of unitary system makes substantially less provision for territorial decentralization of authority and employs rather strict procedures for the central supervision of locally elected governments. The classic example of this type is France. Until March 1982, when a law on decentralization went into effect, the French administrative system was built around *départements,* each headed by a *préfet,* and subdivisions of the *départements,* termed *arrondissements,* each headed by a *sous-préfet.* The *préfets* and *sous-préfets* were appointed by the government in Paris to serve as agents of the central government and also as the executives of the divisional governments, the *conseils généraux,* which were composed of elected officials. The system thus combined central supervision of local affairs through appointed officials with territorial representation through locally elected governments. (Following the passage of the decentralization law, the executive powers of the *préfets* were transferred to the *conseils généraux.*)

Yet a third type of unitary system provides for only token decentralization. In such cases, the officials responsible for managing the affairs of the territorial subdivisions are appointees of the central government, and the role of locally elected officers is either minimal or nonexistent. Examples of this kind of arrangement include Germany under Adolf Hitler and also several formerly Communist countries. The Third Reich was divided into 42 *Gaue,* each headed by a gauleiter chosen for his personal loyalty to Hitler. In eastern Europe, the people's councils or people's committees were named by the centrally organized Communist parties; their appointment was confirmed by elections with one slate of candidates.

Federal systems. In federal systems, political authority is divided between two autonomous sets of governments, one national and the other subnational, both of which operate directly upon the people. Usually a constitutional division of power is established between the national government, which exercises authority over the whole national territory, and provincial governments that exercise independent authority within their own territories. Of the eight largest countries in the world, seven—Argentina, Australia, Brazil, Canada, India, Russia, and the United States—are organized on a federal basis. Federal states also include Austria, Germany, Malaysia, Mexico, Nigeria, Switzerland, and Venezuela.

The governmental structures and political processes found in these federal systems show great variety. One may distinguish, first, a number of systems in which federal arrangements reflect rather clear-cut cultural divisions. A classic case of this type is Switzerland, where the people speak four different languages—German, French, Italian, and Romansh—and the federal system unites 26 historically and culturally different entities, known as cantons and demicantons. The Swiss constitution of 1848, as modified in 1874, converted into the modern federal state a confederation originally formed in the 13th century by the three forest cantons of Uri, Schwyz, and Unterwalden. The principal agencies of federal government are a bicameral legislature, composed of a National Council representing the people directly and a Council of States representing the constituent members as entities; an executive branch (Bundesrat) elected by both houses of the legislature in joint session; and a supreme court that renders decisions on matters affecting cantonal and federal relations. The Russian Federation's arrangements, although of a markedly different kind, also reflect the cultural and linguistic diversity of the country. Depending on their size and on the territories they have historically occupied, ethnic minorities may have their own autonomous republic, province, or district. These divisions provide varying degrees of autonomy in setting local policies and provide a basis for the preservation of the minorities' cultures. Some of these areas were integrated into the Russian Empire centuries ago, after the lands were taken from the Mongols of the Golden Horde, and others resisted occupation even late in the 19th century. It is not uncommon for Russians to constitute a plurality of the population in these areas. The national government consists of the executive branch, led by the nationally elected president; the parliament; and a judicial branch that resolves constitutional matters.

The Swiss and Russian systems

In other systems, federal arrangements are found in

conjunction with a large measure of cultural homogeneity. The Constitution of the United States delegates to the federal government certain activities that concern the whole people, such as the conduct of foreign relations and war and the regulation of interstate commerce and foreign trade; certain other functions are shared between the federal government and the states; and the remainder are reserved for the states. Although these arrangements require two separate bodies of political officers, two judicial systems, and two systems of taxation, they also allow extensive interaction between the federal government and the states. Thus, the election of Congress and the president, the process of amending the Constitution, the levying of taxes, and innumerable other functions necessitate cooperation between the two levels of government and bring them into a tightly interlocking relationship.

SUBNATIONAL POLITICAL SYSTEMS

Basic political communities

Although national government is the dominant form of political organization in the modern world, an extraordinary range of political forms exists below the national level—tribal communities, the intimate political associations of villages and towns, the governments of regions and provinces, the complex array of urban and suburban governments, and the great political and administrative systems of the cities and the metropolises. These subnational entities are, in a sense, the basic political communities—the foundation on which all national political systems are built.

Tribal communities. The typical organization of mankind in its early history was the tribe. Today, in many parts of the world, the tribal community is still a major form of human political organization. Even in advanced systems, traces can still be found of its influence. Some of the *Länder* of modern Germany, such as Bavaria, Saxony, or Westphalia, have maintained their identity since the days of the Germanic tribal settlements. In England, too, many county boundaries can be explained only by reference to the territorial divisions in the period after the end of the Roman occupation.

In many African states, the tribe is still an effective community and a vehicle of political consciousness. Most of these states are the successors to the administrative units established by colonial regimes and owe their present boundaries to the often arbitrary decisions of imperial bureaucracies or to the territorial accommodations of rival colonial powers. The result was often the splintering of the tribal communities or their aggregation in largely artificial entities. Tribal loyalties continue to hamper nation-building efforts in some parts of the world where tribes were once the dominant political structure. Tribes may act through formal political parties like any other interest group. In some cases they simply act out their tribal bias through the machinery of the political system, and in others they function largely outside of formal political structures.

Importance of kinship in tribal societies

In its primary sense, the tribe is a community organized in terms of kinship, and its subdivisions are the intimate kindred groupings of moieties, gentes, and totem groups. Its territorial basis is rarely defined with any precision, and its institutions are typically the undifferentiated and intermittent structures of an omnifunctional social system. The leadership of the tribe is provided by the group of adult males, the lineage elders acting as tribal chiefs, the village headmen, or the shamans, or tribal magicians. These groups and individuals are the guardians of the tribal customs and of an oral tradition of law. Law is thus not made but rather invoked; its repository is the collective memory of the tribal council or chief men. This kind of customary law, sanctioned and hallowed by religious belief, nevertheless changes and develops, for each time it is declared something may be added or omitted to meet the needs of the occasion.

Rural communities. The village has traditionally been contrasted with the city: the village is the home of rural occupations and tied to the cycles of agricultural life, while the inhabitants of the city practice many trades, and its economy is founded on commerce and industry; the village is an intimate association of families, while the city is the locus of a mass population; the culture of the village is simple and traditional, while the city is the centre of the arts and sciences and of a complex cultural development. The village and the city offer even sharper contrasts as political communities. Historically, the village has been ruled by the primitive democracy of face-to-face discussion in the village council or by a headman whose decisions are supported by village elders or by other cooperative modes of government; urban government has never been such a simple matter, and monarchical, tyrannical, aristocratic, and oligarchic forms of rule have all flourished in the city. In the village, the boundaries among political, economic, religious, and other forms of action have not been as clearly drawn as in cities.

The origins and development of the apparatus of government can be seen most clearly in the simple political society of the rural community. The transformation of kinbound societies with their informal, folk-sustained systems of sociopolitical organization into differentiated, hierarchical societies with complex political structures began with the enlargement of the rural community—an increase in its population, the diversification of its economy, or its interaction with other communities. The rudimentary organs of communal government were then elaborated, the communal functions received more specialized direction, and leadership roles were institutionalized. This was sometimes a process that led by gradual stages to the growth of cities. Elsewhere, however, as in the case of ancient Attica, the city was established as the result of a process of *synoikismos*, or the uniting of a number of tribal or village communities. This was undoubtedly the origin of Athens, and, according to its legendary history, Rome also was established as a result of the forcible unification of the tribes that dwelt on the hills surrounding the Palatine Hill.

Even in the nation-states of today's world, the contrasts between the village or the town and the city as centres of human activity are readily apparent. In the country, life is more intimate, human contacts more informal, the structure of society more stable. In the city, the individual becomes anonymous, the contacts between people are mainly formal, and the standing of the individual or the family in society is subject to rapid change. In many contemporary systems, however, the differences in the forms of government of rural and urban communities appear to be growing less pronounced. In the United States, for example, rural institutions have been seriously weakened by the movement of large numbers of people to the city. The township meeting of New England and other forms of direct citizen participation in the affairs of the community have declined in importance and have often been displaced by more formal structures and the growth of local governmental bureaucracies.

Organization of ancient cities

Cities. Cities first emerged as complex forms of social and political organization in the valleys of the Euphrates and the Tigris, the Nile, the Huang Ho, and the Yangtze. These early cities broke dramatically with the patterns of primitive life and the rural societies from which they sprang. Kinship as the basis of society was replaced by status determined by class and occupation; the primitive magical leaders of the tribe were displaced by temple priesthoods presiding over highly developed religious institutions and functioning as important agencies of social control; earlier systems of rule by the tribal chieftains and the simple forms of communal leadership gave way to kingships endowed with magical powers and important religious functions; and specialized functionaries in the royal courts became responsible for supervising new kinds of governmental activity. Many other developments contributed to the growing centralization of power in these city civilizations. Barter was replaced by more effective systems of exchange, and the wealth generated in commerce and the specialized city trades became both an object of taxation and an instrument of power. Class distinctions emerged as the result of a division of labour and advances in technical development. A military order and a professional soldiery were created and trained in new techniques of warfare, and a slave class provided the work force for large-scale projects of irrigation, fortification, and royal architecture. As these developments proceeded, the

city was able to project its power even further into the surrounding countryside, to establish its rule over villages and other cities in its sphere, and finally to become the centre of such early empires as those of Sumeria, Egypt, China, Babylonia, Assyria, and Persia.

The Greek polis

A very different form of city life emerged among the Greeks. The Greek polis also broke with the folkways of primitive society, but its political development was in striking contrast to the despotism of the Oriental city empires and their massive concentrations of power in the hands of king and priest. As the polis transcended its origins in village life, the powers of the tribal chief dwindled and passed into the hands of aristocratic families. The kingship of Homeric tradition vanished, the "kings" who remained became mere dignitaries in the religious and ceremonial life of the city, and new magistracies and other civic offices were founded. These offices became the focus of factional struggle among the aristocratic families and later, with the weakening of aristocratic rule, the chief prizes in a contest of power between the nobility and the common citizens. Eventually, these developments issued in the characteristic form of Greek city government. A citizen body, always a much narrower group than the total population but often as numerous as the population of freeborn males, acquired power in the direction of the city government through the election of its officers and direct participation in the city councils. Although often interrupted by episodes of oligarchic or tyrannical rule and by periods of civic dissension and class rivalry, the main theme of governmental development in the Greek city was the elaboration of structures that permitted the control of political affairs by its citizens.

Autonomous cities also sprang up in Europe in the later Middle Ages. Medieval city life, although it differed from that of the polis and was coloured by the forms of feudal society, also emphasized the principle of cooperative association. Indeed, for the first time in the history of city civilization, the majority of the inhabitants of the city were free. The development of trades, the growth of commerce, and the mobilization of wealth emancipated the city from its feudal environment, and the merchant and craft guilds became the matrices of a new kind of city democracy. In time, the guilds were transformed into closed corporations and became a basis for oligarchic control; and the city's independence was threatened by the rise of the new nation-states. Tempting targets for the ambition of kings, Venice, Genoa, Florence, Milan, Cologne, Amsterdam, Hamburg, and other free cities of Europe eventually succumbed to monarchical control. Theirs was an important legacy, however, for the political order of the medieval city was a powerful influence in the development of the constitutional structures of the modern democratic state.

Modern cities

Although cities are no longer independent, the almost universal increase in urban population has made them more important than ever before as centres of human activity. The political organization of modern cities differs from country to country. Even within the same nation-state, there are often important contrasts in the structures of city government. In the United States, for example, three principal types of city government are usually distinguished: the council–manager form, the mayor–council form, and the commission form.

More than half of all American cities with populations between 25,000 and 250,000 operate under council–manager governments. In council–manager systems the council is generally small, elected at large on a nonpartisan ballot for overlapping four-year terms; no other offices are directly elected, and the mayor, who presides at council meetings and performs mainly ceremonial functions, is chosen by the council from among its members. The manager, a professional city administrator, is selected by the council, serves at the council's pleasure, and is responsible for supervising the city departments and municipal programs, preparing the budget, and controlling expenditures.

Mayor–council governments are found in two basic forms, the "weak" mayor and the "strong" mayor. The former was typical of the 19th-century municipal organization and is now mainly confined to cities of less than 25,000 population; the latter is a common arrangement in cities with significantly larger populations. In weak-mayor–council governments, a number of officials, elected or appointed for lengthy terms, wield important administrative powers; the council, typically elected by divisions of the city called wards, is responsible for the direction of the major city agencies; and the mayor's powers of appointment and removal and his control over the city budget are severely limited. In many cases, strong-mayor–council governments evolved from weak-mayor–council systems as an independently elected mayor won the power of veto over council ordinances, strengthened his control over appointment and removal, and established himself as the city's chief budgetary officer; at the same time, also, the elective administrative officers and the semi-autonomous appointive boards and commissions were often eliminated and the number of councilmen reduced.

The commission plan, which is now found in fewer than 200 cities of more than 10,000 population, concentrates legislative and executive powers in the hands of a small group of commissioners. The commissioners serve individually as the heads of administrative departments and choose one of their number to act as a ceremonial mayor without executive authority.

The variety in the governmental structures of American cities is paralleled in many other countries, for everywhere in the modern world the government of the city continues to challenge man's political invention. Although no longer sovereign, cities are the centres of modern civilization and—both in terms of the services demanded of them and the range and importance of the functions they exercise—the most important of contemporary subnational political systems. Moreover, it is in the cities that most of the problems of modern industrial society seem to have their focus. These problems are not only governmental but also technological, cultural, and economic. They are found in their most acute form in the great metropolitan centres and in that vast urban agglomeration known as the megalopolis. In political terms, the issue that is posed appears to be whether these huge centres of population can continue as effective communities with democratically manageable governments.

Regions. In many contemporary national political systems the forces of history and administrative necessity have joined to produce regional communities at an intermediate level between the local and the national community. In some cases—the Swiss canton, the English county, the German *Land,* and the American state—these regional communities possess their own political institutions and exercise governmental functions. In other cases, however, the territorial community is a product of ethnic, cultural, linguistic, physiographic, or economic factors and maintains its identity without the support of political structures.

Regional communities

As subnational political systems, regional communities are sometimes based in tradition, even tracing their origin to a period prior to the founding of the nation; in other cases, they are modern administrative units created by national governments for their own purposes. Examples of both types may be found in the history of regionalism in France and its complex pattern of internal territorial divisions. Before the French Revolution, France was divided into ancient provinces—Burgundy, Gascony, Brittany, Normandy, Provence, Anjou, Poitou, and others. After the Revolution, in what seems to have been an effort to discourage regional patriotism and threats of separatism, the Napoleonic government superimposed a new regional structure of *départements* on the old provincial map. More than a century and a half later, in the era of rapid communications and national economic planning, the French national government announced a regrouping of the Napoleonic *départements* into much larger Gaullist *régions.* Recognizing, perhaps, the continuing strength of the provincial attachments of Gascon, Breton, Norman, and Provençal and the survival of old regional folk cultures with their distinctive patterns of speech, the new *régions* were given boundaries similar in many cases to the traditional provincial boundaries of pre-republican France.

The history of the French regional communities is not a special case, for political, administrative, economic, and technical forces have led many other national governments

to replace traditional territorial divisions with new regional units. In England, for example, the traditional structure of county governments (52 in number) was replaced by a system of administrative counties (61 plus London); and, in 1969, a royal commission proposed a further reform that would abolish 39 counties. Attempts have also been made to use older regional communities as the infrastructure for new systems of regional government. Thus, the Italian constitution provides for semi-autonomous government in five special regions—Valle d'Aosta, Sardinia, Sicily, Trentino-Alto Adige, and Friuli-Venezia Giulia—which, in different ways, are historically distinct from the rest of Italy. In yet other cases the fear of competition from regional governments or of separatist movements has led national governments to make various efforts to resist the development of regional political structures. Again, Italy provides a convenient example, for its constitution requires the establishment of 14 other autonomous regions, but Italian governments refused to implement this provision of the 1947 constitution until recently. It should be noted that the Italian republic of 1870–1922 and its Fascist successor state also made similar efforts to combat regional political development, the former by the creation of a large number of administrative provinces and the latter by establishing *corporazione* to represent occupations regardless of geographic location.

In several modern states the growth of vast conurbations and the rise of the megalopolis have prompted the development of new kinds of regional governmental structures. The Port of London Authority, the Port of New York Authority, and the San Francisco Bay Area Transit Authority are examples of regional systems designed to serve the needs of urban communities that have outgrown the boundaries of existing city governments. New regional structures have also resulted from the increased responsibility of national governments for the administration of comprehensive social and economic programs. The Tennessee Valley Authority, for example, is both a national agency and a regional government whose decisions affect the lives of the inhabitants of all the states and cities in its sphere. Other examples of new regional administrative structures include the zonal councils established in India for social and economic planning purposes, the districts of the Interstate Commerce Commission in the United States, and the governmental and economic units established in Britain to deal with the problems of industrially depressed areas.

ISSUES OF CLASSIFICATION

The almost infinite range of political systems has been barely suggested in this brief review. Confronted by the vast array of political forms, political scientists have attempted to classify and categorize, to develop typologies and models, or in some other way to bring analytic order to the bewildering variety of data. Many different schemes have been developed. There is, for example, the classical distinction between governments in terms of the number of rulers—government by one man (monarchy or tyranny), government by the few (aristocracy or oligarchy), and government by the many (democracy). There are schemes classifying governments in terms of their key institutions (for example, parliamentarism, cabinet government, presidentialism). There are classifications that group systems according to basic principles of political authority or the forms of legitimacy (charismatic, traditional, rational–legal, and others). Other schemes distinguish between different kinds of economic organization in the system (the laissez-faire state, the corporate state, and Socialist and Communist forms of state economic organization) or between the rule of different economic classes (feudal, bourgeois, and capitalist). And there are modern efforts to compare the functions of political systems (capabilities, conversion functions, and system maintenance and adaptation functions) and to classify them in terms of structure, function, and political culture.

Approaches to the analysis of political systems

Although none is comprehensive, each of these principles of analysis has some validity, and the classifying schemes that are based on them, although in some cases no longer relevant to modern forms of political organization, have often been a major influence on the course of political development. The most influential of such classifying schemes is undoubtedly the attempt of Plato and Aristotle to define the basic forms of government in terms of the number of power holders and their use or abuse of power. Plato held that there was a natural succession of the forms of government: an aristocracy (the ideal form of government by the few) that abuses its power develops into a timocracy (in which the rule of the best men, who value wisdom as the highest political good, is succeeded by the rule of men who are primarily concerned with honour and martial virtue), which through greed develops into an oligarchy (the perverted form of government by the few), which in turn is succeeded by a democracy (rule by the many); through excess, the democracy becomes an anarchy (a lawless government), to which a tyrant is inevitably the successor. Abuse of power in the Platonic typology is defined by the rulers' neglect or rejection of the prevailing law or custom (*nomos*); the ideal forms are thus *nomos* observing (*ennomon*), and the perverted forms are *nomos* neglecting (*paranomon*). Although disputing the character of this implacable succession of the forms of government, Aristotle also based his classification on the number of rulers and distinguished between good and bad forms of government. In his typology it was the rulers' concern for the common good that distinguished the ideal from perverted forms of government. The ideal forms in the Aristotelian scheme are monarchy, aristocracy, and polity (a term conveying some of the meaning of the modern concept of "constitutional democracy"); when perverted by the selfish abuse of power, they are transformed respectively into tyranny, oligarchy, and ochlocracy (or the mob rule of lawless democracy). The concept of the polity, a "mixed" or blended constitutional order, fascinated political theorists for another millennium. To achieve its advantages, innumerable writers from Polybius to St. Thomas Aquinas experimented with the construction of models giving to each social class the control of appropriate institutions of government.

Another very influential classifying scheme was the distinction between monarchies and republics. In the writings of Machiavelli and others, the tripartism of classical typologies was replaced by the dichotomy of princely and republican rule. Sovereignty in the monarchy or the principality is in the hands of a single ruler; in republics, sovereignty is vested in a plurality or collectivity of power holders. Reducing aristocracy and democracy to the single category of republican rule, Machiavelli also laid the basis in his analysis of the exercise of princely power for a further distinction between despotic and nondespotic forms of government. In the work of Montesquieu, for example, despotism, or the lawless exercise of power by the single ruler, is contrasted with the constitutional forms of government of the monarchy and the republic. As a result of the decline of monarchies and the rise of new totalitarian states terming themselves republics, this traditional classification is now, of course, of little more than historical interest.

The usefulness of all the traditional classifications has been undermined by the momentous changes in the political organization of the modern world. Typologies based on the number of power holders or the formal structures of the state are rendered almost meaningless by the standardization of "democratic" forms, the deceptive similarities in the constitutional claims and governmental institutions of regimes that actually differ markedly in their political practices, and the rise of new political orders in the non-Western world. A number of modern writers have attempted to overcome this difficulty by constructing classifying schemes that give primary importance to social, cultural, economic, or psychological factors. The most influential of such schemes is the Marxist typology, which classifies types of rule on the basis of economic class divisions and defines the ruling class as that which controls the means of production in the state. A monistic typology that also emphasized the importance of a ruling class was developed by an Italian theorist of the early 20th century, Gaetano Mosca. In Mosca's writings all forms of government appear as mere facades for oligarchy

Modern classifying systems

or the rule of a political "elite" that centres power in its own hands. Another classification, which distinguishes between "legitimate" and "revolutionary" governments, was suggested by Mosca's contemporary Guglielmo Ferrero. Using a sociopsychological approach to the relations between rulers and ruled, Ferrero held that a legitimate government is one whose citizens voluntarily accept its rule and freely give it their loyalty; in revolutionary systems, the government fears the people and is feared by them. Legitimacy and leadership are also the basis of a typology developed by the German sociologist Max Weber. In Weber's scheme there are three basic types of rule: charismatic, in which the authority or legitimacy of the ruler rests upon some genuine sense of calling and in which the followers submit because of their faith or conviction in the ruler's exemplary character; traditional, in which, as in hereditary monarchy, leadership authority is historically or traditionally accepted; and rational–legal, in which leadership authority is the outgrowth of a legal order that has been effectively rationalized and where there is a prevailing belief in the legality of normative rules or commands. The Weberian typology has been elaborated by a number of recent writers who have found it particularly useful for comparing and classifying the emergent political orders of the non-Western world.

A serviceable classification of political systems must penetrate beneath formal appearances to underlying realities; these realities, however, do not consist only of the facts of social and economic organization. Important differences often exist between political systems having very similar socioeconomic structures. That is why some recent sociological classifications and schemes of analysis fail as tools of political inquiry: they cannot effectively distinguish between certain societies whose political orders are full of contrasts. The political system itself must be the primary focus of inquiry and the phenomena of politics the principal facts of investigation. Such an approach may involve many different kinds of analysis, but it must begin with an examination of the ways in which power is acquired and transferred, exercised, and controlled. This is important for comparing advanced political orders and also for drawing important distinctions between regimes in the underdeveloped areas of the world.

GOVERNMENTS CLASSIFIED BY MODE OF SUCCESSION

A key problem of all political orders is that of succession. "The king is dead; long live the king" was the answer, not always uncontested, of European hereditary monarchy to the question of who should rule after the death of the king. A second, closely related problem is in what manner and by whom a present ruler may be replaced or deprived of power. To this second question hereditary monarchy gave no definite answer, although the concept of *diffidatio,* or the severance of the bond of allegiance between king and feudal lord, was invoked more than once in the medieval period. Political systems, even those of primitive tribal societies, have approached both problems in a variety of ways. Anthropological records show that tribal chiefs or kings were sometimes selected as a result of ritual tests or the display of magical signs and proofs of divine origin, usually as determined by the tribal elders or magical leaders; in other cases, a principle of heredity, often diluted by a choice among heirs in terms of physique or warrior ability, was applied; in still other cases, the chief was elected, often from among the adult males of a select group of families. Techniques for the removal of tribal rulers were equally varied. Sometimes the ruler would be killed after a specified period or when his magical powers weakened or when his physical prowess or health failed; in other cases the chief was exposed to periodic tests of his magical powers or required to accept challenges to combat from other qualified candidates for rule; and in some cases the elders could remove him from office.

Techniques for assuring the succession are also varied in the modern world. Succession procedures range from the complex hieratic process of identifying a reincarnated Dalai Lama, which was practiced until quite recently in Tibet, to the subtle, informal procedures by which parliamentary majorities choose a successor to the office of prime minister in Britain. In fact, however, the succession practices of modern political systems appear to be of four main types: (1) heredity, (2) constitutional prescription, (3) election, and (4) force.

Hereditary succession. Although dictators still occasionally seek to establish their sons as their heirs, they usually rely on force rather than the claims of heredity to achieve their object. Apart from a few states, mostly in the Arab world, where the dynastic ruler is the effective head of the government, the hereditary principle of succession is now almost exclusively confined to the constitutional monarchies of western Europe. There is some irony in the fact that the line of succession is more securely established in these monarchies now than at any point in their earlier history: intradynastic struggle, it appears, is much less likely when kingship is mainly ceremonial. Heredity may be reinforced or modified by constitutional prescription: this was the case, for example, of the famous Act of Settlement that secured the Hanoverian succession in Britain.

Succession by constitutional prescription. A leading example of succession by constitutional prescription is the United States. Article II, Section 1, of the Constitution of the U.S. provides:

> In case of the removal of the President from office, or of his death, resignation, or inability to discharge the powers and duties of the said office, the same shall devolve on the Vice President, and the Congress may by law provide for the case of removal, death, resignation, or inability, both of the President and Vice President, declaring what officer shall then act as President, and such officer shall act accordingly, until the disability be removed, or a President shall be elected.

A constitutional amendment was ratified in 1967 elaborating these procedures to include further arrangements for dealing with the problem of presidential disability. The original language of the Constitution has been the basis for the peaceful succession of Vice Presidents John Tyler, Millard Fillmore, Andrew Johnson, Chester Arthur, Theodore Roosevelt, Calvin Coolidge, Harry S. Truman, and Lyndon Johnson. Constitutionally prescribed arrangements for assuring the succession are not always so successful, and many states whose constitutions contain very similar provisions have experienced succession crises that were resolved only by violence.

Succession by election. Election is a principle of succession also frequently combined with force. In cases of closely contested elections or where there is doubt as to the validity or proper form of the election, the result is often a disputed succession. The Great Schism in the papacy in the 14th century and the disputed succession to the elective kingship of Hungary in the 16th century are examples of the failure of elective systems to assure an orderly succession. Force is the effective basis of succession in several contemporary states in which *pro forma* electoral confirmation is given to a ruler who seizes power.

The problem of succession imposes great strains on any political order: the continuity of rule is broken, established patterns of action are interrupted, and the future suddenly becomes uncertain. This political crisis tests the character of regimes in ways that are of some importance for comparative political analysis. A number of interesting comparisons may be drawn from the study of succession practices, but perhaps the most important is the distinction between those systems in which the problem is resolved primarily by force and those systems in which heredity, constitutional prescription, or election assure a peaceful and orderly succession.

The transfer of power

Political orders are subjected to another kind of strain when the rule of their present power holders is challenged and the question arises of depriving them of authority. This is the problem of the transfer of power: whether, in what way, and by whom a present ruler may be displaced. Like succession, it is a recurrent problem in all political systems, and, as in the case of succession practices, the ways in which political systems respond to the strains involved offer important clues to their character. It is, in a sense, the fundamental political crisis, for all systems are in some way shaken, often violently, sometimes to the point of destruction, by the struggles between established rulers and their rivals.

Succession by force. Revolutions, which are the result of the crisis in its most extreme form, involve the overthrow not merely of the government but of the political order itself. Typically, a revolution is preceded by a series of strains within the system: challenges to the authority of the government mount, and its legitimacy is increasingly questioned; the exercise of power becomes coercive, and the challenge to rule assumes ever more violent forms; eventually, the struggle comes to a dramatic climax in the destruction of the old order. The coup d'etat is another form of violent response to the crisis of rule, but it is distinguished from the revolution in that it involves the overthrow only of the government: the political order is not immediately affected, for the coup is managed by an individual or group within the government or within the ruling class. In some cases, however, the coup d'etat is merely a preliminary stage to revolution. Sometimes this happens when the new ruler leads a governmentally imposed revolution: this was the role played by Napoleon I, Napoleon III, Mussolini, and Hitler. At other times, coups are actually prompted by fear of revolution but succeed only in further weakening the claims to legitimacy of the existing order: this has recently been the case in some countries in the non-Western world where conservative-led coups were quickly overthrown by revolutionary movements. In addition to revolutions and coups d'etat, the crisis of rule may prompt other forms of violent political reaction, including civil war and secession, resistance movements and rebellions, guerrilla warfare and terrorism, class warfare, and peasant revolts.

Causes of the overthrow of governments

The causes of internal conflict leading to the forcible overthrow of governments are extremely varied. They include tensions created by rapid social and economic development; the rise of new social classes and the refusal of established elites to share their power; problems of the distribution of wealth and the grievances of different economic groups and interests; the rise of corrosive social and political philosophies and the estrangement of intellectuals; conflict of opinions over the ends of government; factional struggles among power holders or within the ruling class; the rise of a charismatic leader; oppressive rule that alienates powerful groups; weak rule that tolerates antigovernmental or revolutionary movements; and many different combinations of these and other social, economic, and political factors. All political systems experience some of these conditions with some frequency. Yet there are a number of modern states that have avoided internal wars and the forcible overthrow of their governments for considerable periods.

It appears that rulers in the contemporary world are generally safe from violent challenges if they possess an effective monopoly of military, economic, and political power, linked with certain important social controls; or, alternatively, if they are obliged to exercise limited powers for specified periods and are required to yield office to rivals who meet certain qualifications. The first is the definition of a modern totalitarian regime, fully and efficiently organized; the second describes the governments of several contemporary constitutional democracies. In the first case, the government secures itself by force combined with social and psychological means of preventing the formation of opposition. In the second case, alternatives to internal war are provided by the opportunities for oppositions to influence the exercise of power and ultimately to replace the government. The great achievement of constitutional democracy has been to give reasonable security to governments from forcible overthrow by compelling them to accept limitations on their power, by requiring them to forgo the use of force against rivals who agree to accept the same limitations, and by establishing well-known legal procedures through which these rivals may themselves constitute the government.

Autocratic versus nonautocratic rule. The foregoing discussion has suggested a distinction among political systems in terms of the role played by force in the acquisition and transfer of power. The role of force is vital, also, in distinguishing among political systems in terms of the exercise and control of power. Here the contrast is essentially between "autocratic" and "nonautocratic" governments, for totalitarianism is only a recent species of autocracy, to which constitutionalism is the principal contemporary antithesis. Autocracy is characterized by the concentration of power in a single centre, be it an individual dictator or a group of power holders such as a committee or a party leadership. This centre relies on force to suppress opposition and to limit social developments that might eventuate in opposition. The power of the centre is not subject to effective controls or limited by genuine sanctions: it is absolute power. In contrast, nonautocratic government is characterized by the existence of several centres, each of which shares in the exercise of power. Nonautocratic rule allows the development of social forces that generate a variety of interests and opinions. It also subjects the power holders to reciprocal controls and to effective sanctions of law.

Autocracy

In appearance, autocracy may sometimes be difficult to distinguish from nonautocratic rule. Often, autocracies attempt to borrow legitimacy by adopting the language of the constitutions of nonautocratic regimes or by establishing similar institutions. It is a common practice, for example, in many modern totalitarian states to establish institutions—parliaments or assemblies, elections and parties, courts and legal codes—that differ little in appearance from the institutional structures of constitutional democracies. Similarly, the language of totalitarian constitutions is often couched in terms of the doctrines of popular rule or democracy. The difference is that in totalitarian regimes neither the institutions nor the constitutional provisions act as effective checks on the power of the single centre: they are essentially facades for the exercise of power through hierarchical procedures that subject all the officials of the state to the commands of the ruling individual or group. The underlying realities of autocratic rule are always the concentration of power in a single centre and the mobilization of force to prevent the emergence of opposition.

Totalitarianism, as already noted, has been a chief form of autocratic rule; it is distinguished from previous forms in its use of state power to impose an official ideology on its citizens. Nonconformity of opinion is treated as the equivalent of resistance or opposition to the government, and a formidable apparatus of compulsion, including various kinds of state police or secret police, is kept in being to enforce the orthodoxy of the proclaimed doctrines of the state. A single party, centrally directed and composed exclusively of loyal supporters of the regime, is the other typical feature of totalitarianism. The party is at once an instrument of social control, a vehicle for ideological indoctrination, and the body from which the ruling group recruits its members.

Constitutional democracy

In the modern world, constitutional democracy is the chief type of nonautocratic government. The minimal definition in institutional terms of a constitutional democracy is that it should provide for a regularized system of periodic elections with a free choice of candidates, the opportunity to organize competing political parties, adult suffrage, decisions by majority vote with protection of minority rights, an independent judiciary, constitutional safeguards for basic civil liberties, and the opportunity to change any aspect of the governmental system through agreed procedures. Two features of constitutional democracy require emphasis in contrasting it with modern totalitarian government: the constitution, or basic law, and the political party. A constitution, as the example of British constitutional democracy suggests, need not be a single written instrument; indeed, the essence of a constitution is that it formalizes a set of fundamental norms governing the political community and determining the relations between the rulers and the people and the interaction among the centres of power. In most modern constitutional democracies, however, there is a constitutional document providing for fixed limitations on the exercise of power. These provisions usually include three major elements: an assignment of certain specified state functions to different state organs or offices, the delimitation of the powers of each organ or office, and the establishment of arrangements for their cooperative interaction; a list of individual rights or liberties that are protected against the

exercise of state power; and a statement of the methods by which the constitution may be amended. With these provisions a concentration of power in the hands of a single ruler is prevented, certain areas of political and social life are made immune to governmental intervention, and peaceable change in the political order is made possible. The political party is the other chief instrument of constitutional democracy, for it is the agency through which the electorate is involved in both the exercise and transfer of power. In contrast with the centralized, autocratic direction of the totalitarian single-party organization, with its emphasis on ideological conformity and restricted membership, the political parties of constitutional democracy are decentralized, concerned with the integration of many interests and beliefs, and open to public participation. In constitutional democracies there is usually some measure of competition among two or more parties, each of which, if it cannot hope to form a future government, has some ability to influence the course of state action. The party in a constitutional democracy is at once a means of representing a mass electorate in the exercise of power and also a device for allowing the peaceful replacement of one set of power holders with another.

The distinction between autocratic and nonautocratic rule should not conceal the existence of intermediate types that combine elements of both. In these cases, also, the best procedure for comparative purposes is to investigate the power configurations underlying the formal structures and to examine the extent to which power is concentrated in a single centre or the role that is played by force in the maintenance of the regime. It is a type of analysis that, by guiding attention to the relative weight of coercive and consensual power and the scope of individual freedom in the political order, allows comparisons between systems in terms of their most important attributes.

GOVERNMENTS CLASSIFIED BY STAGE OF DEVELOPMENT

The difficulties of analyzing political change

Political life is shaped by a wide variety of factors, including social and cultural conditions, economic organization, intellectual and philosophical influences, geography or climate, and historical circumstance. Recurrent attempts have been made to reduce this range of variables to analytically manageable dimensions. This is partly the motive, for example, of Marxist and other efforts to relate specific types of political systems to stages of economic development or particular kinds of socioeconomic organization. Although interesting interrelations between political and economic development have been discovered, such monistic, or single-factor, approaches are inadequate to the task of explaining political change. The problem is not only that there are many factors that should be examined but also that they are found in different combinations from one society to another. All political orders are unique as products of history and creations of the peculiar forces and conditions of their environment. A second problem that confronts comparative analysis is the difficulty of devising measures of political development. The definition of what is modern or what constitutes an advanced or developed political system has troubled many recent writers. Clearly, the older notions of development toward the goals of constitutionalism or democracy must now be seriously questioned, and to judge the maturity of a political system in terms of the extent to which it adopts any particular set of institutions or techniques of rule is an equally doubtful procedure. Another difficulty is that political change is not simply a reaction to "objective" factors such as economic forces but also the product of conscious manipulation. In explaining the growth and development of political systems it is impossible to ignore the fact that men, having considered the advantages and disadvantages of different forms of government, often decide to adopt one form rather than another. A similar problem arises from the fact that the nature of the interaction between political systems and their environment is extremely complex. For example, to treat the political system as merely the outgrowth of particular patterns of social or economic organization is to ignore the fact that changes in social and economic structures are often the product, sometimes the intended product, of governmental action.

These difficulties of analysis have prevented the emergence of any satisfactory theory to explain the processes of political change or growth. In the absence of such a theory, however, several writers have recently attempted to identify certain basic phases in the development of national political systems. For example, five major steps in the emergence of the advanced nation-states of the modern world are often distinguished: (1) unification and independence or autonomy; (2) development and differentiation of political institutions and political roles; (3) transfer of power from traditional elites; (4) further institutional and political role differentiation accompanied by the development of a number of organized social interests and growth in governmental functions; and (5) use of state power in attempts to guide or control social and economic activity, extensive exploitation of resources as the result of technological development, and full participation in the international political system.

Steps in the emergence of advanced nation-states

Other writers distinguish among "traditional," "transitional," and "modern" societies in an effort to identify differences and regularities in social, economic, cultural, and political development. The social structure of the traditional society is described as hierarchical, class bound, based on kinship, and divided into relatively few effectively organized social groupings; its economic basis is primarily agricultural, and industry and commerce are relatively undeveloped; its political institutions are those of sacred monarchy, rule by a nobility, and various forms of particularism. The social system of the transitional society is typified by the formation of new classes, especially a middle class and a proletariat, and conflict among ethnic, religious, and cultural groupings; its economic system experiences major tensions as the result of technological development, the growth of industry, urbanization, and the use of rapid communications; its political institutions are typically authoritarian, although constitutional forms also make their appearance. Modernity is seen as the age of high social mobility, equality, universal education, mass communications, increasing secularism, and sociocultural integration; in its economic system, the modern society experiences a further technological revolution, massive urbanization, and the development of a fully diversified economy; its political institutions are those of democracy and modified totalitarianism, and, in either case, a specialized bureaucracy is used to carry on the expanding functions of government.

These efforts to identify stages of "modernization" are poor substitutes for a general theory of political change, but they serve to emphasize the increasing complexity of all the structures—social, economic, and political—of the modern state. The elaboration of the institutions and procedures of modern government appears to be partly a reflection of the social and economic forces at work in the contemporary world and partly the result of efforts to control these forces through governmental action. The complex structures of advanced political orders are treated in the discussion that follows.

The structure of government

The study of governmental structures must be approached with great caution, for political systems having the same kind of legal arrangements and using the same type of governmental machinery often function very differently. A parliament, for example, may be an important and effective part of a political system; or it may be no more than an institutional facade of little practical significance. A constitution may provide the framework within which the political life of a state is conducted; or it may be no more than a piece of paper, its provisions bearing almost no relationship to the facts of political life. Political systems must never be classified in terms of their legal structures alone: the fact that two states have similar constitutions with similar institutional provisions and legal requirements should never, by itself, lead to the conclusion that they represent the same type of political system.

To be useful, the study of governmental structures must always proceed hand in hand with an investigation of the actual facts of the political process: the analyst must exer-

cise the greatest care in distinguishing between form and reality and between prescription and practice. Approached in this way, an examination of the organizational arrangements that governments use for making decisions and exercising power can be a valuable tool of political inquiry.

CONTEMPORARY FORMS OF GOVERNMENT

The rapidity of change

Few states in the modern world have constitutional arrangements that are more than a century old. Indeed, the vast majority of all the world's states have constitutions written in the 20th century. This is true of states, such as Germany, Italy, and Japan, that were defeated in World War II and of other states, such as the successor states of the Soviet Union, Spain, and China, that have experienced civil war and revolutions in the course of the century. Great Britain and the United States are almost alone among major contemporary nation-states in possessing constitutional arrangements that predate the 20th century.

Even in Britain and the United States, the 20th century has seen much change in the governmental system. In the United States, for example, the relationship of legislature and executive at both the national and the state levels has been significantly altered by the growth of bureaucracies and the enlargement of the executive's budgetary powers. In Britain, even more far-reaching changes have occurred in the relationship between the prime minister and Parliament and in Parliament's role in supervising the executive establishment. In both countries, the appearance of the welfare state, the impact of modern technology on the economy, and international crises have resulted in major alterations in the ways in which the institutions of government function and interact.

The modern student of constitutional forms and institutional arrangements confronts an endlessly changing world. In many parts of the world, in countries as different as France, Pakistan, Argentina, and Tanzania, there have been continuing experiments with new constitutions. The adoption of new constitutions is also a major aspect of political change in almost all of the states of eastern Europe and in the successor states of the Soviet Union. All systems, moreover, even without formal constitutional change, undergo a continual process of adjustment and mutation as their institutional arrangements respond to and reflect changes in the social order and the balance of political forces.

Decline of traditional monarchy

Monarchy. The ancient distinction among monarchies, tyrannies, oligarchies, and constitutional governments, like other traditional classifications of political systems, is no longer very descriptive of political life. A king may be a ceremonial dignitary in one of the parliamentary democracies of western Europe, or he may be an absolute ruler in one of the emerging states of North Africa, the Middle East, or Asia. In the first case his duties may be little different from those of an elected president in many republican parliamentary regimes; in the second his role may be much the same as that of countless dictators and strongmen in autocratic regimes throughout the less-developed areas of the world.

It may be said of the reigning dynasties of modern Europe that they have survived only because they failed to retain or to acquire effective powers of government. Royal lines have been preserved only in those countries of Europe in which royal rule was severely limited prior to the 20th century or in which royal absolutism had never firmly established itself. More successful dynasties, such as the Hohenzollerns in Germany, the Habsburgs in Austria-Hungary, and the Romanovs in Russia, which continued to rule as well as to reign at the opening of the 20th century, have paid with the loss of their thrones. Today in countries such as Great Britain or The Netherlands or Denmark the monarch is the ceremonial head of state, an indispensable figure in all great official occasions and a symbol of national unity and of the authority of the state, but is almost entirely lacking in power. Monarchy in the parliamentary democracies of modern Europe has been reduced to the status of a dignified institutional facade behind which the functioning mechanisms of government—cabinet, parliament, ministries, and parties—go about the tasks of ruling.

The 20th century has also seen the demise of most of the hereditary monarchies of the non-Western world. Thrones have toppled in Turkey, in China, in most of the Arab countries, in the principates of India, in the tribal kingdoms of Africa, and in several countries of Southeast Asia. The kings who maintain their position do so less by the claim of legitimate blood descent than by their appeal as popular leaders responsible for well-publicized programs of national economic and social reform or as national military chieftains. In a sense, these kings are less monarchs than monocrats, and their regimes are little different from several other forms of one-man rule found in the modern world.

Dictatorship. While royal rule, as legitimized by blood descent, had almost vanished as an effective principle of government in the modern world, monocracy—a term that comprehends the rule of the remaining non-Western royal absolutists, of the generals and strongmen of Latin America and Asia, of the messianic leaders of postcolonial Africa, and of the totalitarian heads of Communist states—still flourished. Indeed, the 20th century, which has witnessed the careers of Atatürk, Benito Mussolini, Adolf Hitler, Joseph Stalin, Francisco Franco, Mao Tse-tung, Juan Perón, Tito, Gamal Abdel Nasser, Sukarno, Kwame Nkrumah, and Charles de Gaulle, could appear in history as the age of plebiscitary dictatorship.

Monocratic rule

In many of the states of Africa and Asia, for example, dictators quickly established themselves on the ruins of constitutional arrangements inherited from Western colonial powers. In some of these countries, presidents and prime ministers captured personal power by banning opposition parties and building primitive replicas of the one-party systems of the Communist world. In other new countries, the armies seized power, and military dictatorships were established. Whether as presidential dictatorships or as military dictatorships, the regimes that came into being appear to have had common roots in the social and economic problems of the new state. The constitutional systems inherited from the colonial powers proved unworkable in the absence of a strong middle class; local traditions of autocratic rule retained a powerful influence; the army, one of the few organized forces in society, was also often the only force capable of maintaining order; and a tiny intellectual class was impatient for economic progress, frustrated by the lack of opportunity, and deeply influenced by the example of authoritarianism in other countries. The dictatorships that resulted proved highly unstable, and few of the individual dictators were able to satisfy for long the demands of the different groups that supported their bids for power.

Although similar in some respects to the dictatorships of the new nations, the caudillos of 19th- and 20th-century Latin America represent a very different type of monocratic rule. In its 19th-century form, *caudillismo* was the result of the breakdown of central authority. After a brief period of constitutional rule, each of the former Spanish colonies in the Americas experienced a collapse of effective national government. A self-proclaimed leader, usually an army officer, heading a private army typically formed from the peasantry with the support of provincial landowners, established his control over one or more provinces, and then marched upon the national capital. The famous 19th-century caudillos—Antonio López de Santa Anna of Mexico or Juan Manuel de Rosas of Argentina, for example—were thus essentially provincial leaders who seized control of the national government to maintain the social and economic power of provincial groups. The 20th-century dictatorships in Latin-American countries have had different aims. The modern caudillo is less a provincial than a national leader. The Perón regime, for example, was established by nationalistic army officers committed to a program of national reform and ideological goals. Often, too, recent dictators in Latin America have allied themselves with a particular social class, attempting either to maintain the interests of established economic groupings or to press social reforms.

Dictatorship in the technologically advanced, totalitarian regimes of modern Communism is distinctively different from the authoritarian regimes of either

Modern totalitarian regimes

Latin America or the new states of Africa and Asia. Nazi Germany under Hitler and the Soviet Union under Stalin are the leading examples of modern totalitarian dictatorships. The crucial elements of both were the identification of the state with the single mass party and of the party with its charismatic leader, the use of an official ideology to legitimize and maintain the regime, the employment of a terroristic police force and a controlled press, and the application of all the means of modern science and technology to control the economy and individual behaviour. The two systems, however, may be distinguished in several ways. Fascism, in its National Socialist form, was primarily a counterrevolutionary movement that mobilized middle- and lower middle-class groups to pursue nationalistic and militaristic goals and whose sole principle of organization was obedience to the Führer. By contrast, Soviet Communism grew out of a revolutionary theory of society, pursued the goal of revolutionary overthrow of capitalist systems internationally, and employed the complex bureaucratic structures of the Communist Party as mechanisms of governmental organization.

Western constitutional democracies have provided examples of another type of contemporary dictatorship. At various points in the 20th century, during periods of domestic or foreign crisis, most constitutional regimes have conferred emergency powers on the executive, suspending constitutional guarantees of individual rights or liberties or declaring some form of martial law. Indeed, the constitutions of some Western democracies explicitly provide for the grant of emergency powers to the executive in a time of crisis to protect the constitutional order. In many cases, of course, such provisions have been the instruments with which dictators have overthrown the regime. Thus, the proclamation of emergency rule was the beginning of the dictatorships of Mussolini in Italy, of Kemal Atatürk in Turkey, of Józef Piłsudski in Poland, of Antônio de Olveira Salazar in Portugal, of Franz von Papen and Hitler in Germany, and of Engelbert Dollfuss and Kurt von Schuschnigg in Austria. In other democracies, however, constitutional arrangements have survived quite lengthy periods of crisis government. After World War II, for example, in both the United States and Britain, the use of extraordinary powers by the executive came to a halt with the end of the wartime emergency. Similarly, although the 1958 constitution of the Fifth Republic of France contained far-reaching emergency powers conferred on the president—"when the institutions of the Republic, the independence of the nation, the integrity of its territory or the fulfillment of its international obligations are threatened with immediate and grave danger, and when the regular functioning of the constitutional authority is interrupted"—their implicit threat to the constitutional order has not been realized.

Many forces at work in the 20th century appear to lend impetus to the rise of monocratic forms of rule. In nearly all political systems, the powers of chief executives have increased in response to the demanding social, economic, and military crises of the age. The complex decisions required of governments in a technological era, the perfectionist impulses of the great bureaucratic structures that have developed in all industrialized societies, and the imperatives of national survival in a nuclear world continue to add to the process of executive aggrandizement. The question for many constitutional regimes is whether the limitation and balance of power that are at the heart of constitutional government can survive the growing enlargement of executive power.

Oligarchy. In the Aristotelian classification of government, there were two forms of rule by the few: aristocracy and its debased form, oligarchy. Although the term oligarchy is rarely used to refer to contemporary political systems, the phenomenon of irresponsible rule by small groups has not vanished from the world.

Small leadership groups

Many of the classical conditions of oligarchic rule were found until recently in those parts of Asia in which governing elites were recruited exclusively from a ruling caste—a hereditary social grouping set apart from the rest of society by religion, kinship, economic status, prestige, and even language. In the contemporary world, in some countries that have not experienced the full impact of industrialization, governing elites are still often recruited from a ruling class—a stratum of society that monopolizes the chief social and economic functions in the system. Such elites have typically exercised power to maintain the economic and political status quo.

The simple forms of oligarchic rule associated with preindustrial societies are, of course, rapidly disappearing. Industrialization produces new, differentiated elites that replace the small leadership groupings that once controlled social, economic, and political power in the society. The demands of industrialization compel recruitment on the basis of skill, merit, and achievement rather than on the basis of inherited social position and wealth. New forms of oligarchic rule have also made their appearance in many advanced industrial societies. Although governing elites in these societies are no longer recruited from a single class, they are often not subjected to effective restraints on the exercise of their power. Indeed, in some circumstances, the new elites may use their power to convert themselves into a governing class whose interests are protected by every agency of the state.

Oligarchic tendencies of a lesser degree have been detected in all the great bureaucratic structures of advanced political systems. The growing complexity of modern society and its government thrusts ever greater power into the hands of administrators and committees of experts. Even in constitutional regimes, no fully satisfactory answer has been found to the question of how these bureaucratic decision makers can be held accountable and their powers effectively restrained without, at the same time, jeopardizing the efficiency and rationality of the policy-making process.

Constitutional government. Constitutional government is defined by the existence of a constitution—which may be a legal instrument or merely a set of fixed norms or principles generally accepted as the fundamental law of the polity—that effectively controls the exercise of political power. The essence of constitutionalism is the control of power by its distribution among several state organs or offices in such a way that they are each subjected to reciprocal controls and forced to cooperate in formulating the will of the state. Although constitutional government in this sense flourished in England and in some other historical systems for a considerable period, it is only recently that it has been associated with forms of mass participation in politics. In England, for example, constitutional government was not harnessed to political democracy until after the Reform Act of 1832 and subsequent 19th-century extensions of the suffrage. In the contemporary world, however, constitutional governments are also generally democracies, and in most cases they are referred to as constitutional democracies or constitutional-democratic systems.

Political parties

The contemporary political systems that combine constitutionalism and democracy share a common basis in the primacy they accord to the will of the majority of the people as expressed in free elections. In all such systems, political parties are key institutions, for they are the agencies by which majority opinion in a modern mass electorate is mobilized and expressed. Indeed, the history of the political party in its modern form is coincidental with the development of contemporary constitutional-democratic systems. In each case, the transition from the older forms of constitutionalism to modern constitutional democracy was accompanied by the institutionalization of parties and the development of techniques of party competition. The essential functions of political parties in a constitutional democracy are the integration of a multitude of interests, beliefs, and values into one or more programs or proposals for change and the nomination of party members for elective office in the government. In both functions, the party serves as a link between the rulers and the ruled: in the first case by allowing the electorate to register an opinion on policy and in the second by giving the people a chance to choose their rulers. Of course, the centralized, autocratically directed, and ideologically orthodox one-party systems of totalitarian regimes perform neither of these functions.

The two major types of constitutional democracy in the modern world are exemplified by the United States and Great Britain. The United States is the leading example of the presidential system of constitutional democracy; Britain, although its system is sometimes referred to as a cabinet system in recognition of the role of the Cabinet in the government, is the classic example of the parliamentary system. The U.S. presidential system is based on the doctrine of separation of powers and distinguishes sharply between the personnel of the legislature and the executive; the British parliamentary system provides for the integration or fusion of legislature and executive. In the U.S. system the separation of legislature and executive is reinforced by their separate election and by the doctrine of checks and balances that provides constitutional support for routine disagreements between the branches; in the British system the integration of legislature and executive is reinforced by the necessity for their constant agreement, or for a condition of "confidence" between the two, if the normal processes of government are to continue. In the U.S. system reciprocal controls are provided by such devices as the presidential veto of legislation (which may be overridden by a two-thirds majority in Congress), the Senate's role in ratifying treaties and confirming executive nominations, congressional appropriation of funds and the exclusive ability to declare war, and judicial review of legislation; in the British system the major control device is the vote of "no confidence" or the rejection of legislation that is considered vital.

The prestige of constitutional democracy was once so great that many thought all the countries of the world would eventually accede to the examples of the United States or Britain and establish similar arrangements. However, the collapse of the Weimar Constitution in Germany in the 1930s and the recurrent political crises of the Fourth Republic in France after World War II suggested that constitutional democracy carries no guarantee of stability. The failure of both presidential and parliamentary systems to work as expected in less-advanced countries that modelled their constitutions on those of the United States and Britain resulted in a further diminution in the prestige of both systems. Functioning examples are located throughout the world, though these are generally poorly institutionalized outside of those countries with direct historical ties to western Europe. Japan is a notable exception to this generalization, as are Costa Rica, India, and several other states to a lesser degree.

CONTEMPORARY LEVELS OF GOVERNMENT

Most national societies have passed through a stage in their social and political development, usually referred to as feudalism, in which a weak and ineffectively organized national government competes for territorial jurisdiction with local power holders. In medieval England and France, for example, the crown was perennially threatened by the power of the feudal nobles, and a protracted struggle was necessary before the national domain was subjected to full royal control. Elsewhere, innumerable societies continued to experience this kind of feudal conflict between local magnates and the central government well into the modern era. The warlords of 19th- and 20th-century China, for example, were just as much the products of feudal society as the warring barons of 13th-century England and presented the same kind of challenge to the central government's claim to exercise sovereign jurisdiction over the national territory. By the 1970s, feudalism was almost extinct. The social patterns that had formerly supported the power of local landowners were rapidly disappearing, and central governments had generally acquired a near monopoly of communications and military technology, enabling them to project their power into areas once controlled by local rulers.

Feudalism

In nearly all national political systems, central governments are better equipped than ever before to exercise effective jurisdiction over their territories. In much of the developing world, nationalist political movements and a variety of modern economic forces have swept away the traditional structures of local government, and the quasi-autonomous governments of village and tribe and province have been replaced by centrally directed systems of subnational administration. Even in the heavily industrialized states of the modern world, there has been an accelerating tendency toward greater centralization of power at the national level. In the United States, for example, the structure of relationships among the governments at the national, state, and local levels has changed in a number of ways to add to the power of the federal government in Washington. Even though the system of national grants-in-aid appears to have been designed as a means of decentralizing administration, the effect has been decidedly centralist, for the conditional character of the grants has allowed the federal government to exercise influence on state policies in fields that were once invulnerable to national intervention.

National government. The nation-state is the dominant type of political system in the contemporary world, and nationalism, or the creed that centres the supreme loyalty of the people upon the nation-state, is the dominating force in international politics. The national ideal triumphed as a result of the wars of the 19th and 20th centuries. The Napoleonic Wars, which spread the doctrines of the French Revolution, unleashed nationalism as a force in Europe and led to the Risorgimento in Italy and the emergence of Bismarck's Germany. The two world wars of the 20th century carried the principles of national self-determination and liberal democracy around the world and gave birth to the independence movements that resulted in the foundation of new states in eastern Europe in 1919 and the emergence from colonial status of countries in Asia and Africa after 1945. The collapse of the Warsaw Pact and the Soviet Union itself completed this process of moving from multinational empires to truly sovereign national states.

The nation-state

All the major forces of world politics—*e.g.*, war, the development of national economies, and the demand for social services—have reinforced the national state as the primary focus of people's loyalties. Wars have played the major part in strengthening national governments and weakening political regionalism and localism. The attachments that people have to subnational political communities are loosened when they must depend for their security on the national power. Even in the new age of total war—which few nations are capable of waging and even fewer of surviving—people look for their security to national governments rather than to international organizations. In nearly all contemporary states, the national budget is dominated by expenditures for defense, the military employs the largest fraction of the work force, and questions of national security pervade the discussion of politics.

One of the lessons of the late 20th century is that national sovereignty continues to be the most important obstacle not only to the emergence of new forms of supranational government but to effective international cooperation as well. Almost everywhere, attempts to achieve federation and other forms of multinational communication have foundered on the rocks of nationalism. The collapse of the Federation of Rhodesia and Nyasaland and the Federation of Malaya, for example, were paralleled by the seeming ineffectiveness of the Organization of American States and the Arab League. On another level is the collapse of the Warsaw Pact when the nations of eastern Europe reclaimed their sovereignty in the late 1980s, after decades of domination by the Soviet Union. In western Europe, nations have joined together to form a confederation known as the European Union (EU). These countries are united not only by a long history and a common cultural inheritance but also by the expectation of mutual economic advantage. Even in this case, nationalism has proved to be an obstacle to the most ambitious goals of unification, which would severely limit national sovereignty in some spheres.

At the international level, anarchy is the principal form of contemporary rule, for the nation-state's freedom of action is limited only by its power. While the state's freedom of action may not be directly threatened, the effectiveness of the state's action in the economic realm is increasingly being called into question. The development of national industries in the 19th and early 20th centuries played a major part in strengthening national as against regional

and local political entities, but the scale of economic activity has now outgrown national markets. Industrial combines and commercial groupings have emerged that cross national frontiers and require international markets. This tight integration of the world economy has limited the effectiveness of some traditional instruments used to influence national trends in capitalist economies.

It is increasingly clear that some aspects of traditional sovereignty may be affected by serious efforts to confront some issues that act on the entire international system. National frontiers can no longer be adequately defended in an era of intercontinental ballistic missiles, especially with the rapid diffusion of the technology required for delivery systems as well as for nuclear weapons themselves. Action in this area is, by definition, an attempt to shape the national security policy of states, something very near the core of a state's sovereignty. Concern over environmental matters could lead to more restrictive regimes than any arms-control provisions, ultimately shaping the way in which nations evolve economically. Destruction of major ecosystems, wasteful use of energy, and industrialization based on the use of fossil fuels are all national policies with international repercussions. As technology empowers more countries to directly affect the state of the planet as well as other nations, there are increasing incentives to limit the domestic policy choices of all nations.

Although the failure of efforts to achieve world government and to develop an effective system of international law may be regretted, it should perhaps be remembered that the nation-state continues to function as an extremely effective system for maintaining order within its boundaries. In some cases, this is achieved with remarkably little coercion and in such a way that the progress of civilization is encouraged. Under present-day conditions, world government might well involve much higher levels of coercion and much less civilization.

Decline of regional and state government

Regional and state government. The 18th-century political philosopher Montesquieu wrote that governments are likely to be tyrannical if they are responsible for administering large territories, for they must develop the organizational capacity characteristic of despotic states. It was partly this fear that led the American founding fathers to provide for a federal system and to divide governmental functions between the government in Washington and the state governments. Modern technology and mass communication are often said to have deprived Montesquieu's axiom of its force. Yet the technology that makes it possible for large areas to be governed democratically also holds out the spectre of an even greater tyranny than Montesquieu foresaw.

In all political systems the relationships between national and regional or state governments have been affected by technology and new means of communication. In the 18th century Thomas Jefferson—in arguing that local government, or the government closest to the people, was best—could claim that citizens knew most about their local governments, somewhat less about their state governments, and least about the national government. In the present-day United States, however, the concentration of the mass media on the issues and personalities of national government has made nonsense of this proposition. As several recent studies have demonstrated, people know much less about local government than national government and turn out to vote in much larger numbers in national elections. The necessity for employing systems for the devolution of political power is reduced when a central government can communicate directly with citizens in all parts of the national territory, and the vitality of subnational levels of government is sapped when public attention is focussed on national problems.

Another general development that has lessened the importance of regional or state government is the rise of efficient national bureaucracies. In nearly all political systems, there has been some tendency toward bureaucratic centralization, and in some cases national bureaucracies have almost completely replaced older systems of regional and provincial administration. In the United States, for example, complex programs of social security, income taxes, agricultural subsidies, and many others that bear directly on individuals are centrally administered.

Even in systems in which a division of functions between national and subnational governments is constitutionally prescribed, the prevailing trend in intergovernmental relations is toward increasing involvement of the national government in areas once dominated by regional or state governments. Thus, the original constitutional arrangements prescribed by the Allied powers for the West German republic in 1949 won general acclaim at the time because they provided for greater decentralization than had the Weimar Constitution; but, as soon as Germany was free to amend its own constitution, several state functions were reassigned to the national government. In the United States, also, the collapse of the doctrine of "dual federalism," according to which the powers of the national government were restricted by the powers reserved to the states, signalled the end of an era in which the states could claim exclusive jurisdiction over a wide range of functions. Today, forms of cooperative federalism involving joint action by national and state governments are increasingly common. Such cooperative relationships in the United States include programs of public assistance, the interstate highway system, agricultural extension programs, and aid to education. In some areas, such as school desegregation, the national government has used broad powers to compel states to conform to national standards.

Efforts made to halt the trend toward centralization and to reinvigorate regional or state governments have met with little success. In the United States a Commission on Intergovernmental Relations established by President Eisenhower in 1953 concluded that it could recommend no major reversion of functions to state governments. Similarly, efforts in France and Italy to decentralize parts of the national administrations have had few practical results. Political regionalism appears to be in steep decline almost everywhere, whether in China or in the American South. The attachments that bind people to localities and allow the growth of genuine subnational political communities have weakened under the impact of technology and the growth of national economies. Only where political regionalism has always been in reality a cloak for movements of national independence—for example, in Scotland, Wales, Northern Ireland, Quebec, and Brittany—are there popular attempts to reverse the trend toward national centralization.

The urban metropolis

City and local government. Political scientists since Aristotle have recognized that the nature of political communities changes when their populations grow larger. One of the central problems of contemporary government is the vast increase in urban population and the progression from "polis to metro-polis to mega-polis." The catalog of ills that have resulted from urban growth includes political and administrative problems of extraordinary complexity. Aging infrastructure has become an issue of pressing national importance in the United States, with the major cities obviously suffering in this area. Grave social problems—for example, violent crime (especially that committed by youths in poverty-stricken areas), drug trafficking, unemployment, and homelessness—are concentrated to such a degree that they directly shape the environment in many large urban areas. The majority of cities are ill equipped to handle these problems without significant assistance from the national government.

The tax base of U.S. cities has dwindled with the flight of the middle classes to the suburbs and the relocation of industry. Largely as a result of this, political power has begun to follow wealth out of the cities and into adjoining suburbs. These outlying areas have not only increased significantly in population, but, compared with that of many large cities, it is a population more likely to vote and otherwise lobby for its interests. This has served to reduce the national government's activism in the cities at the very time when most cities are suffering from a drastically reduced capacity to act. Aside from such fiscal and political pressures, however, the national government is inevitably concerned with the threats posed by racial conflict, ghetto violence, and other kinds of social chaos in the cities.

The metropolis suffers from several acute governmental and administrative failures. Responsibility for the issues

that transcend the boundaries of local governments has not been defined, for representative institutions have failed to develop at the metropolitan level. In most cases, there are no effective governmental structures for administering area-wide services or for dealing comprehensively with the common problems of the metropolitan community. The result has been the appearance of a new class of problems created by government itself, including uneven levels of service for metropolitan residents, inequities in financing government services and functions, and variations in the democratic responsiveness of the governments scattered through the metropolitan area. The tangled pattern of local governments, each operating in some independent sphere, does not allow the comprehensive planning necessary to deal with the escalating problems of urban life.

Efforts to create new governing structures for metropolitan communities are among the most interesting developments in contemporary government. In the United States these efforts include the creation of special districts to handle specific functions, area-wide planning agencies, interstate compacts, consolidated school and library systems, and various informal intergovernmental arrangements. Although annexation of outlying areas by the central city and city–county consolidations have been attempted in many cases, the reluctance of urban areas to surrender their political independence or to pay for central-city services has been an obstacle. The Los Angeles plan, by which the county has assumed responsibility for many area-wide functions, leaving the local communities with substantial political autonomy, may represent a partial solution to the problem of urban–suburban tensions. In other cases, "metropolitan federation" has been attempted. One of the earliest and most influential examples of a federated system of metropolitan government is the Greater London Area in Britain, which encompasses 32 London boroughs and places effective governing powers in the hands of an elected city council. In Canada the city of Toronto and its suburbs adopted a metropolitan "constitution" in 1953 under which mass transit, highways, planning, and several other functions are controlled by a council composed of elected officials from the central city and surrounding governments. In 1957 Miami and Dade County in Florida chose "metro" government: Miami and 27 suburban cities retained control of local functions, while area-wide functions such as fire and police protection and transportation were allocated to the new federal structure. In 1962 Nashville, Tennessee, combined with Davidson County to form a single metropolitan government; and in 1968 Jacksonville, Florida, joined with Duval County in a similar arrangement. Other examples of various degrees of area-wide consolidation in the United States include Baton Rouge, Louisiana; Seattle, Washington; Portland, Oregon; and Indianapolis, Indiana.

Most of the major problems of contemporary politics seem to have found their focus in the metropolis, and there is almost universal agreement that new governing systems must be devised for the metropolitan community if the problems are ever to be resolved.

CONTEMPORARY DIVISIONS OF GOVERNMENT

Legislative, executive, and judicial functions

In his *Politics,* Aristotle differentiated three categories of state activity—deliberations concerning common affairs, decisions of executive magistrates, and judicial rulings—and indicated that the most significant differences among constitutions concerned the arrangements made for these activities. This threefold classification is not precisely the same as the modern distinction among legislature, executive, and judiciary. Aristotle intended to make only a theoretical distinction among certain state functions and stopped short of recommending that they be assigned as powers to separate organs of government. Indeed, since Aristotle held that all power should be wielded by one man, pre-eminent in virtue, he never considered the concept of separated powers. In the 17th century the English political philosopher John Locke also distinguished the legislative from the executive function but, like Aristotle, failed to assign these to separate organs or institutions. Montesquieu was the first to make the modern division among legislative, executive, and judiciary. Arguing that the purpose of political association is liberty, not virtue, and that the very definition of liberty's great antagonist, tyranny, is the accumulation of all power in the same hands, he urged the division of the three functions of government among three separate institutions. After Montesquieu, the concept of separation of powers became one of the principal doctrines of modern constitutionalism. Nearly all modern constitutions, from the document written at Philadelphia in 1787 through the French Declaration of the Rights of Man and of the Citizen of August 1789 up to the constitutions of the new states of Africa and Asia, provide for the separate establishment of legislative, executive, and judiciary. The functional division among the branches of government is never precise. In the American system, for example, the doctrine of checks and balances justifies several departures from the strict assignment of functions among the branches. Parliamentary forms of government depart even further from the concept of separation and integrate both the personnel and the functions of the legislature and the executive. Indeed, the principle of shared rather than separated powers is the true essence of constitutionalism. In the constitutional state, power is controlled because it is shared or distributed among the divisions of government in such a way that they are each subjected to reciprocal checks and forced to cooperate in the exercise of political power. In the nonconstitutional systems of totalitarianism or autocracy, although there may be separate institutions such as legislatures, executives, and judiciaries, power is not shared but rather concentrated in a single organ. Because this organ is not subjected to the checks of shared power, the exercise of political power is uncontrolled or absolute.

The legislature. The characteristic function of all legislatures is the making of law. In most systems, however, legislatures also have other tasks, such as selection and criticism of the government, supervision of administration, appropriation of funds, ratification of treaties, impeachment of executive and judicial officials, acceptance or refusal of executive nominations, determination of election procedures, and public hearings on petitions. Legislatures, then, are not simply lawmaking bodies. Neither do they monopolize the function of making law. In most systems the executive has a power of veto over legislation, and, even where this is lacking, the executive may exercise original or delegated powers of legislation. Judges, also, often share in the lawmaking process, through the interpretation and application of statutes or, as in the U.S. system, by means of judicial review of legislation. Similarly, administrative officials exercise quasi-legislative powers in making rules and deciding cases that come before administrative tribunals.

Varieties of legislatures

Legislatures differ strikingly in their size, the procedures they employ, the role of political parties in legislative action, and their vitality as representative bodies. In size, the British House of Commons is among the largest; the Icelandic lower house, the New Zealand House of Representatives, and the Senate of Nevada are among the smallest. Most legislatures are bicameral, although New Zealand, Denmark, the state of Queensland, in Australia, and Nebraska, in the United States, have all abolished their second chambers. The procedures of the United States House of Representatives, which derive from a manual of procedure written by Thomas Jefferson, are among the most elaborate of parliamentary rules, requiring study and careful observation over a considerable period before members become proficient in their manipulation. Voting procedures range from the formal procession of the division or teller vote in the British House of Commons to the electric voting methods employed in the California legislature and in some other American states. Another point of difference among legislatures concerns their presiding officers. These are sometimes officials who stand above party and, like the speaker of the British House of Commons, exercise a neutral function as parliamentary umpires; sometimes they are the leaders of the majority party and, like the speaker of the United States House of Representatives, major political figures; and sometimes they are officials who, like the vice president of the United States in his role as presiding officer of the Senate, ex-

ercise a vote to break ties and otherwise perform mainly ceremonial functions.

Legislative parties are of various types and play a number of roles or functions. In the United States House of Representatives, for example, the party is responsible for assigning members to all standing committees; the party leadership fills the major parliamentary offices, and the party membership on committees reflects the proportion of seats held by the party in the House as a whole. The congressional party, however, is not disciplined to the degree found in British and some other European legislative parties, and there are relatively few "party line" votes in which all the members of one party vote against all the members of the other party. In the House of Commons, party-line voting is general; indeed, it is very unusual to find members voting against their party leadership, and, when they do, they must reckon with the possibility of penalties such as the "withdrawal of the whip" or the loss of their official status as party members.

It is often said that the 20th century has dealt harshly with legislatures and that this is an age of executive aggrandizement. Certainly, executives in most countries have assumed an increasingly large role in the making of law, through the initiation of the legislation that comes before parliaments, assemblies, and congresses, through the exercise of various rule-making functions, and as a result of the growth of different types of delegated legislation. It is also true that executives have come to predominate in the sphere of foreign affairs and, by such devices as executive agreements, which are frequently used in place of treaties, have freed themselves from dependence upon legislative approval of important foreign-policy initiatives. Moreover, devices such as the executive budget and the rise of specialized budgetary agencies in the executive division have threatened the traditional fiscal controls of legislatures. This decline in legislative power, however, is not universal. The United States Congress, for example, has preserved a substantial measure of its power. Indeed, congressional oversight of the bureaucracy is an area in which it has added to its power and has developed new techniques for controlling the executive. The difficulties of presidents in the late 20th century with legislative programs of foreign aid and the perennial congressional criticism of executive policies in foreign affairs also suggest that Congress continues to play a vital role in the governing process.

The executive. Political executives are government officials who participate in the determination and direction of government policy. They include heads of state and government leaders—presidents, prime ministers, premiers, chancellors, and other chief executives—and many secondary figures, such as cabinet members and ministers, councillors, and agency heads. By this definition, there are several thousand political executives in the U.S. national government, including the president, dozens of political appointees in the Cabinet departments, in the agencies, in the commissions, and in the White House staff, and hundreds of senior civil servants. The same is true of most advanced political systems, for the making and implementation of government policy require very large executive and administrative establishments.

The crucial element in the organization of a national executive is the role assigned to the chief executive. In presidential systems, such as in the United States, the president is both the political head of the government and also the ceremonial head of state. In parliamentary systems, such as in Great Britain, the prime minister is the national political leader, but another figure, a monarch or elected president, serves as the head of state. In mixed presidential–parliamentary systems, such as that established in France under the constitution of 1958, the president serves as head of state but also wields important political powers, including the appointment of a prime minister and Cabinet to serve as the government.

Varieties of executives

The manner in which the chief executive is elected or selected is often decisive in shaping his role in the political system. Thus, although he receives his seals of office from the monarch, the effective election of a British prime minister usually occurs in a private conclave of the leading members of his party in Parliament. Elected to Parliament from only one of more than 630 constituencies, he is tied to the fortunes of the legislative majority that he leads. By contrast, the American president is elected by a nationwide electorate, and, although he leads his party's ticket, his fortunes are independent of his party. Even when the opposition party controls the Congress, his fixed term and his independent base of power allow him considerable freedom of manoeuvre. These contrasts explain many of the differences in the roles of the two chief executives. The British prime minister invariably has served for many years in Parliament and has developed skills in debate and in political negotiation. His major political tasks are the designation of the other members of the Cabinet, the direction of parliamentary strategy, and the retention of the loyalty of a substantial majority of his legislative party. The presidential chief executive, on the other hand, often lacks prior legislative and even national-governmental experience, and his main concern is with the cultivation of a majority in the electorate through the leadership of public opinion. Of course, since the president must have a legislative program and often cannot depend on the support of a congressional majority, he may also need the skills of a legislative strategist and negotiator.

Another important area of contrast between different national executives concerns their role in executing and administering the law. In the U.S. presidential system, the personnel of the executive branch are constitutionally separated from the personnel of Congress: no executive officeholder may seek election to either house of Congress, and no member of Congress may hold executive office. In parliamentary systems the political management of government ministries is placed in the hands of the party leadership in parliament. In the U.S. system the president often appoints to Cabinet positions persons who have had little prior experience in politics, and he may even appoint members of the opposition party. In the British system, Cabinet appointments are made to consolidate the prime minister's personal ascendancy within the parliamentary party or to placate its different factions. These differences extend even further into the character of the two systems of administration and the role played by civil servants. In the U.S. system a change in administration is accompanied by the exodus of a very large number of top government executives—the political appointees who play the vital part in shaping day-to-day policy in all the departments and agencies of the national government. In Britain, when political control of the House of Commons changes, only the ministers, their parliamentary secretaries, and one or two other top political aids are replaced. For all practical purposes, the ministries remain intact and continue under the supervision of permanent civil servants.

In nearly all political systems, even in constitutional democracies where executive responsibility is enforced through free elections, the 20th century has seen an alarming increase in the powers of chief executives. The office of the presidency in the United States, like the office of prime minister in Britain, has greatly enlarged the scope of its authority. One of the challenges of representative government is to develop more constitutional restraints on the abuse of executive powers while retaining their advantages for effective rule.

The judiciary. Like legislators and executives, judges are major participants in the policy-making process; and courts, like legislatures and administrative agencies, promulgate rules of behaviour having the nature of law. The process of judicial decision making, or adjudication, is distinctive, however, for it is concerned with specific cases in which an individual has come into conflict with society by violating its norms or in which individuals have come into conflict with one another, and it employs formal procedures that contrast with those of parliamentary or administrative bodies.

Varieties of court systems

Established court systems are found in all advanced political systems. Usually there are two judicial hierarchies, one dealing with civil and the other with criminal cases, each with a large number of local courts, a lesser number at the level of the province or the region, and one or more courts at the national level. This is the pattern of judicial organization in Britain, for example. In some countries—

for example, in France—although there is a double hierarchy, the distinction is not between courts dealing with criminal cases and other courts dealing with civil cases but rather between those that handle all civil and criminal cases and those that deal with administrative cases or challenges to the administrative authority of the state. Reflecting the federal organization of its government, the United States has two court systems: one set of national courts and 50 sets of state courts. By contrast, Germany, which is federal in governmental organization, possesses only a single integrated court system.

Local courts are found in all systems and are usually of two types. The first type deals with petty offenses and may include a traffic court, a municipal court, a small-claims court, and a court presided over by a justice of the peace or a local magistrate. The second type, sometimes called trial courts, are courts of first instance in which most cases of major importance are begun. These are the state superior courts in the United States, the county courts and quarter sessions in Britain, the *tribunal de grande instance* in France, and the district courts, or *Landgerichte,* in Germany. In some systems there is a level above the local court, usually referred to as assize courts, in which exceptionally serious crimes, such as homicide, are tried. Courts of appeal review the procedures and the law in the lower court and, in some instances, return the case for a new trial. In all systems there are national supreme courts that hear appeals and exercise original jurisdiction in cases of the greatest importance, such as those involving conflict between a state and a national government. Outside the regular court systems, there are sometimes found specialized judicial tribunals, such as administrative courts, or courts of claims that deal with special categories of cases.

The functions of government

In all modern states, governmental functions have greatly expanded with the emergence of government as an active force in guiding social and economic development. In Socialist countries, government has a vast range of responsibilities for many types of economic behaviour. Even in the United States, where there remains a much greater attachment than in most societies to the idea that government should be only an umpire adjudicating the rules by which other forces in society compete, such governmental activities as the Tennessee Valley Authority or the use of credit controls to prevent economic fluctuations are now accepted with relatively little question. Government has thus become the major or even the dominant organizing power in all contemporary societies.

The role of government

The historical stages by which governments have come to exercise their contemporary functions make an interesting study in themselves. The scope of government in the ancient polis involved the comprehensive regulation of the ends of human existence. As Aristotle expressed it, what was not commanded by government was forbidden. The extent of the functions of government in the ancient world was challenged by Christianity and its insistence on a division of those things that belong separately to Caesar and to God. When the feudal world succeeded the Roman Empire, however, the enforcement of the sanctions of religion became one of the first objects of political authority. The tendencies that began in the 18th century separated church from state and state from society, and the modern concept of government came into being. The American colonies' Declaration of Independence expresses the classic modern understanding of those ends that governmental functions exist to secure. The first aim of government is to secure the right to life; this comprehends the safety of fellow citizens as regards one another and the self-preservation of the nation as regards foreign powers. Life exists for the exercise of liberty, in terms of both natural and civil rights, and these, along with other specific functions of government, provide those conditions upon which men may pursue happiness, an end that is finally entirely private and beyond the competence of government.

With the advent of the Marxist conception of the state, the ends of human existence once again became the objects of comprehensive government regulation. Marxism sees the state as a product of class warfare that will pass out of existence in the future age of perfect freedom. Aristotle believed human perfection to be possible only within political society; Marx believed that the perfection of man would follow upon the abolition of political society. Before the final disposal of the state, however, many Marxists believe that forceful use of governmental power is justified in order to hasten mankind's progress toward the last stage of history.

THE TASKS

The maintenance of authority

Self-preservation. The first right of men and nations is self-preservation. The task of maintaining the nation, however, is more complex than the individual's duty of self-preservation, for the nation must seek to command the attachment of a community of citizens as well as to preserve itself from external violence. As Thomas Hobbes insisted, civil war constitutes the greatest threat to governments, for it represents the dissolution of the "sovereign power." In modern terms, civil war signifies that the government has lost one of the basic attributes of political authority: its monopoly of force and its control over the use of violence. In a fundamental sense, political authority may be preserved from the threat of civil war only when there exists in the political community an agreement on the basic principles of the regime. Such a consensus is the result, among other things, of a shared "ideology" that gives fellow citizens a sense of communal belonging and recognizes interlocking values, interests, and beliefs. Ideology, in this sense, may be the product of many different forces. Sometimes it is associated with ancient customs, sometimes with religion, sometimes with severe dislocations or the sort of common need that has led to the formation of many nation-states, and sometimes with the fear of a common enemy. The ideological commitment that people call patriotism is typically the product of several of these forces.

Governments neglect at their peril the task of strengthening the ideological attachment of their citizens to the regime. In this sense, civic education should be counted among the essential functions of the state, for it is primarily through systems of education that citizens learn their duties. Indeed, as a number of sociological studies have shown, the process of political socialization that transforms people into citizens begins in kindergarten and grade school. Even more than this, education is the instrument by which governments further the cohesion of their societies and build the fundamental kinds of consensus that support their authority. It is not surprising, therefore, that national systems of education are often linked to central elements of the regimes. In France public education was traditionally mixed with the teachings of the Roman Catholic church; in Great Britain a private system of education supported the class divisions of society; and in the United States a primarily secular form of public education traditionally used constitutional documents as the starting point of children's training in patriotism.

The preservation of the authority of the state also requires a governmental organization capable of imposing its jurisdiction on every part of the national territory. This involves the maintenance of means of communication, the use of administrative systems, and the employment of police forces capable of controlling domestic violence. The police function, like education, is often a key to the character of a regime. In Nazi Germany, Hitler's Brownshirts took over the operation of local and regional police systems and often supervised the administration of law in the streets. In the Soviet Union the security police acted to check any deviation from the policy of the party or state. In the United States the police powers are left in the hands of the 50 states and the local agencies of government. With the exception of certain offenses created by the McCarran Act and some parallel statutes, political crimes as such are unknown. In addition, there are state militias that act, under the control of the governors of the various states, in moments of local emergency, such as riots or natural catastrophes. The Federal Bureau of Investigation (FBI), the only equivalent of a national police force, is an agency

Police functions and regime type

established to carry out specific assignments dealing with a limited but important class of crimes. Since there is no comprehensive federal criminal code, there is not, strictly speaking, a federal police.

Governments must preserve themselves against external as well as domestic threats. For this purpose they maintain armed forces and carry on intelligence activities. They also try to prevent the entry of aliens who may be spies or saboteurs, imprison or expel the agents of foreign powers, and embargo the export of materials that may aid a potential enemy. The ultimate means of preserving the state against external threats, of course, is war. In war, governments usually enlarge the scope of their domestic authority; they may raise conscript forces, imprison conscientious objectors, subject aliens to internment, sentence traitors to death, impose extraordinary controls on the economy, censor the press, compel settlement of labour disputes, impose internal-travel limitations, withhold passports, and provide for summary forms of arrest.

Many forces generate clashes between nations, including economic rivalry and disputes over trade, the desire to dominate strategic land or sea areas, religious or ideological conflict, and imperialistic ambition. All national governments develop organizations and policies to meet these and other situations. They have foreign ministries for the conduct of diplomatic relations with other states, for representing them in international organizations, and for negotiating treaties. Some governments conduct programs such as foreign aid, cultural exchange, and other activities designed to win goodwill abroad.

In the 20th century, relationships among governments have been affected by a developing awareness that world peace and prosperity depend on multinational and international cooperation. The League of Nations and the United Nations, together with their associated agencies, have represented major efforts to establish substitutes for traditional forms of diplomacy. Regional alliances and joint efforts, such as the Organization of American States, the North Atlantic Treaty Organization, the European Economic Community, and the Organization of African Unity, represent another type of cooperation among nations.

Multilateral organizations

Supervision and resolution of conflicts. The conflict of private interest is the leading characteristic of the political process in constitutional democracies, and the supervision, mediation, arbitration, and adjudication of such conflicts are among the key functions of their governments. Representative institutions are themselves a device for the resolution of conflict. Elections in constitutional democracies provide opportunities for mass participation in a process of open debate and public decision; assemblies, congresses, and other parliamentary institutions provide for public hearings on major issues of policy and require formal deliberative procedures at different stages of the legislative process; and political parties integrate a variety of interests and effect compromises on policy that win acceptance from many different groups.

If the interests that compete in the political process are too narrow or restricted, efforts may be made to control or change the rules of competition. Thus, laws have been enacted that seek to prevent discrimination from locking racial and other minorities out of the democratic process; the franchise has been extended to all groups, including minorities such as women, blacks, and 18-year-olds; and government bodies such as courts and administrative agencies enforce legislation against groups considered to be too large or monopolistic.

Judicial processes offer a means by which some disputes in society are settled according to rule and legal authority, rather than by political struggle. In all advanced societies, law is elaborated in complex codes governing rights and duties and procedural methods, and court systems are employed that adjudicate disputes in terms of the law. In constitutional systems such as the United States, the judiciary is deeply involved in the process of public decision making; the courts actually produce much of the substantive law that bears on private individuals and economic groups in society.

Regulation of the economy. Government regulation of economic life is not a new development. The national mercantilist systems of the 18th century provided for regulation of the production, distribution, and export of goods by government ministries; even during the 19th century, governments continued to intervene in the economy. The government of the United States, for example, from its inception in 1789, allotted funds or subsidies for the support of agriculture, maintained a system of tariffs for its own revenue and the support of domestic manufacturers, patronized the arts and sciences, and engaged in various kinds of public works to advance commerce and promote the general welfare. In France even more elaborate governmental schemes of economic regulation were practiced throughout the 19th century, including a variety of Socialist experiments such as the Public Workshops that Louis Blanc established in Paris in 1848. In Britain the various factory acts of the 19th century represented an effort by government to improve slightly the working conditions in industry.

Support of specific activities

After World War II the ability of a government to regulate or control the economy became one of the chief tests of its success, and regulatory agencies multiplied to the point at which they are now often referred to as "the fourth branch of government." The extent of the controls imposed on the economy is one of the principal distinctions among capitalist, Socialist, and Communist systems. In Communist countries it is a matter of doctrine that the means of production should be owned and therefore controlled by the state. In Britain the Labour governments nationalized some major industries, including coal, steel, and the railroads, prompted partly by Socialist doctrine and partly by the failure of British industry to remain competitive in international markets. This process was then reversed when the Conservative Party became ascendent. In the United States the government has involved itself in the economy primarily through its regulatory powers. In France the government has gone further and has engaged in national economic planning in cooperation with private business organizations.

The regulation of industrial conditions and of labour–management relations has been a major concern of most Western governments. In the United States the first regulatory efforts in this field were made during the Progressive era at the turn of the 20th century, when the wages, hours, and working conditions of women and children in industry became a matter of public scandal. A little later the conditions, hours, food, and wages of merchant seamen were brought under government regulation; an eight-hour day was set for railway crews; and workmen's compensation laws were instituted. With the Great Depression in the 1930s, minimum wages were introduced for workers in many industries, hours of work were set, and the right to collective bargaining was given legal sanction.

Regulation of transportation has been another major activity in most Western political systems, beginning with the railroads. In the United States their monopolistic practices attracted the criticism of agricultural interests and led eventually to the Interstate Commerce Act of 1887, which regulated railroad rates; subsequent legislation covered the hours, conditions, and wages of railroad employees, among other things. Other modes of land and air transportation have since been brought under regulatory controls implemented by government agencies.

In most European countries, facilities of communication—the telegraph, telephone, radio, and television—are owned and operated by the government. In the United States, most of these facilities have remained in private ownership, although they are regulated by the Federal Communications Commission. The regulation by government of important instruments of public opinion such as radio, television, and newspapers has important implications for the freedoms of speech and press and other individual rights. In the United States and Great Britain, government censorship of the press and other media has been restricted to matters of national security. This is also generally true of other Western constitutional democracies, although the celebrated *Der Spiegel* affair in Germany in 1962 and some extraordinary controls imposed on the media in France have been widely criticized. In many

Government role in communications

of the less-developed nations with authoritarian governments, very extensive controls are imposed on the press, and government-owned newspapers are often the principal sources of political news.

Other forms of government regulation of the economy involve the use of taxes and tariffs, the regulation of weights and measures, and the issuance of money.

Protection of political and social rights. To some extent, all modern governments assume responsibility for protecting the political and social rights of their citizens. The protection of individual rights has taken two principal forms: first, the protection of liberty in the face of governmental oppression; second, the protection of individual rights against hostile majorities and minorities. From the 1960s to the mid-1980s the sphere of public discussion in the Soviet Union was gradually, though erratically, widened. While this widening never extended to the purely political, some of the sociological discussion allowed served to set the intellectual stage for the Gorbachev period of radical reform. Although the Soviet Union suppressed most expression of political opinion until the late 1980s, domination by a central authority protected the basic human rights of some groups from violent rival minorities. On a larger scale, this is also what suppressed many long-standing rivalries in eastern Europe during the Cold War. The degree of repression in the former Communist states varied from country to country and changed with time after the death of Stalin.

Rights in the former U.S.S.R.

In the United States the Supreme Court expanded the rights of the criminally accused; and after 1954 the national government acted to bar legal discrimination against ethnic minorities. Indeed, almost all the freedoms detailed in the first 10 amendments to the Constitution have been extended since World War II.

Another type of government regulation bearing on the individual concerns the law of immigration and emigration. The great mass migrations of the 19th and early 20th centuries came to an abrupt halt after 1914 with the proliferation of government controls on the freedom of movement across national boundaries. After some later liberalization, immigration to the industrialized states again saw increased restrictions near the end of the 20th century.

Provision of goods and services. All modern governments participate directly in the economy, purchasing goods, operating industries, providing services, and promoting various economic activities. One of the indispensable functions of government—national defense—has made governments the most important consumers of goods, and they have not hesitated to use their resulting pricing, purchasing, and contracting powers to achieve various economic aims. In efforts to avoid dependence on private sources of strategic goods and defense materials, some governments have taken a further step and established their own military production plants. In wartime, governments have assumed control over entire industries and have subjected the work force to military direction in addition to rationing goods and regulating prices.

Economic impact of national defense

In nearly all political systems, certain functions are recognized as public, or belonging to the government. In addition to national defense, these include the maintenance of domestic peace, public education, fire protection, traffic control, aid to the indigent, conservation of natural resources, flood control, and postal services. But governments have assumed responsibility for many other commercial operations, even in non-Socialist countries. The sale of electric power, for example, is one of the established enterprising functions of national, state, and local governments in the United States and Canada. Municipally owned power utilities exist in about 2,000 cities in North America; 14 states of the U.S. have public power districts; and the U.S. government markets power through the Tennessee Valley Authority, the Bureau of Reclamation, and several other agencies. Another range of functions is performed by other national agencies, including the Rural Electrification Administration, which loans money to rural cooperatives to finance local power projects; the Export–Import Bank of Washington, which makes loans to finance export and import trade; the Maritime Administration, which holds mortgages on ships; and the Veterans Administration, which makes loans for farm or home purchases to military veterans. The United States government, acting through agencies such as the Social Security Board, the Federal Deposit Insurance Corporation, and the Federal Housing Administration, is also the largest insurer in the nation. Through other agencies, such as the Housing and Home Finance Agency, the Urban Renewal Administration, and the Public Works Administration, the national government has also developed a major role in the construction and rental of residential housing for low-income persons.

Other miscellaneous enterprises in which governments are involved include the provision of health care, the operation of public transport facilities, the development of public works, airport and port maintenance, and water-supply systems. In Great Britain the government operates hospitals and provides medical care under the National Health Service. In the United States many state and local governments operate hospitals on a commercial basis, although providing some charity care. At the local level in the United States the Port of New York Authority constructs and operates bridges, terminals, and airports. The states in the Delaware Basin have joined in a compact to establish an agency to control the use of water from the basin, institute programs to prevent pollution, provide recreation facilities, transmit and sell hydroelectric power, and provide watershed management. Cities in the United States and Canada operate more than 70 urban transit systems, 600 municipal gas utilities, and more than 4,000 water-supply systems. Cities are also generally responsible for garbage disposal and sometimes operate commercial slaughterhouses, coal yards, laundries, ice plants, and golf courses. Finally, packaged liquor sales, either wholesale or retail or both, are often made by the state governments.

PUBLIC ADMINISTRATION

While the functional objectives of government administration vary from system to system, all countries that are technologically developed have evolved systems of public administration. A number of common features may be detected in all such systems. The first is the hierarchical, or pyramidal, character of the organization by which a single chief executive oversees a few subordinates, who in turn oversee their chief subordinates, who are in turn responsible for overseeing other subordinates, and so on until a great structure of personnel is integrated and focussed on the components of a particular program. A second common feature is the division of labour or specialization within the organization. Each individual in the hierarchy has specialized responsibilities and tasks. A third feature is the maintenance of detailed official records and the existence of precise paper procedures through which the personnel of the system communicate with each other and with the public. Finally, tenure of office is also characteristic of all public bureaucracies.

Civil service

The various national civil services, despite their similarities, also show important differences, particularly in the way in which individuals are recruited and in the status accorded them in the political system. The British civil service, for example, has traditionally been composed of three classes, or grades—clerical, executive, and administrative. Administrative civil servants, the highest grade, are recruited by examination from among recent university graduates. The top managers of the different government ministries are drawn from this elite group. They remain in office despite changes in government and are accorded immense prestige. The U.S. civil service is organized into 18 grades. Although promotion from the lower grades is the typical means by which positions in the top grades are filled, there is also a flow of individuals into senior positions from private business and the professions. The U.S. equivalent of the administrative civil servant in Britain is usually a political appointee recruited by each new administration from private life or from a position in politics.

Development and change in political systems

The student of political systems grapples with a subject matter that is today in constant flux. He must deal not

only with the major processes of growth, decay, and breakdown but also with a ceaseless ferment of adaptation and adjustment. The magnitude and variety of the changes that occurred in the world's political systems between the second and eighth decades of the 20th century suggest the dimensions of the problem. Great empires disintegrated; nation-states emerged, flourished briefly, and then vanished; world wars twice transformed the international system; new ideologies swept the world and shook established groups from power; all but a few nations experienced at least one revolution and many nations two or more; domestic politics in every system were contorted by social strife and economic crisis; and everywhere the nature of political life was changed by novel forms of political activity, new means of mass communication, the enlargement of popular participation in politics, the rise of new political issues, the extension of the scope of governmental activity, the threat of nuclear war, and innumerable other social, economic, and technical developments.

CAUSES OF STABILITY AND INSTABILITY

Although it is possible to identify a number of factors that obviously have a great deal to do with contemporary development and change in the world's political systems—industrialization, population growth, the "revolution of rising expectations" in the less developed countries, and international tensions—there is no agreed theory to explain the causes of political change. Some social scientists have followed Aristotle's view that political instability is generally the result of a situation in which the distribution of wealth fails to correspond with the distribution of political power and have echoed his conclusion that the most stable type of political system is one based on a large middle class. Others have adopted Marxist theories of economic determinism that view all political change as the result of changes in the mode of production. Still others have focussed on governing elites and their composition and have seen in the alienation of the elite from the mass the prime cause of revolutions and other forms of violent political change.

In the discussion that follows, a distinction is drawn between unstable and stable political systems, and an attempt is made to suggest ways of understanding the processes of political development and change.

Crisis situations

Unstable political systems. In modern times the great majority of the world's political systems have experienced one form or another of internal warfare leading to violent collapse of the governments in power. Certain crisis situations seem to increase the likelihood of this kind of breakdown. Wars and, more particularly, national military defeats have been decisive in prompting many revolutions. The Paris Commune of 1871, the Russian revolutions of 1905 and 1917, Hitler's overthrow of the Weimar Constitution in Germany, and the revolutions in China all occurred in the aftermath of national military disasters. Many factors in such a situation, including the cheapening of human life, the dislocation of population, the ready availability of arms, the disintegration of authority, the discrediting of the national leadership, material scarcities, and a sense of wounded national pride, contribute to the creation of an atmosphere in which radical political change and violent mass action are acceptable to large numbers of people. Economic crises are another common stimulus to revolutionary outbreaks, for they produce not only the obvious pressures of material scarcity and deprivation but also a threat to the individual's social position, a sense of insecurity and uncertainty as to the future, and an aggravation of the relationships among social classes. A severe national economic crisis works, in much the same way as a military disaster, to discredit the existent leadership and the present regime. Another triggering factor is the outbreak of revolutions in other political systems. Revolutions have a tendency to spread: the Spanish Revolution of 1820 had repercussions in Naples, Portugal, and Piedmont; the French Revolution of July 1830 provoked similar outbreaks in Poland and Belgium; the Russian Revolution of 1917 was followed by a dozen other revolutions; and the colonial liberation movements in Africa, Southeast Asia, and elsewhere after World War II appear to have involved a similar chain reaction.

Crisis situations test the stability of political systems in extremely revealing ways, for they place extraordinary demands on the political leadership and the structure and processes of the system. Since the quality of the political leadership is often decisive, those systems that provide methods of selecting able leaders and replacing them possess important advantages. Although leadership ability is not guaranteed by any method of selection, it is more likely to be found where there is free competition for leadership positions. The availability of established methods of replacing leaders is equally, if not more, important, for the result of crises is often to disgrace the leaders in power, and, if they cannot be replaced easily, their continued incumbency may discredit the whole regime. The stamina and resolve of the ruling elite are also important. It is often said that a united elite, firmly believing in the justice of its own cause and determined to employ every measure to maintain its power, will not be overthrown. Most revolutions have gotten under way not when the oppression was greatest but only after the government had lost confidence in its own cause.

Other conditions of the survival of political systems relate to the effectiveness of the structures and processes of government in meeting the demands placed on them. Political systems suffer violent breakdown when channels of communication fail to function effectively, when institutional structures and processes fail to resolve conflicts among demands and to implement acceptable policies, and when the system ceases to be viewed as responsive by the individual and groups making demands on it. Usually, a system has failed over a period of some time to satisfy persistent and widespread demands; then, exposed to the additional strains of a crisis situation, it is unable to maintain itself. Revolutions and other forms of violent collapse are thus rarely sudden catastrophes but rather the result of a process of considerable duration that comes to its climax when the system is most vulnerable.

Unstable political systems are those that prove vulnerable to crisis pressures and that break down into various forms of internal warfare. The fundamental causes of such failures appear to be the lack of a widespread sense of the legitimacy of state authority and the absence of some general agreement on appropriate forms of political action. Governments suffer their gravest handicap when they must govern without consent or when the legitimacy of the regime is widely questioned. This is often the case in systems that have experienced prolonged civil war, that are torn by tensions among different national or ethnic groups, or in which there are divisions along sharply drawn ideological or class lines. The problem is often most acute where there is a pretender to the throne, a government in exile, a neighbouring state sympathetic to a rebel cause, or some other focus for the loyalty of dissidents. To some degree, also, the problem of legitimacy confronts all newly established regimes. Many of the new states of Africa and Asia, for example, have found it a source of great difficulty. Often they have emulated the form of Western institutions but failed to achieve their spirit: borrowing eclectically from Western political philosophies and systems of law, they have created constitutional frameworks and institutional structures that lack meaning to their citizens and that fail to generate loyalty or a sense that government exercises rightful powers.

Causes of the breakdown of political systems

Closely related to the problem of legitimacy as a cause of the breakdown of political systems is the absence of a fundamental consensus on what is appropriate political behaviour. A regime is fortunate if there are well-established, open channels of political action and settled procedures for resolving grievances. Although the importance of such "rules of the game" is that they allow change to occur in mainly peaceful ways, stable political systems often show surprising tolerance for potentially violent forms of political behaviour, such as strikes, boycotts, and mass demonstrations. Such forms of political behaviour are not permitted in systems where there are no agreed limits to the role of violence and where there is a high risk that violence may escalate to the point of actual warfare. If the government cannot count upon widespread support for

peaceful political procedures, it must restrict many kinds of political action. Such restriction, of course, inhibits still more the development of open methods of citizen participation in politics and adds to tension between the government and the people.

Managed and evolutionary systems

Stable political systems. The simplest definition of a stable political system is one that survives through crises without internal warfare. Several types of political systems have done so, including despotic monarchies, militarist regimes, and other authoritarian and totalitarian systems. After 1868, in the period of the restoration regime under the Meiji emperor, Japan succeeded, without major political breakdowns, in building an industrial state and developing commercial structures that transformed traditional Japanese society. This achievement was based on the development of centralized patterns of political control and the growth of a type of authoritarianism involving the rule of a military elite. Similarly, some of the totalitarian regimes of the contemporary world have demonstrated an impressive capability for survival. The key to their success is their ability to control social development, to manage and prevent change, and to bring under governmental direction all the forces that may result in innovations that are threatening to the system.

In some systems, survival does not depend on the detailed management of the society or close governmental control over social processes but is the result of sensitive political response to the forces of change, of flexible adjustment of the structures of the system to meet the pressures of innovation, and of open political processes that allow gradual and orderly development. Much of the Western democratic world has achieved peaceful progress in this way, despite new political philosophies, population increases, industrial and technological innovations, and many other social and economic stresses.

Such evolutionary change is possible when representative institutions provide effective channels for the communication of demands and criticisms to governments that rely upon majority support. The election of legislators and executive officials, competition between political parties, constitutional guarantees of freedom of speech and press, the right of petition, and many other structures and procedures perform this function in contemporary constitutional democracies. In such systems, social and economic problems are quickly transformed into issues in the open arenas of politics; governments are obliged to shape policies that reflect a variety of pressures and effect compromises among many conflicting demands.

The representative mechanisms that have produced evolutionary change in Western constitutional democracies are themselves subject to a continuous process of adjustment and mutation. Indeed, representative institutions must develop in ways that reflect social and economic developments in the society or they will lose their legitimacy in the minds of the people. In political systems such as the United States, for example, subtle shifts in the function and relative power of different institutions are continuously being made and, over time, produce entirely new structures and very different patterns of institutional behaviour. It is as a result of this process that the presidency has accumulated a range of new powers that have given it primacy among the branches of American government. This process also explains the growth of administrative agencies that perform both legislative and judicial functions. This process of dynamic adjustment is crucial, for institutions that remain static in a changing society are unable to serve as agencies of evolutionary change.

TYPES OF POLITICAL CHANGE

The study of political change is difficult, for change occurs in many different ways and at many different points in the political system. One may distinguish several major types of change.

Radical revolution. First are changes of the most fundamental type—transformations not only of the structure of government but of the whole polity. Such change is not limited to political life but transforms also the social order, the moral basis, and the values of the whole society. Drastic change of this kind occurred in the four great revolutions of the modern era—the English Revolution of the 17th century, the American Revolution, the French Revolution, and the Russian Revolution. These movements had the most profound effect on social and political life, permanently altering the beliefs by which men live. Their consequences were felt not only in the societies in which they occurred but also in many other political systems, in which, as a result of their example, revolutions of an equally fundamental character occurred.

Each of these major revolutions was something of a world revolution, for it resulted in a basic change in the ways in which men in all political systems viewed the nature of politics and the purpose of political life. The independence movements in the colonial empires after World War II, for example, were fuelled by those principles of individual liberty and representative government that were once the slogans of 18th-century American and French revolutionaries. Marxist revolutionary concepts emphasizing economic progress and radical social change have shaped the development of many of the new nations. The continuing impact of such ideas is an example of another way in which fundamental political change occurs. The nature of a political system may be transformed not suddenly or violently in the course of revolution but by the gradual, corrosive influence of ideas and by the accumulating impact of different political philosophies.

Structural revision. A second type of change involves alterations to the structure of the political system. Such change is not fundamental, in the sense of a basic transformation of the nature of the regime, but it may produce great shifts in policy and other political outcomes. Because the structure of a political system—that is, its formal and informal institutional arrangements—is a major determinant of policy outcomes, it is frequently the target of political action of various kinds. The political activist, the reformer, and the revolutionary share the recognition that the policies of a government may be effectively changed by adjusting the institutional forms through which the government acts. In some systems, structural change has been accomplished by legal means. In the United States, for example, such major institutional reforms as the direct election of the Senate and the limitation on presidential terms were made by constitutional amendment; and in Britain the various reforms of Parliament were accomplished by statute. In other systems, structural changes are often achieved by revolution and other violence.

Change of leaders. A third type of political change involves the replacement of leaders. Again, the recognition that to change the personnel of a government may be an effective way of changing government policy prompts many kinds of political action, ranging from election contests to political assassination and various forms of coup d'etat. In some systems the existence of established means of changing political leaderships works to prevent violent types of political action. In the United States the quadrennial contests for the presidency afford a constitutional opportunity to throw the whole executive leadership out of office. At the other extreme, the coup d'etat leads to the abrupt, often violent replacement of national executives. Although it is a type of revolution, the coup d'etat usually does not involve prolonged struggle or popular participation; after seizing office, the principal aim of the leaders of the coup is usually the restoration of public order. The coup d'etat occasionally develops into much more than the replacement of one set of governmental leaders by another and may prove to be the initial stage of a truly revolutionary process; *e.g.*, the coups d'etat that initiated Communist rule in Czechoslovakia in 1948 and ended King Farouk I's regime in Egypt in 1952.

Change of policies. Government policy itself may be an important agency of political change. The social and economic policies of Franklin D. Roosevelt's New Deal and the Socialist programs implemented by the British Labour Party after 1945 are examples. In both cases, government policies resulted in far-reaching modifications to the functioning of the political system: a vast expansion in the role of government in the economy, the use of taxes to redistribute wealth, an increase in the political influence of organized labour, and the implementation of

national programs of social welfare. Major policy change of this type, of course, is often a response to widespread pressures and demands that, if not satisfied by the system, may intensify and lead to various forms of violent political action. At other times, however, policy changes are imposed by a government to achieve the political, social, and economic goals of a single class, of an elite, or of the political leadership itself.

Many important questions remain as to the reasons for change, the ways in which change occurs, and the effects of change. Political scientists are still not completely certain, for example, why some systems have managed to avoid violent political change for considerable periods, while in other systems change is typically accomplished through coups d'etat, revolutions, and other forms of internal warfare. As suggested above, the explanation may have much to do with the existence in countries such as the United States and Great Britain of well-established political institutions that permit peaceful change, the presence in the population of widely shared attitudes toward the government, and the existence of basic agreement on the legitimacy of state authority. Clearly, however, other factors are also involved. Perhaps one of the chief goals of the study of political systems should be to determine as exactly as possible the conditions and prerequisites of those forms of change that permit the peaceful and evolutionary development of human society. (D.A.He./Ed.)

ELECTORAL PROCESSES

Elections provide a means of making political choices by voting. They are used in the selection of leaders and in the determination of issues. This conception of elections implies that the voters are presented with alternatives, that they can choose among a number of proposals designed to settle an issue of public concern. The presence of alternatives is a necessary condition, for, although electoral forms may be employed to demonstrate popular support for incumbent leaders and their policies, the absence of alternatives disqualifies such devices as genuine elections.

The widespread use of elections in the modern world is due in large part to the gradual emergence of representative government from the 17th century on. For proper appraisal, however, it is important to distinguish between the form and substance of elections. Electoral forms may be present but the substance may be missing. The substance is, of course, that the voter has a free and genuine choice between at least two alternatives. In a purely formal sense, the great majority of the more than 150 contemporary nations have what are called "elections," but probably only about a third of these have more or less competitive elections; perhaps a fifth have one-party elections; and in some others the electoral situation may be highly ambiguous.

The discovery of the individual as the unit to be counted was, from the 17th century on, the critical factor in the emergence of modern electoral processes. The counting of individuals, in turn, was a by-product of the change from the holistic conception of representation in the Middle Ages to an individualistic conception. The British Parliament, for instance, was seen no longer as representing estates, corporations, and vested interests but rather as standing for actual human beings. The movement abolishing "rotten boroughs"—boroughs of small population controlled by a single person or family—that culminated in the Reform Act of 1832 was a direct consequence of the individualistic conception of representation. Once governments were not only believed to derive their powers from the consent of the governed but were expected to seek consent, the only remaining problem was to decide who was to be included among the governed whose consent was to be sought. The democratic answer was, of course, universal adult suffrage.

Although it is common to equate representative government, elections, and democracy and although competitive elections under universal suffrage are in many respects a defining characteristic of political democracy, universal suffrage is not a necessary condition of competitive electoral politics. An electorate may be limited by formal legal requirements—as was the case before universal adult suffrage—or it may be limited by citizens' failure to take advantage of the vote, as is often the case in American municipal and other elections. Although such legal or self-imposed exclusion affects the democratic quality of elections and may ultimately affect the legitimacy of government, it does not impede decision making by election, provided the voter is presented with alternatives among which to choose.

Access to the political arena during the 18th century depended largely on membership in some aristocracy, and participation in elections was regulated mainly by local customs and arrangements. With the American and French revolutions, every citizen was declared formally equal to every other citizen, but the vote remained an instrument of political power possessed by very few.

Even with the arrival of universal suffrage, the ideal of "one man, one vote" was not achieved. Systems of plural voting were maintained in some countries, giving certain social groups an electoral advantage. In Great Britain, for example, university graduates and owners of businesses in constituencies other than those in which they lived continued to have an extra vote until 1948. Before World War I both Austria and Prussia had three classes of weighted votes that effectively kept electoral power in the hands of the upper social strata.

Whereas in the Western nations of the 19th and 20th centuries the increasing use of competitive mass elections in selecting governments had the purpose and the effect of institutionalizing the diversity of modern societies, in the Eastern, one-party, Communist regimes mass elections came to have quite different purposes and consequences. They differed from competitive elections in that each voter usually had only the choice of voting for or against the official candidate. Indeed, they were in the nature of the 19th-century Napoleonic plebiscites, in that they were intended to demonstrate the unity rather than the diversity of the people.

Elections in authoritarian systems

Dissent could be registered by crossing out the name of the candidate on the ballot, as several million Soviet citizens did in each election before 1989. As voting was not private, however, this invited reprisals. It may well be that some portion of dissenting votes were cast not so much because of dislike of Communism but because of grievances involving the conduct of minor officials. This may have served to weed out some of the worst officials at the very lowest levels. Even not voting was a form of protest, especially because local Communist Party activists were under extreme pressure to get nearly a 100 percent turnout. Before the revolutions of 1989, not all elections in eastern Europe followed the Soviet model exactly. In Poland, for instance, more names appeared on the ballot than there were offices to fill, and some degree of electoral choice was possible.

Many authoritarian regimes throughout the world have attempted to gain some level of legitimacy through the holding of elections. This may be done when it is clear that, because of repression, no substantive opposition is remotely feasible. Often, however, the process is more subtle in order to maximize the regime's gains. Elections may be scheduled when economic factors favour the regime and, more importantly, when election laws have been written to the severe disadvantage of competing parties. The opposition may be given little time to prepare, while the government already has various networks of supporters in place. Challengers also may be forced to campaign in an atmosphere of intimidation that precludes the effective organization of many potential supporters. A regime may cite unrelated reasons for postponing an election if it perceives a significant chance of losing. Also, it is not uncommon for government intervention to occur once balloting has begun, either in the form of voter intimidation (not infrequently actually attacking voters) or manipulating the count of votes freely cast.

Functions of elections. Fundamental to the use of elections is the contribution that they make to democratic government. Where the members of the body politic cannot themselves govern and must entrust government to representatives, elections serve not only to select leaders

acceptable to the voters but also to hold the leaders accountable for their performance in office. Accountability, however, is greatly jeopardized in electoral situations in which elected leaders, for want of ambition, do not care whether or not they are re-elected or in situations in which, for historical or other reasons, one party is so predominant as to preclude effective choice among alternate candidates or policies. Nevertheless, the possibility of controlling leaders by requiring them to submit to regular and periodic elections contributes to solving the problem of succession in leadership and, thereby, to the continuation of democracy. Moreover, where the electoral process is competitive and forces candidates or parties to expose their record of accomplishment and future intentions to popular scrutiny in election campaigns, elections serve as forums for the discussion of public issues, facilitate the expression of public opinion, and, more generally, permit an exchange of influence between governors and governed.

Elections also serve to reinforce the stability and legitimacy of the political community in which they take place. Like national holidays commemorating common experiences, elections serve to link the members of a body politic to each other and thereby confirm the viability of the political community. By mobilizing masses of voters in a common act of governance, elections lend authority and legitimacy to the acts of those who wield power in the name of the people.

Elections can also confirm the worth and dignity of the individual citizen as a human being. Whatever other needs he may have, participation in an election serves to gratify the voter's sense of self-esteem and self-respect. It gives him an opportunity to have his say, and he can, through expressing partisanship and even through nonvoting, satisfy his sense of belonging to or alienation from the political community. It is for just this reason that the age-long battle for the right to vote and the demand for equality in electoral participation can be seen as the manifestation of a profound human craving for personal fulfillment.

In all forms of government, from the most democratic to the most totalitarian alike, elections have a ritualistic aspect. Elections and the campaigns preceding them are dramatic events which, depending on cultural or historical circumstances, may exude the gay atmosphere of a circus or the sombre atmosphere of a funeral. Rallies, banners, posters, buttons, headlines, and television call attention to the importance of participation in the event. Candidates and parties, from right to left, in addition to propagating their policy objectives through rhetoric and slogans, invoke the symbols of nationalism or patriotism, reform or revolution, past glory or future promise. Whatever the peculiar national, regional, or local variations, elections are events that, by arousing emotions and channelling them toward collective symbols, break the monotony of daily life and focus attention on the common fate.

Systems of counting votes. Individual votes are totalled into collective decisions by a wide variety of rules of counting that voters and leaders have accepted as legitimate prior to the election. These decision rules may call for plurality voting, which requires that, among three or more alternatives, the winner need have only the highest number of votes; simple majority voting, which requires that the winner receive more than 50 percent of the vote; extraordinary majority voting, which requires some higher proportion for the winner, such as a two-thirds vote; or unanimity.

Plurality voting

Plurality and majority decision. The simplest means of deciding an election is the plurality rule. To win, a candidate need only poll more votes than any other single opponent; he need not, as required by the majority formula, poll more votes than the combined opposition. The more candidates contesting a constituency seat, the greater the probability that the winning candidate will receive less than 50 percent of the vote. The plurality formula is used in the national elections of such countries as Great Britain, Canada, and New Zealand. One-half of the former West German Bundestag was elected according to the plurality system, while the other half was elected according to a proportional representation formula (see below).

Under the majority rule, the party or candidate winning more than 50 percent of the vote in a constituency is awarded the contested seat. The winning party or candidate must poll more votes than the combined opposition. This system thus ensures that the elected representative has the support of the majority of the voters. The majority formula is employed in Australia and France. It is usually applied only within single-member electoral constituencies.

Both the plurality and the majority-decision rules are employed in the election of U.S. presidents. The composition of the electoral college, which actually elects the president, is determined by a plurality vote taken within each state. Voters choose between the names of the presidential candidates, but they are in effect choosing the electors who will elect the president by means of a majority vote in the electoral college. All of a state's electoral votes (which are equal in number to its seats in Congress) are given to the presidential candidate who gains a plurality of the vote in the state election. It is thus possible for a president to be elected on the basis of a minority of the popular vote.

A critical difficulty with the majority formula is that, in a multiparty political system, the formula may produce an electoral deadlock if no candidate secures 50 percent of the total vote. In order to break such deadlocks a second round of elections (second ballot) is required, if no candidate obtains a majority on the first round. In Australia, voters rank the candidates on an alternate preference ballot. If a majority is not achieved on the first round of elections, the weakest candidate is eliminated, and the votes he accrued are distributed to the other candidates according to the second preference on the ballot. This redistributive process is repeated until one candidate collects a majority of the votes. If no candidate secures a majority in the first round of the French National Assembly elections, another round of elections is required. In this second round, the candidate securing a plurality of the popular vote is declared winner.

Neither the majority nor the plurality formulas distribute legislative seats in proportion to the share of the popular vote won by the competing parties. Both formulas tend to award the strongest party disproportionately and to handicap the weaker parties. The latter is particularly true if small parties are rooted in ethnic, religious, or social minorities; small parties escape the inequities of the electoral system only if they have a regionally concentrated base. The plurality formula distorts the distribution of seats more than the majority system.

Advantage of the strongest party

Proportional representation. Proportional representation requires that the distribution of offices be proportional to the distribution of the popular vote among competing political parties or candidates. It seeks to overcome the distribution imbalances that result from majority and plurality formulas and to create a representative body that reflects the distribution of opinion within the electorate. Because of the use of multimember constituencies in proportional representation, parties with neither a majority nor plurality of the popular vote can still win legislative representation.

Proportional representation is an ideal that is sought after, but only approximated: the size of electoral districts is the critical factor; the larger the electoral district in terms of seats, the more proportional the representation will be. The number of seats assigned to a constituency is in fact a greater determinant of the proportionality of the outcome than is the specific type of proportional formula used. The different formulas of proportional representation are basically similar in their effect on the conversion of votes to political representation.

Developed in the 19th century in Denmark and in Britain, the single transferable vote formula—or Hare system (after one of its English developers, T. Hare)—employs a ballot that allows the voter to rank the competing candidates in order of preference. When the ballots are counted, any candidate receiving the necessary quota of first preference votes is awarded a seat. In the electoral calculations, votes received by a winning candidate in excess of the quota are transferred to other candidates according to the second preference marked on the ballot. Any candidate who then has the necessary quota is also

awarded a seat. This process is repeated, with subsequent surpluses also being transferred, until all the remaining seats have been awarded. Five-member constituencies are considered optimal for the working of the Hare system.

The single transferable vote formula, because it involves the aggregation of ranked preferences, necessitates complex electoral computations. This factor, plus the fact that the Hare system limits the influence of political parties, probably accounts for its infrequent use; it has been used in Northern Ireland, Ireland, and Malta in the selection of the Australian and South African senates. The characteristic of the Hare formula that distinguishes it from other proportional representation formulas is its emphasis on candidates, not parties. The party affiliation of the candidates has no bearing on the computations.

The basic difference between the transferable vote formula and the list systems is that, in the latter, voters choose among party-compiled lists of candidates rather than among individual candidates. Although voters may have some limited choice among individual candidates, electoral computations are on the basis of party affiliation; seats are awarded in respect to party rather than candidate totals. The seats that a party wins are allocated to its candidates in the order in which they appear on the party list. Several types of electoral formulas are used, although they fall into two main categories: largest average formulas and greatest remainder formulas. All employ some type of electoral quota.

Largest average formula

In the largest average formula, the number of votes won by each party is divided by the number of seats held by the party, plus one. The first seat is awarded to the party with the highest number of votes, since, no seats yet having been allocated, the initial denominator is one. When a party wins a seat, its formula denominator is increased by one and hence the party's chances of winning the next seat are reduced. The available seats are awarded one at a time to the party with the greatest average. Party totals, not candidate totals, are used in the calculations. No transfer of ballots takes place. Frequently named after its Belgian inventor, Victor d'Hondt, the largest average formula is used in Austria, Belgium, Finland, and Switzerland.

The d'Hondt formula has a slight tendency to overreward large parties and to reduce the chance of small parties gaining legislative representation. The so-called Lague variation of the d'Hondt formula—used in Denmark, Norway, and Sweden—reduces the reward to large parties but increases further the handicap to small parties. By adjusting the denominator of the d'Hondt formula, the Lague formula increases both the cost of the initial seat of a party and that of additional seats. The Lague formula thus aids middle-size political parties and reduces the number of legislatively represented small parties.

The greatest remainder method establishes a vote quota for each seat in an electoral district by dividing the total vote in the district by the number of competing parties. The total popular vote won by each party is then divided by the quota, and a seat is awarded as many times as the party total contains the full quota. If all the seats are awarded in this manner, the election is complete. Such an outcome is unlikely, however. Seats that were not won by full quotas are subsequently awarded to the parties with the largest remainder of votes after the quota has been subtracted from each party's total vote for each seat it was awarded. Seats are awarded sequentially to the parties with the largest remainder until all the district's allocated seats have been awarded.

Of all the proportional representation formulas, the greatest remainder formula, given large enough constituencies, yields results closest to the proportional ideal. Small parties fare better when the greatest remainder formula is used than when the largest average formula is employed. The greatest remainder formula is used in Israel, Italy, and Luxembourg and in some elections in Denmark. Italy, however, uses a special variety of the formula, called the Imperiali formula, whereby the electoral quota is established by dividing the total popular vote by the number of parties plus two. This modification increases the legislative representation of small parties but leads to a greater distortion of the proportional ideal.

The choice of majority and plurality or proportional systems is, of course, not simply a matter of pure theory. Different methods of counting, just as different conceptions of representation, usually reflect cultural, social-structural, and political circumstances in a particular jurisdiction. Majority or plural methods of voting are most likely to be acceptable in relatively stable political cultures. In such cultures, fluctuations in electoral support, given to one party or another from one election to the next, reduce polarization and make for political centrism. Thus the "winner take all" implications of the majority or plurality formulas are not experienced as unduly deprivational or restrictive. Proportional representation, on the other hand, is more likely to be found in societies with traditional ethnic, linguistic, and religious cleavages or in societies experiencing pervasive class and ideological conflicts.

Types of elections. *Election of officeholders.* Rather than deciding public policy, election to public office bestows the right to decide on a small group of elected officials. The elections grant officials the authority to make public policy decisions, thereby legitimizing the decisions that are made. The policy voice of the electorate in these elections may be weak. An electoral mandate results not so much in the implementation of the abstract preferences of the majority of the citizens but rather in the implementation of the policy of those leaders who come to power in elections.

Role of parties

Most officeholder elections are interpretable only in terms of the role of political parties in the political process. Parties are tied so closely to the democratic functions of elections that their competition for the electorate's support is considered the main criterion by which democratic elections are distinguished from nondemocratic elections. Because of the difficulty in fixing individual responsibility in the complexity and vastness of modern governments, political parties, rather than representatives, have become the vehicle of democratic accountability.

Political parties shape the nature of the electoral decision. The selection and nomination of candidates, a vital first stage of the electoral process, lies in the hands of the political parties; the election serves only as the final process in the recruitment to political office. The party system can thus be regarded in a true sense as an extension of the electoral process. Political parties provide the pool of talent from which candidates are drawn, and they simplify and direct the electoral choice and mobilize the electorate at the registration and election stage. Political parties have transformed elections from a choice of people to a choice of governments.

But the fusion of political parties with the electoral process has not gone unchallenged. In North America a number of modifications have been made in the electoral structure to limit the influence of political parties. Nonpartisan elections (elections in which party affiliations are not formally indicated on ballots) are the norm for some municipalities in the United States and Canada and represent an attempt to remove political parties from the electoral scene. Nonpartisanship started as a reform movement in the early part of the 20th century and was designed to put local government on a more businesslike basis and to isolate local politics from state and national politics. The exclusion of party labels from the ballot, however, does not guarantee that political parties will be inactive or without influence in nonpartisan elections.

Recall. The practice of recalling officeholders was, like most populist innovations, an attempt to minimize the influence of political parties on representatives. The recall was an electoral modification designed to ensure that an elected representative would act as a spokesman for his constituency rather than for his political party or the dictates of his own conscience or interests. It enabled constituents to recall their elected representative before the end of his term of office if, in their opinion, he was no longer acting in accord with the best interests of the constituency. The actual instrument of recall was often a letter of resignation signed by the elected representative before assuming office. During the term of office, this letter could be evoked by a quorum of constituents if the representative's legislative performance failed to meet their

expectations. In practice the recall is not often or widely used today, even where it is provided for constitutionally.

Referendum. A referendum is an election, held either at the initiative of the governors or the governed, in which the preferences of the community regarding a particular issue are assessed. The use of the referendum reflects a reluctance to entrust full decision-making power to elected representatives. Because of the small voting turnout for referendum elections, and the absence of affixed party labels to the referendum alternatives, voting in referendum elections is more open to leadership and interest group influence than is voting in officeholder elections.

Referenda usually concern the raising and spending of public money, although they are occasionally used to decide moral issues in which the elected bodies are deemed to possess no special competence. They may be legislatively binding, or merely consultative, but even consultative referenda are likely to be considered legislative mandates.

Plebiscite. Plebiscites are elections held to decide two paramount types of political issue: the form of government and the national boundaries adopted by the state. In the first case, the plebiscite is a means of establishing the legitimacy to rule, not a means of providing a choice among alternate governmental teams or legislative proposals. The incumbent government, seeking a popular mandate as a basis for legitimacy, employs a plebiscite to establish its right to speak for the nation. The incumbent rulers seek a single identity with the nation, and thus a vote for the nation is interpreted as a vote for the incumbent regime. Plebiscites of this nature are thought to establish a direct linkage between the rulers and the ruled; intermediaries such as political parties are bypassed in plebiscites, and in this respect they are antithetical to pluralism and competitive politics. Following the Revolution of 1789, the plebiscite had a great appeal in France, rooted as it was in the ideas of nationalism and popular sovereignty. In more recent times totalitarian regimes have employed plebiscites to legitimize their power.

Plebiscites have also been used as a device for establishing the boundaries of political units according to criteria of nationality. After World War I the plebiscite was employed as a means of fitting political boundaries to boundaries of nationality. Eleven plebiscites, based on universal suffrage, were proposed by the League of Nations, the most successful of which was held in 1935 in the Saar (which chose to remain a German territory rather than go to France). The potential of plebiscites for establishing or changing political boundaries is, however, severely limited, because they entail the prior agreement of two governments over what is usually a very contentious issue.

Constituencies: districting and apportionment. The drawing up of constituencies—those subdivisions of the total electorate that send representatives to the local or central assembly—is inextricably linked with questions concerning the nature of representation and methods of voting. Of these questions, that of representation is primary. The problem of electoral representation turns on the question of what is to be represented. In this connection the functions of constituencies and the problem of determining them (districting) become relevant. The central problem is that the constituency, as a geographical area, may contain within its boundaries diverse, and sometimes incompatible, social, economic, religious, or ethnic interests, all of which seek to be represented.

Districting

The solution to this problem has been largely historically determined. Where the interests of electors have not been totally incompatible, and where ethnic, religious, social, and economic differences have been relatively free of passionate conflict, as in the Anglo-American countries, geographical areas (electoral districts) have usually been considered the constituency, and the method of counting has been some system of majority or plurality voting. The elected person represents the geographical unit as a whole, irrespective of internal divisions. In general, the district is a single-member constituency, although multimember constituencies are also possible.

On the other hand, where the electorate is composed of several minorities, none of which can hope to obtain a majority, or perhaps even a plurality sufficiently large to obtain representation, the geographical district can only be regarded as an administrative unit for counting votes. The effective constituency is the group of electors that can be identified as having voted for a given candidate. In this case, election is usually by some method of proportional counting whereby any candidate, party, or group receiving a requisite number of votes is entitled to a proportionate number of representatives.

The drawing up, or delimitation, of electoral districts is inextricably linked with differing conceptions of representation; and conceptions of representation are inextricably linked with alternate methods of vote counting. The virtues and vices of different electoral arrangements are still debated by constitution makers, but the arguments advanced in favour of any one set of arrangements remain inconclusive. Although proportional systems treating electoral districts as merely administrative conveniences and providing for multiple representation are believed to approximate more closely the one man, one vote principle of democratic theory than do single-member geographic constituencies providing for majority or plurality voting, they are probably conducive to governmental instability, and they place an inordinately heavy burden of conflict resolution on representative bodies. This often makes for political stalemate or immobilism, as has been the case at times in France and Italy. But it is doubtful that in nations with many interests seeking representation any other method of electoral districting and vote counting would be satisfactory.

Representation of geographic constituencies with majority or plurality voting, on the other hand, disguises many differences in the electorate's composition and preferences. In single-member, majority-vote electoral districts, the minority losing at the polls may feel unrepresented. But it is often a precarious assumption that the majority itself is truly represented, since it is likely to be a loose coalition of diverse interests, and no single representative can usually do justice to the diversity of interests involved. In general, therefore, the representative, in so far as he is responsive, is more likely to play the role of broker among his district's diverse interests, rather than the role of spokesman for the district as a whole.

Whatever the role of representative, and however electoral units are conceived—as geographic areas with stable, compatible interests, or as purely administrative areas for the purpose of vote counting—districting must be distinguished from apportionment. In general, if the district is considered as only an administrative unit for counting votes, its boundaries can be drawn rather arbitrarily without injustice. The district as such has no stakes of its own to be represented. On the other hand, if the district is considered a true constituency, in the sense that it has unique interests that can be geographically and, as a result, socially defined, districting should not be arbitrary. The area involved should at least be contiguous and compact so that its presumed interests can be fairly recognized.

The difficulty in treating geographic areas as genuine constituencies is, of course, that internal transformations in their economy and social structure may make their historical boundaries obsolete. As a result, historical areas serving as electoral districts (such as provinces, states, or counties), which once served as genuine communities of interest to be represented, may no longer do so. As these districts become increasingly diverse, making it difficult to identify particular constituency interests, representatives find it increasingly impossible to be responsive to constituency wishes or needs.

Apportionment

Problems of apportionment, in contrast to problems of districting, stem from efforts to reconcile the territorial and population bases of representation. If geographic areas, for instance, are assumed to have an equal right to be represented because the area is considered to be a viable constituency, malapportionment in terms of population size is inevitable. New York with a population of nearly 20,000,000 has, in the U.S. Senate, the same number of representatives as Nevada with a population of less than 1,000,000. This kind of "constitutional malapportionment" must not be confused with the kind of "electoral malapportionment" that defies the one-man, one-vote

principle of equal representation. The latter is usually the result of population shifts. Under modern conditions of increasing urbanization, failure to reapportion the number of seats in representative bodies to take account of population changes has largely benefitted rural electoral districts. More recently, the shift of city people to the suburbs has led to possible underrepresentation of suburban vis-à-vis urban populations.

From a political rather than a legal perspective, malapportionment is usually considered "undemocratic" because it results in overrepresentation and underrepresentation of certain sectors of the population and, as a result, may culminate in public policies not acceptable to the true majority.

Apportionment is not a simple matter. In particular, the definition of the population among which a specified number of legislative seats is to be apportioned is not easy to arrive at. Any of several distinct quantities—total population, number of citizens of voting age, number of registered voters, or number of actual voters—may serve as bases of apportionment. If there were a constant relation between these groupings, apportionment would not be a problem, and any one of them could serve as a basis. In fact, however, the relation is not constant because of variations in registration and voting turnout.

Gerrymandering

Malapportionment, whether of the constitutional or electoral sort, must not be confused with gerrymandering—a form of arbitrary districting used to benefit the party that at a given time controls the apportionment process. Gerrymandering is named after governor of Massachusetts Elbridge Gerry (1744–1814), who recognized the possibility of influencing electoral outcomes by manipulating the boundaries of electoral districts. The practice involves concentrating large percentages of the opposite party's votes into a few districts, and drawing the boundaries of the other districts in such a way that the gerrymandering party wins in them all, even though the majority, or, in three-cornered elections, the plurality, is relatively small.

Gerrymandering comes to politicians easily. Party managers prefer "safe" districts to those that are competitive. By not contesting seriously in the other party's safe districts, by maintaining one's own safe districts, and by carving competitive districts up in such a way that one's own party is favoured, party managers use gerrymandering to create artificial district boundaries that, though advantageous to the parties, may harm other social, economic, and political interests that need representation.

Voting practices. There is a direct relation between the size of an electorate and the formalization and standardization of its voting practices. In very small voting groups, in which political encounters are face-to-face and the members are bound together by ties of friendship or common experience, voting is mostly informal and may not even require counting, because the "sense of the meeting" emerges from the group's deliberations. The consensus of the Quaker meeting is of this order. An issue is discussed until a solution emerges to which all participants can agree or, at least, from which any one participant will not dissent.

By way of contrast, in modern mass electorates, in which millions of individual votes are aggregated into the collective choice, formalization and standardization of voting practices and vote counting are the rule. This is necessary in order to guarantee that the outcome can be considered valid, reliable, and legitimate. Validity means that the collective choice in fact expresses the sense of the electorate. Reliability means that each vote is accurately recorded and effectively counted into the total. Legitimacy means that the criteria of validity and reliability have been met, so that the result of the voting is acceptable and provides authoritative guidelines in subsequent political conduct.

The development of routinized and standardized electoral practices in mass electorates is a surprisingly recent phenomenon, not much older than 100 years. It is as much a corollary of the growth of rapid communication through telephone and telegraph as of the growth of the electorate and rational insistence on making electoral processes fair and equitable. Nevertheless, even today electoral practices around the world differ a great deal, depending not just on formal institutional arrangements, but even more on a country's political culture.

Secret voting. Once suffrage rights had been extended to masses of voters who, in theory, were assumed to be equal but who, in fact, were unequal (in order of birth, in intellectual endowment and educational accomplishment, in social status and the possession of property, and so on), open voting was no longer tolerable precisely because it could and often did involve undue influence, ranging from hidden persuasion and bribery to intimidation, coercion, and punishment. Equality, at least in voting, was not something given but something that had to be engineered; the secrecy of the vote was a first and necessary administrative step toward the one man, one vote principle. Equality in voting was possible only if each vote was formally independent of every other vote, and this suggested the need for strict secrecy.

The slow progress made in introducing secret voting, from the French Revolution well into the 20th century, attests to the fact that social engineering, no matter how desirable or lofty in purpose, depends on favourable conditions for success. One need not assume that the obstacles placed in the way of secret voting were the result of some conspiracy on the part of those who recognized, quite accurately, that the mobilization of large numbers of voters fundamentally changed the distribution of political power and who, through opposing the secret vote, hoped to maintain a stranglehold on the newly enfranchised electorates. Rather, the success of secret voting depended on the reduction of illiteracy and, at the cultural level, on the spread of the individualistic norms of privacy and anonymity to certain classes of the population, notably peasants and workers. Traditionally these groups took their cues from those accepted as superior in status, or from their peers. Secret voting required learning to free oneself as a citizen from one's customary associations and from pressures for conformity. The difficulty of introducing and practicing the secret vote in today's politically and economically less developed nations mirrors the tortuous advance of the secret ballot in the Western nations during the 19th and early 20th centuries.

Benefits of secret voting

Secret voting reduces drastically the possibility of undue influence on the voter. Without it, influence can range from outright purchase of votes to social chastisement or economic sanctions. This is not to say that bribery in voting is automatically eliminated by secret voting. Laws prohibiting and punishing the purchase of votes are on the statute books of many countries and undoubtedly contribute to discouraging the practice.

Although informal social pressures on the voter are probably unavoidable and, in some respects, useful in reducing political rootlessness and contributing to political stability, secrecy in voting permits voters to break away from their social moorings and gives them a considerable degree of independence if they wish to take advantage of this electoral freedom. As a result it becomes ever more difficult for interest groups, whether labour unions, farmers' organizations, commercial or industrial associations, ethnic leadership groups, or even criminal syndicates, to "deliver the vote." The extent to which "deviant voting" occurs depends partly on the degree of rigidity in the social structure. In countries where caste or class barriers are high or where traditional social, economic, religious, or regional cleavages remain strong, deviant voting is less likely than in countries where social mobility is possible and where political conflicts cut across traditional social cleavages. In Western nations the increasing difficulties of labour, farm, religious, or ethnic leaders in influencing the voter attests to the success that secret voting seems to have had on freeing the individual from electoral bondage to his traditional or economic affiliations.

Balloting. The ballot makes secret voting possible. Its initial use seems to have been a means to reduce irregularities and deception in elections. This objective, however, could be achieved only if the ballot was not supplied by the voter himself, as was the case in much early voting by secret ballot, or by political parties, as is still the case in some countries. Ballot procedures differ widely, ranging from marking the names of preferred candidates to

crossing out those not preferred or writing in the names of persons who are not formal candidates. Ballots also differ according to the type of voting system employed. Where plurality or majority voting is practiced, most elections employ classified ballots whereby the voter casts his vote for only one candidate or list of candidates. Where proportional methods are used, the ballot enables the voter to rank the candidates according to his preferences.

It is generally believed that the nature of the ballot influences a voter's choice. In jurisdictions where electors are called upon to vote not only for higher offices but also for a multitude of local positions and where the election may include propositions in the nature of referenda, the length of the ballot can affect the results. Overwhelmed by a ballot's length, voters may be discouraged from expressing their preferences for candidates of whom they have not heard, or from deciding on propositions that they do not understand. For example, in some U.S. jurisdictions participation is often lower in contests for county coroners, tax assessors, and other local positions for which there is often little coverage in the media. Election data show a rapid decline from votes cast for higher offices to those cast for lower offices and referendum-type propositions, a phenomenon referred to as ballot roll-off.

Ballot roll-off

Ballot position also has an effect on the votes cast for particular candidates, especially in the absence of cues as to party affiliation or other identifications. The first position on the ballot may be favoured, and on long ballots both first and last names may benefit, with candidates in the middle of the ballot suffering slightly. Ballot position is likely to have its greatest impact in nonpartisan elections, primaries, and elections for minor offices.

The manner in which candidates are listed—by party column or by office bloc—is likely to affect election outcomes. On party-column ballots it is possible to vote a "straight ticket" for all of a party's candidates by entering a single mark, although voting for individual candidates is usually possible. Conversely, on the office-bloc ballot, voters choose individual candidates grouped by office rather than party, which discourages voting exclusively for members of one party, though some jurisdictions that use the office-bloc ballot allow voters to cast a straight ticket. The type of ballot can have important consequences for the structure of government, especially in systems with separated powers and federal territorial organization. If different offices are controlled by different parties, the governmental process may be marked by greater conflict than would otherwise be the case.

The introduction of voting machines and computer technology has not substantially changed the balloting process, though it generally has made it faster and more economical. Voting machines are not without problems, in that they may marginally depress the level of voting owing to improper use, a problem that can be overcome through improved machines and voter education. Indeed, analysis in the U.S. state of Florida found that a controversial ballot design in one county and improperly punched ballots across the state may well have been critical to the overall outcome of the national presidential election in 2000. The results, which highlighted the problems associated with various voting methods and the disparate distribution of voting technologies according to socioeconomic status (*i.e.*, wealthier areas generally had more advanced technology and fewer invalidated ballots), spurred extensive debate about election reform.

Electoral abuses. Corrupt electoral practices are not limited to bribery or voter intimidation. They include disseminating scurrilous rumours and false campaign propaganda, tampering with election machinery by stuffing ballot boxes with fraudulent returns, counting or reporting the vote dishonestly, and disregarding electoral outcomes by incumbent officeholders (*e.g.*, by mobilizing the military to thwart an election loss). The existence of these practices depends more on a population's adherence to political civility and the democratic ethos than on legal prohibitions and sanctions.

The integrity of the electoral process can be maintained by a variety of devices and practices, including a permanent and up-to-date register of voters and procedures designed to make the registration process as simple as possible. In most jurisdictions elections are held on a single day rather than on staggered days. Polling hours in all localities are generally the same, and opening and closing hours are fixed and announced, so that voters have an equal opportunity to participate. Polling stations are operated by presumably disinterested government officials or polling clerks under governmental supervision. Political party agents or party workers are given an opportunity to observe the polling process, enabling them to challenge irregularities and prevent abuses. Efforts are made to maintain order in polling stations, directly through police protection or indirectly through such practices as closing bars and liquor stores. The act of voting itself takes place in voting booths to protect privacy. Votes are counted and often recounted by tellers, who are watched by party workers to assure an honest count. The transmission of voting results from local polling stations to central election headquarters is safeguarded and checked.

Compulsory voting. In some countries, notably Australia and Belgium, electoral participation is legally required, and nonvoters can face fines. The concept of compulsory voting reflects a strain in democratic theory in which voting is considered not merely a right but a duty. Its purpose is to ensure the electoral equality of all social groups. However, whether created through laws or through social pressure, it is doubtful that high voter turnout is a good indication of an electorate's capability for intelligent social choice. On the other hand, high rates of abstention or differential rates of abstention by different social classes are not necessarily signs of satisfaction with governmental processes and policies and in fact may indicate the contrary.

Participation in elections. The electoral participation rate depends on many factors, including the type of electoral system, the social groupings to which voters belong, the voters' personalities and beliefs, their place of residence, and a host of other idiosyncratic factors.

The level and type of election has a great impact on the rate of electoral participation. Electoral turnout is greater in national than in state or provincial elections, and greater in the latter than in local elections. If local elections are held concurrently with provincial or national elections, generally a higher voter turnout is achieved than for nonconcurrent elections. Whether an election is partisan or nonpartisan also affects turnout, with fewer people participating in nonpartisan elections. Supporters of political parties vote more often than those without a partisan identification. Participation is also generally greater in candidate elections than in noncandidate elections such as referenda. There is evidence that elections based on proportional representation have higher electoral turnouts than majority or plurality elections. Noncompetitive or safe electoral districts tend to depress the level of voter turnout, whereas marginal elections increase it. The perceived closeness of an electoral contest and the degree of ideological polarization between parties or candidates can affect the competitiveness of the election and consequently its turnout. The frequency of elections is also related to voter participation, with fewer people generally participating in countries where elections are more frequent.

Effects of level and type of election

Technicalities in electoral law may disenfranchise many potential voters. For example, people who change their legal residence may temporarily lose their vote because of residence requirements for voters in their new electoral district. Complicated voter registration procedures, combined with a high level of geographical mobility, significantly reduce the size of the active electorate in the United States, whereas in many other countries the size of the electorate is maximized by government-initiated registration immediately prior to an election. Voter registration in the United States is largely left to the initiative of individuals and political parties, though the implementation of "motor-voter laws" during the 1990s attempted to increase voter registration by allowing a citizen to register to vote when he received or renewed his driver's license.

Voter registration requirements

Relatively low levels of electoral participation are associated with low levels of education, occupational status, and income. Those groups in society that have been most recently enfranchised also tend to vote at lower rates. For a

significant period of time in the 20th century, women voted less frequently than men, though the difference had been erased by the end of the century in most countries. The rates of participation of racial minorities are generally lower than those of majority groups, and members of the working class vote less frequently than members of the middle class. In many countries, participation by young people is significantly lower than that of older voters. The failure of certain types of people to vote in elections has important implications; most analyses have found that if all eligible voters cast ballots, the balance of electoral power would favour the recently enfranchised and less privileged members of society.

A small group of people are conscientious nonvoters. Others, perceiving the vote more as an instrument of censure than of support, may not vote because they are satisfied with the current government. This group of voluntary nonvoters is also small, however. In fact, nonvoters generally are less satisfied with the political status quo than are voters. The vote is a rather blunt and ineffectual instrument for expressing dissatisfaction, and nonvoting is more likely to be symptomatic of alienation from the political system, than of satisfaction with it.

A number of random factors influence individual participation in specific elections. Election campaigns vary in their intensity. A crisis atmosphere may induce a large number of people to vote on one occasion, whereas on another the chance to vote for an extremist candidate may increase the participation of the normally disinterested. Even the weather can affect election turnout.

National differences in voter participation rates

Voter participation varies from country to country. For example, approximately half of the voting age population participates in presidential elections in the United States. In contrast, many European countries have participation rates exceeding 80 percent. Even within Europe, however, participation varies significantly. For example, post-World War II Italy has averaged around 90 percent, whereas fewer than 40 percent of the electorate participates in elections in Switzerland. Research has suggested a long-term decline in turnout at national elections in western democracies since the 1970s; it seems most likely that this is a consequence of partisan dealignment (*i.e.*, a weakening of partisan identification), the erosion of social cleavages based on class and religion, and increasing voter discontent.

Influences on voting behaviour. The electoral choices of voters are influenced by a range of factors, especially social-group identity, which helps to forge enduring partisan identification. In addition, voters are to a greater or lesser extent susceptible to the influence of more short-term and contingent factors such as campaign events, issues, and candidate appeals. In particular, the perceived governing competence of candidates and political parties often weighs heavily on voters' choices.

Research suggests that partisan dealignment has reduced the proportion of voters in western democracies who retain their long-term partisan identities. In conjunction with the declining impact of social-group influences, voter choice is now more heavily affected by short-term factors relevant to specific election campaigns. This shift from long-term predisposition to short-term evaluation has been facilitated in part by the phenomenon of "cognitive mobilization," a supposed enhancement of the political independence and intelligence of voters who are both better educated and better informed than earlier generations. Nevertheless, many independents and nonvoters are poorly informed politically and relatively uninterested and uninvolved in politics. Whether cognitively mobilized or not, however, independent voters are often a decisive factor in elections.

(H.E./R.G./P.D.W./Ed.)

BIBLIOGRAPHY

General works. Perhaps the single best introduction to the subjects covered in this article is MARY HAWKESWORTH and MAURICE KOGAN (eds.), *The Encyclopedia of Government and Politics*, 2 vol. (1992). Overviews of various political systems can be found in CHARLES HAUSS, *Comparative Politics: Domestic Responses to Global Challenges*, 3rd ed. (2000); and ANTON BEBLER and JIM SEROKA (eds.), *Contemporary Political Systems* (1990). Other excellent general works include ROBERT A. DAHL, *Modern Political Analysis*, 5th ed. (1991). GABRIEL A. ALMOND and G. BINGHAM POWELL, JR. (eds.), *Comparative Politics Today: A World View*, 5th ed. (1992), is also of interest. Country-specific data can be found in J. DENIS DERBYSHIRE and IAN DERBYSHIRE, *Political Systems of the World*, 2nd ed. (1996).

Types of political systems. Totalitarianism is addressed in HANNAH ARENDT, *The Burden of Our Time* (1951; also published as *The Origins of Totalitarianism*, new ed., 1973, reprinted 1986); and ELLEN FRANKEL PAUL (ed.), *Totalitarianism at the Crossroads* (1990). Fascism is addressed in ERNST NOLTE, *Three Faces of Fascism*, trans. by Leila Vennewitz (1965; originally published in German, 1963).

Systems theory. DAVID EASTON, *The Political System: An Inquiry into the State of Political Science*, 2nd ed. (1971), is the classic statement of systems theory in political science. Other useful sources for the theory in general include PETER M. SENGE, *The Fifth Discipline* (1993); and FRITJOF CAPRA, *The Turning Point: Science, Society, and the Rising Culture* (1982). IMMANUEL WALLERSTEIN, *The Modern World-System* (1974–), provides an insightful understanding of the historical antecedents to the current array of political and economic relations from a decidedly non-realist perspective. THOMAS FRIEDMAN, *The Lexus and the Olive Tree*, rev. ed. (2000), is an excellent popular use of systems theory in analyzing globalization.

Supranational organizations. LEROY BENNET, *International Organizations*, 7th ed. (2002), provides a relatively short overview of all surpanational organizations. LAWRENCE ZIRING, ROBERT E. RIGGS, and JACK C. PLANO, *The United Nations: International Organization and World Politics*, 3rd ed. (2000), explores the structure of the largest international organization. A detailed overview of the European Union is DESMOND DINAN, *Ever Closer Union*, 2nd ed. (1999). (Ch.Ha./Ed.)

Voting and elections. Electoral processes are examined in CHARLES SEYMOUR and DONALD PAIGE FRARY, *How the World Votes: The Story of Democratic Development in Elections*, 2 vol. (1918), still the most complete review of the history of elections in the English language; ANTHONY DOWNS, *An Economic Theory of Democracy* (1957); SEYMOUR MARTIN LIPSET and STEIN ROKKAN (eds.), *Party Systems and Voter Alignments: Cross-National Perspectives* (1967); GERALD L. CURTIS, *Election Campaigning, Japanese Style* (1971, reissued 1983); RICHARD S. KATZ, *A Theory of Parties and Electoral Systems* (1980); IVOR CREWE and DAVID DENVER (eds.), *Electoral Change in Western Democracies: Patterns and Sources of Electoral Volatility* (1985); PAUL W. DRAKE and EDUARDO SILVA (eds.), *Elections and Democratization in Latin America, 1980–1985* (1986); JANE J. MANSBRIDGE (ed.), *Beyond Self-Interest* (1990); MICHAEL LAVER and NORMAN SCHOFIELD, *Multiparty Government: The Politics of Coalition in Europe* (1990); WILLIAM A. BARNETT, MELVIN J. HINICH, and NORMAN SCHOFIELD (eds.), *Political Economy: Institutions, Competition, and Representation* (1993); SAMUEL L. POPKIN, *The Reasoning Voter*, 2nd ed. (1994); and IVOR CREWE, ANTHONY FOX, and NEIL DAY, *The British Electorate, 1963–1992*, rev. ed. (1995).

Political parties, elections, and interest groups are comparatively treated in MAURICE DUVERGER, *Political Parties*, 3rd ed. rev. (1964, reprinted 1972; originally published in French, 1951); MARY B. WELFLING, *Political Institutionalization: Comparative Analyses of African Party Systems* (1973); GIOVANNI SARTORI, *Parties and Party Systems*, vol. 1 (1976); RICHARD S. KATZ (ed.), *Party Government: European and American Experiences* (1987); EDMUND TERENCE GOMEZ, *Politics in Business: UMNO's Corporate Investments* (1990); PETER MAIR and GORDON SMITH (eds.), *Understanding Party System Change in Western Europe* (1990); and AREND LIJPHART, *Electoral Systems and Party Systems: A Study of Twenty-Seven Democracies, 1945–1990* (1994).

(H.E./R.G./P.D.W./Ed.)

Jackson Pollock

Jackson Pollock was a leading exponent of Abstract Expressionism, an art movement characterized by the free-associative gestures in paint sometimes referred to as "action painting." During his lifetime he received widespread publicity and serious recognition for the radical poured, or "drip," technique he used to create his major works. Among his contemporaries, he was respected for his deeply personal and totally uncompromising commitment to the art of painting. His work and example had enormous influence on them and on many subsequent art movements in the United States. He is also one of the first American painters to be recognized during his lifetime and after as a peer of 20th-century European masters.

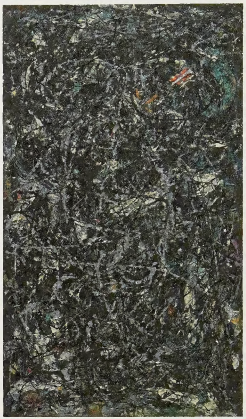
Early life and work. Paul Jackson Pollock, the fifth and youngest son of Stella May McClure and LeRoy Pollock, was born on Jan. 14, 1912, in Cody, Wyoming. The family left Cody 11 months after Jackson's birth; he would know his birthplace only through family photographs. Over the next 16 years his family lived in California and Arizona, eventually moving nine times. In 1928 they moved to Los Angeles, where Pollock enrolled at Manual Arts High School. There he came under the influence of Frederick John de St. Vrain Schwankovsky, a painter and illustrator who was also a member of the Theosophical Society, a sect that promoted metaphysical and occult spirituality. Schwankovsky gave Pollock some rudimentary training in drawing and painting, introduced him to advanced currents of European modern art, and encouraged his interest in theosophical literature. Pollock, who had been raised an agnostic, also attended the camp meetings of the former messiah of the theosophists, Jiddu Krishnamurti, a personal friend of Schwankovsky. These spiritual explorations prepared him to embrace the theories of the Swiss psychologist Carl Jung and the exploration of unconscious imagery in his paintings in subsequent years.

Spiritual explorations

In the fall of 1930 Pollock followed his brother Charles, who left home to study art in 1922, to New York City, where he enrolled at the Art Students League under his brother's teacher, the regionalist painter Thomas Hart Benton. (Jackson dropped his first name, Paul, about the time he went to New York in 1930.) He studied life drawing, painting, and composition with Benton for the next two and one-half years, leaving the league in the early months of 1933. For the next two years Pollock lived in poverty, first with Charles and, by the fall of 1934, with his brother Sanford. He would share an apartment in Greenwich Village with Sanford and his wife until 1942.

Early days in New York City

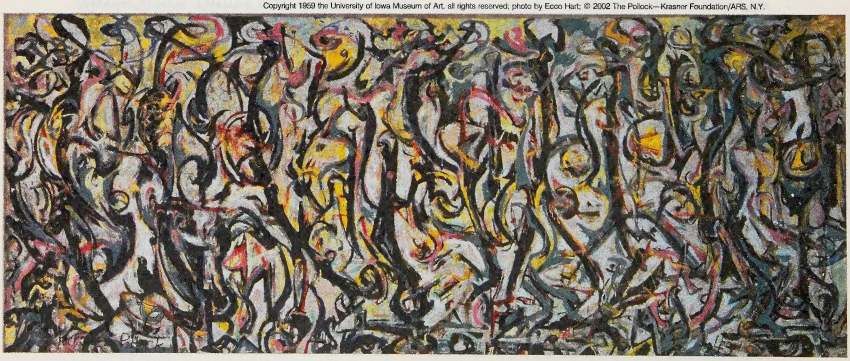
Pollock was employed by the WPA Federal Art Project in the fall of 1935 as an easel painter. This position gave him economic security during the remaining years of the Great Depression as well as an opportunity to develop his art. From his years with Benton through 1938, Pollock's work was strongly influenced by the compositional methods and



"Full Fathom Five," oil on canvas with mixed media by Jackson Pollock, 1947. In the Museum of Modern Art, New York City; gift of Peggy Guggenheim (186.1952). 129.2 × 76.5 cm.

Digital scan © The Museum of Modern Art/ Licensed by Scala/Art Resource; © 2002 The Pollock—Krasner Foundation/ARS, N.Y.

Copyright 1959 the University of Iowa Museum of Art, all rights reserved; photo by Ecco Hart; © 2002 The Pollock—Krasner Foundation/ARS, N.Y.



"Mural," oil on canvas by Jackson Pollock, *c.* 1943–44. In the collection of the University of Iowa Museum of Art, Iowa City, Iowa; gift of Peggy Guggenheim (1959.6). 6.05 × 2.47 m.

regionalist subject matter of his teacher and by the poetically expressionist vision of the American painter Albert Pinkham Ryder. It consisted mostly of small landscapes and figurative scenes such as "Going West" (1934–35), in which Pollock utilized motifs derived from photographs of his birthplace at Cody.

In 1937 Pollock began psychiatric treatment for alcoholism, and he suffered a nervous breakdown in 1938, which caused him to be institutionalized for about four months. After these experiences, his work became semi-abstract and showed the assimilation of motifs from the modern Spanish artists Pablo Picasso and Joan Miró, as well as the Mexican muralist José Clemente Orozco. Jungian symbolism and the Surrealist exploration of the unconscious also influenced his works of this period; indeed, from 1939 through 1941 he was in treatment with two successive Jungian psychoanalysts who used Pollock's own drawings in the therapy sessions. Characteristic paintings from this period include "Bird" (*c.* 1941), "Male and Female" (*c.* 1942), and "Guardians of the Secret" (1943).

Coming into maturity. In 1943, after the liquidation of the Federal Art Project, Pollock was given a contract by Peggy Guggenheim at her Art of This Century Gallery in New York, and his first one-man show was held there in November. Very late in 1943, possibly in the early weeks of 1944, Pollock painted his first wall-size work, called "Mural." This painting represents Pollock's breakthrough into a totally personal style in which Benton's compositional methods and energetic linear invention are fused with the Surrealist free association of motifs and unconscious imagery. Pollock's evolution from this point throughout the 1940s shows a struggle to find a process by which he could translate his entire personality into painting. The figurative character of works such as "Totem Lesson 1" (1944) and "The Blue Unconscious" (1946) contrasts with the heavily painted, all-over design of "Shimmering Substance" (1946) and "Eyes in the Heat" (1946), indicating the range of imagery and technique he employed during this period.

"Mural"

In 1945 Pollock married the painter Lee Krasner and moved to East Hampton, on the shore of Long Island, New York. Krasner, whom Pollock respected as an artist, had already proved her ability to handle his affairs with Guggenheim. She also provided a stabilizing factor that he sorely needed, given his drinking and social awkwardness.

"Poured" works. In 1947 Pollock first used the process of pouring or dripping paint onto a flat canvas in stages, often alternating weeks of painting with weeks of contemplating before he finished a canvas. This process allowed him to record the force and scope of his physical gesture in trajectories of enamel or aluminum paint. At the time, he said these abstract trajectories "veiled the image," or the traces of figuration, that had often been apparent in his earlier work. Recent research has indicated that his "veiling" constituted a form of free association from which he began most of his major paintings. The results, in effect, were huge areas covered with complex linear patterns that fused image and form; these works engulfed the spectator in their scale and intricacy. A whole series of paintings—beginning with "Full Fathom Five" (1947) and "Lucifer" (1947) and proceeding through "Summertime" (1948), "Number Ten, 1949" (1949), the mural-sized canvases of 1950 such as "One, Autumn Rhythm," and "Lavender Mist," and the black and white "Number Thirty-two, 1950" (1950)—display the infinite variety of effect and expression he achieved through the method of "poured" painting.

Free association

During the late 1940s and early '50s, Pollock had one-man shows of new paintings nearly every year in New York, his work being handled by Peggy Guggenheim through 1947, the Betty Parsons Gallery from 1947 to 1952, and finally by the Sidney Janis Gallery.

In 1951–52 Pollock painted almost exclusively in black enamel on unsized canvas, creating works in which his earlier imagery is evident. The configuration in "Number Twenty-three, 1951/'Frogman' " (1951), for example, relates to "Bird" and also to drawings Pollock did for his second Jungian analyst. Other important paintings from this phase are "Echo" (1951) and "Number Seven, 1952" (1952). In 1952 he returned to colour and mural scale in "Convergence" (1952) and "Blue Poles" (1952). He created his last series of major works in 1953; "Portrait and a Dream," "Easter and the Totem," "Ocean Greyness," and "The Deep," among other works, recapitulate many aspects of his former styles and images. Though his production waned and his health deteriorated after 1953, he did produce important paintings such as "White Light" (1954) and "Scent" (1955) in his last years. He died in an automobile accident in East Hampton, New York, on Aug. 11, 1956.

Later work

Assessment. As a man, Pollock was described by his contemporaries as gentle and contemplative when sober, violent when drunk. These extremes found equilibrium in his art. He was highly intelligent, widely read, and, when he chose, incisively articulate. He believed that art derived from the unconscious, saw himself as the essential subject of his painting, and judged his work and that of others on its inherent authenticity of personal expression.

During his lifetime, Pollock's critical reception ranged from the supportive criticism of Clement Greenberg in *The Nation* during the 1940s to *Time* magazine's pejorative reference to him as "Jack the Dripper" a few months before his death in 1956. Despite occasional attempts in the art press to understand his work seriously, his name became synonymous with extreme artistic caprice, since the novelty of his "pouring" technique overshadowed his obsession with the deeply personal expression that the technique permitted. Ironically, he did not profit financially from his fame. He never sold a painting for more than $10,000 in his lifetime and was often hard-pressed for cash. His work was more appreciated abroad. It was seen in Europe, for example, at the Venice Biennales of 1948, 1950, and 1956 and in a one-man show in Paris in 1952. In 1949 the French abstract artist Georges Mathieu stated that he considered Pollock the "greatest living American painter."

After Pollock's death, artists active in the American art movements immediately following Abstract Expressionism—such as "happenings," Pop art, Op art, and Colour Field painting—looked back, with more or less cause, to Pollock's example as fundamental to their departures. For these artists, he became the model of a painter who had successfully fused art and life. For critics, however, the psychologically oriented iconography in Pollock's work, prompted by Jungian psychotherapy and so important to Pollock and to a balanced understanding of his contribution, has generally been misunderstood or ignored. Posthumous critical opinion and biographical scholarship, strongly influenced by Greenberg and postmodernist theory, has tended largely to emphasize the formal elements of his work and his affinities with European art movements and artists; to exploit his drinking and alleged homosexuality; and to make unwarranted claims concerning his social identity as a renegade artist or pawn of the psychoanalytic community. Now considered an "iconic" master of mid-century Modernism, he has become all things to all interpreters, often in spite of the actual facts of his art and life. (F.V.O'C.)

Critical debate

BIBLIOGRAPHY. Biographies include FRANCIS V. O'CONNOR, *Jackson Pollock* (1967); B.H. FRIEDMAN, *Jackson Pollock: Energy Made Visible* (1972, reissued 1995); JEFFREY POTTER (compiler), *To a Violent Grave: An Oral Biography of Jackson Pollock* (1985, reprinted 1987); DEBORAH SOLOMON, *Jackson Pollock: A Biography* (1987, reissued 2001); and STEVEN NAIFEH and GREGORY WHITE SMITH, *Jackson Pollock: An American Saga* (1989, reissued 1992).

FRANCIS V. O'CONNOR and EUGENE VICTOR THAW (eds.), *Jackson Pollock*, 4 vol. (1978), is a catalogue raisonné; a supplement to this catalog, ed. by O'Connor, was published in 1995. BERNICE ROSE, *Jackson Pollock: Works on Paper* (1969), is a study of the drawings and small paintings. Interpretative monographs include FRANK O'HARA, *Jackson Pollock* (1959); BRYAN ROBERTSON, *Jackson Pollock* (1960); HANS NAMUTH, *Pollock Painting*, ed. by BARBARA ROSE (1980); ELIZABETH FRANK, *Jackson Pollock* (1983); ELLEN G. LANDAU, *Jackson Pollock* (1989, reissued 2000); and KIRK VARNEDOE and PEPE KARMEL, *Jackson Pollock* (1998).

An excellent short colour film of Pollock in the act of painting, *Jackson Pollock* (1951), was produced by PAUL FALKENBERG and HANS NAMUTH, with music by MORTON FELDMAN and narration by the artist. (F.V.O'C.)

Population

Few aspects of human societies are as fundamental as the size, composition, and rate of change of their populations. Such factors affect economic prosperity, health, education, family structure, crime patterns, language, culture—indeed, virtually every aspect of human society is touched upon by population trends.

The study of human populations is called demography—a discipline with intellectual origins stretching back to the 18th century, when it was first recognized that human mortality could be examined as a phenomenon with statistical regularities. Demography casts a multidisciplinary net, drawing insights from economics, sociology, statistics, medicine, biology, anthropology, and history. Its chronological sweep is lengthy: limited demographic evidence for many centuries into the past, and reliable data for several hundred years are available for many regions. The present understanding of demography makes it possible to project (with caution) population changes several decades into the future.

This article is divided into the following sections:

The basic components of population change 1034
- Fertility
- Mortality
- Marriage
- Migration
- Natural increase and population growth

Population composition 1038
- Age distribution
- Sex ratio
- Ethnic or racial composition
- Geographical distribution and urbanization

Population theories 1039
- Population theories in antiquity
- Mercantilism and the idea of progress
- Physiocrats and the origins of demography
- Utopian views
- Malthus and his successors
- Marx, Lenin, and their followers
- The Darwinian tradition
- Theory of the demographic transition

Trends in world population 1041
- The developing countries since 1950
- The industrialized countries since 1950

Population projections 1042

Bibliography 1042

THE BASIC COMPONENTS OF POPULATION CHANGE

Closed and open populations

At its most basic level, the components of population change are few indeed. A closed population (that is, one in which immigration and emigration do not occur) can change according to the following simple equation: the population (closed) at the end of an interval equals the population at the beginning of the interval, plus births during the interval, minus deaths during the interval. In other words, only addition by births and reduction by deaths can change a closed population. The notion of a closed population is not an abstraction; unless one believes that there has been substantial migration to and from this planet, the world population as a whole is closed.

Populations of nations, regions, continents, islands, or cities, however, are rarely closed in the same way. If the assumption of a closed population is relaxed, in- and out-migration can increase and decrease population size in the same way as do births and deaths; thus, the population (open) at the end of an interval equals the population at the beginning of the interval, plus births during the interval, minus deaths, plus in-migrants, minus out-migrants. Hence the study of demographic change requires knowledge of fertility (births), mortality (deaths), and migration. These, in turn, affect not only population size and growth rates but also the composition of the population in terms of such attributes as sex, age, ethnic or racial composition, and geographic distribution.

Fertility. Demographers distinguish between fecundity, the underlying biological potential for reproduction, and fertility, the actual level of achieved reproduction. (Confusingly, these English terms have opposite meanings from their parallel terms in French, where *fertilité* is the potential and *fécondité* is the realized; similarly ambiguous usages also prevail in the biological sciences, thereby increasing the chance of misunderstanding.) The difference between biological potential and realized fertility is determined by several intervening factors, including the following: (1) most women do not begin reproducing immediately upon the onset of puberty, which itself does not occur at a fixed age; (2) some women with the potential to reproduce never do so; (3) some women become widowed and do not remarry; (4) various elements of social behaviour restrain fertility; and (5) many human couples choose consciously to restrict their fertility by means of sexual abstinence, contraception, abortion, or sterilization.

Potential and realized fertility

The magnitude of the gap between potential and realized fertility can be illustrated by comparing the highest known fertilities with those of typical European and North American women in the late 20th century. A well-studied high-fertility group is the Hutterites of North America, a religious sect that views fertility regulation as sinful and high fertility as a blessing. Hutterite women who married between 1921 and 1930 are known to have averaged 10 children per woman. Meanwhile, women in much of Europe and North America averaged about two children per woman during the 1970s and 1980s—a number 80 percent less than that achieved by the Hutterites. Even the highly fertile populations of developing countries in Africa, Asia, and Latin America produce children at rates far below that of the Hutterites.

The general message from such evidence is clear enough: in much of the world, human fertility is considerably lower than the biological potential. It is strongly constrained by cultural regulations, especially those concerning marriage and sexuality, and by conscious efforts on the part of married couples to limit their childbearing.

Dependable evidence on historical fertility patterns in Europe is available back to the 18th century, and estimates have been made for several earlier centuries. Such data for non-European societies and for earlier human populations are much more fragmentary. The European data indicate that even in the absence of widespread deliberate regulation there were significant variations in fertility among different societies. These differences were heavily affected by socially determined behaviours such as those concerning marriage patterns. Beginning in France and Hungary in the 18th century, a dramatic decline in fertility took shape in the more developed societies of Europe and North America, and in the ensuing two centuries fertility declines of fully 50 percent took place in nearly all of these countries. Since the 1960s fertility has been intentionally diminished in many developing countries, and remarkably rapid reductions have occurred in the most populous, the People's Republic of China.

There is no dispute as to the fact and magnitudes of such declines, but theoretical explanation of the phenomena has proved elusive. (See below *Population theories.*)

Biological factors affecting human fertility. Reproduction is a quintessentially biological process, and hence all fertility analyses must consider the effects of biology. Such factors, in rough chronological order, include:

the age of onset of potential fertility (or fecundability in demographic terminology);

the degree of fecundability—*i.e.*, the monthly probability of conceiving in the absence of contraception;

the incidence of spontaneous abortion and stillbirth;
the duration of temporary infecundability following the birth of a child; and
the age of onset of permanent sterility.

The age at which women become fecund apparently declined significantly during the 20th century; as measured by the age of menarche (onset of menstruation), British data suggest a decline from 16–18 years in the mid-19th century to less than 13 years in the late 20th century. This decline is thought to be related to improving standards of nutrition and health. Since the average age of marriage in western Europe has long been far higher than the age of menarche, and since most children are born to married couples, this biological lengthening of the reproductive period is unlikely to have had major effects upon realized fertility in Europe. In settings where early marriage prevails, however, declining age at menarche could increase lifetime fertility.

Variations in fecundability

Fecundability also varies among women past menarche. The monthly probabilities of conception among newlyweds are commonly in the range of 0.15 to 0.25; that is, there is a 15–25-percent chance of conception each month. This fact is understandable when account is taken of the short interval (about two days) within each menstrual cycle during which fertilization can take place. Moreover, there appear to be cycles during which ovulation does not occur. Finally, perhaps one-third or more of fertilized ova fail to implant in the uterus or, even if they do implant, spontaneously abort during the ensuing two weeks, before pregnancy would be recognized. As a result of such factors, women of reproductive age who are not using contraceptive methods can expect to conceive within five to 10 months of becoming sexually active. As is true of all biological phenomena, there is surely a distribution of fecundability around average levels, with some women experiencing conception more readily than others.

Spontaneous abortion of recognized pregnancies and stillbirth also are fairly common, but their incidence is difficult to quantify. Perhaps 20 percent of recognized pregnancies fail spontaneously, most in the earlier months of gestation.

Following the birth of a child, most women experience a period of temporary infecundability, or biological inability to conceive. The length of this period seems to be affected substantially by breast-feeding. In the absence of breast-feeding, the interruption lasts less than two months. With lengthy, frequent breast-feeding it can last one or two years. This effect is thought to be caused by a complex of neural and hormonal factors stimulated by suckling.

A woman's fecundability typically peaks in her 20s and declines during her 30s; by their early 40s as many as 50 percent of women are affected by their own or their husbands' sterility. After menopause, essentially all women are sterile. The average age at menopause is in the late 40s, although some women experience it before reaching 40 and others not until nearly 60.

Contraception. Contraceptive practices affect fertility by reducing the probability of conception. Contraceptive methods vary considerably in their theoretical effectiveness and in their actual effectiveness in use ("use-effectiveness"). Modern methods such as oral pills and intrauterine devices (IUDs) have use-effectiveness rates of more than 95 percent. Older methods such as the condom and diaphragm can be more than 90-percent effective when used regularly and correctly, but their average use-effectiveness is lower because of irregular or incorrect use.

The effect upon fertility of contraceptive measures can be dramatic: if fecundability is 0.20 (a 20-percent chance of pregnancy per month of exposure), then a 95-percent effective method will reduce this to 0.01 (a 1-percent chance).

Abortion. Induced abortion reduces fertility not by affecting fecundability but by terminating pregnancy. Abortion has long been practiced in human societies and is quite common in some settings. The officially registered fraction of pregnancies terminated by abortion exceeds one-third in some countries, and significant numbers of unregistered abortions probably occur even in countries reporting very low rates.

Sterilization. Complete elimination of fecundability can be brought about by sterilization. The surgical procedures of tubal ligation and vasectomy have become common in diverse nations and cultures. In the United States, for example, voluntary sterilization has become the most prevalent single means of terminating fertility, typically adopted by couples who have achieved their desired family size. In India, sterilization has been encouraged on occasion by various government incentive programs and, for a short period during the 1970s, by quasi-coercive measures.

Mortality. As noted above, the science of demography has its intellectual roots in the realization that human mortality, while consisting of unpredictable individual events, has a statistical regularity when aggregated across a large group. This recognition formed the basis of a wholly new industry—that of life assurance, or insurance. The basis of this industry is the life table, or mortality table, which summarizes the distribution of longevity—observed over a period of years—among members of a population. This statistical device allows the calculation of premiums—the prices to be charged the members of a group of living subscribers with specified characteristics, who by pooling their resources in this statistical sense provide their heirs with financial benefits. Life table

Overall human mortality levels can best be compared by using the life-table measure life expectancy at birth (often abbreviated simply as life expectancy), the number of years of life expected of a newborn baby on the basis of current mortality levels for persons of all ages. Life expectancies of premodern populations, with their poor knowledge of sanitation and health care, may have been as low as 25–30 years. The largest toll of death was that exacted in infancy and childhood: perhaps 20 percent of newborn children died in their first 12 months of life and another 30 percent before they reached five years of age.

In the developing countries by the 1980s, average life expectancy lay in the range of 55 to 60 years, with the highest levels in Latin America and the lowest in Africa. In the same period, life expectancy in the developed countries of western Europe and North America approached 75 years, and fewer than 1 percent of newborn children died in their first 12 months.

Greater longevity of females

For reasons that are not well understood, life expectancy of females usually exceeds that of males, and this female advantage has grown as overall life expectancy has increased. In the late 20th century this female advantage was seven years (78 years versus 71 years) in the industrial market economies (comprising western Europe, North America, Japan, Australia, and New Zealand). It was eight years (74 years versus 66 years) in the nonmarket economies of eastern Europe.

The epidemiologic transition. The epidemiologic transition is that process by which the pattern of mortality and disease is transformed from one of high mortality among infants and children and episodic famine and epidemic affecting all age groups to one of degenerative and man-made diseases (such as those attributed to smoking) affecting principally the elderly. It is generally believed that the epidemiologic transitions prior to the 20th century (*i.e.,* those in today's industrialized countries) were closely associated with rising standards of living, nutrition, and sanitation. In contrast, those occurring in developing countries have been more or less independent of such internal socioeconomic development and more closely tied to organized health care and disease control programs developed and financed internationally. There is no doubt that 20th-century declines in mortality in developing countries have been far more rapid than those that occurred in the 19th century in what are now the industrialized countries.

Infant mortality. Infant mortality is conventionally measured as the number of deaths in the first year of life per 1,000 live births during the same year. Roughly speaking, by this measure worldwide infant mortality approximates 80 per 1,000; that is, about 8 percent of newborn babies die within the first year of life.

Variations in infant mortality

This global average disguises great differences. In certain countries of Asia and Africa, infant mortality rates exceed 150 and sometimes approach 200 per 1,000 (that is, 15 or 20 percent of children die before reaching the age of

one year). Meanwhile, in other countries, such as Japan and Sweden, the rates are well below 10 per 1,000, or 1 percent. Generally, infant mortality is somewhat higher among males than among females.

In developing countries substantial declines in infant mortality have been credited to improved sanitation and nutrition, increased access to modern health care, and improved birth spacing through the use of contraception. In industrialized countries in which infant mortality rates were already low the increased availability of advanced medical technology for newborn—in particular, prematurely born—infants provides a partial explanation.

Infanticide. The deliberate killing of newborn infants has long been practiced in human societies. It seems to have been common in the ancient cultures of Greece, Rome, and China, and it was practiced in Europe until the 19th century. In Europe, infanticide included the practice of "overlaying" (smothering) an infant sharing a bed with its parents and the abandonment of unwanted infants to the custody of foundling hospitals, in which one-third to four-fifths of incumbents failed to survive.

In many societies practicing infanticide, infants were not deemed to be fully human until they underwent a rite of initiation that took place from a few days to several years after birth, and therefore killing before such initiation was socially acceptable. The purposes of infanticide were various: child spacing or fertility control in the absence of effective contraception; elimination of illegitimate, deformed, orphaned, or twin children; or sex preferences.

With the development and spread of the means of effective fertility regulation, infanticide has come to be strongly disapproved in most societies, though it continues to be practiced in some isolated traditional cultures.

Mortality among the elderly. During the 1970s and 1980s in industrialized countries there were unexpectedly large declines in mortality among the elderly, resulting in larger-than-projected numbers of the very old. In the United States, for example, the so-called frail elderly group aged 85 years and older increased nearly fourfold between 1950 and 1980, from 590,000 to 2,461,000. Given the high incidence of health problems among the very old, such increases have important implications for the organization and financing of health care.

Marriage. One of the main factors affecting fertility, and an important contributor to the fertility differences among societies in which conscious fertility control is uncommon, is defined by the patterns of marriage and marital disruption. In many societies in Asia and Africa, for example, marriage occurs soon after the sexual maturation of the woman, around age 17. In contrast, delayed marriage has long been common in Europe, and in some European countries the average age of first marriage approaches 25 years. Another aspect of historical marriage patterns in Europe is the high incidence of permanent celibacy: as many as 10 percent of women never marry. In contrast, nearly all women marry in most traditional developing countries of Asia and Africa.

Disruption of marriage

In the 20th century dramatic changes have taken place in the patterns of marital dissolution caused by widowhood and divorce. Widowhood has long been common in all societies, but the declines of mortality (as discussed above) have sharply reduced the effects of this source of marital dissolution on fertility. Meanwhile, divorce has been transformed from an uncommon exception to an experience terminating a large proportion (sometimes more than a third) of marriages in some countries. Taken together, these components of marriage patterns can account for the elimination of as little as 20 percent to as much as 50 percent of the potential reproductive years.

Many Western countries have experienced significant increases in the numbers of cohabiting unmarried couples. In the 1970s some 12 percent of all Swedish couples living together aged 16 to 70 were unmarried. When in the United States in 1976 the number of such arrangements approached 1,000,000, the Bureau of the Census formulated a new statistical category—POSSLQ—denoting persons of the opposite sex sharing living quarters. Extramarital fertility as a percentage of overall fertility accordingly has risen in many Western countries, accounting for one in five births in the United States, one in five in Denmark, and one in three in Sweden.

Migration. Since any population that is not closed can be augmented or depleted by in-migration or out-migration, migration patterns must be considered carefully in analyzing population change. The common definition of human migration limits the term to permanent change of residence (conventionally, for at least one year), so as to distinguish it from commuting and other more frequent but temporary movements.

Human migrations have been fundamental to the broad sweep of human history and have themselves changed in basic ways over the epochs. Many of these historical migrations have by no means been the morally uplifting experiences depicted in mythologies of heroic conquerors, explorers, and pioneers; rather they frequently have been characterized by violence, destruction, bondage, mass mortality, and genocide—in other words, by human suffering of profound magnitudes.

Early human migrations. Early humans were almost surely hunters and gatherers who moved continually in search of food supplies. The superior technologies (tools, clothes, language, disciplined cooperation) of these hunting bands allowed them to spread farther and faster than had any other dominant species; humans are thought to have occupied all the continents except Antarctica within a span of about 50,000 years. As the species spread away from the tropical parasites and diseases of its African origins, mortality rates declined and population increased. This increase occurred at microscopically small rates by the standards of the past several centuries, but over thousands of years it resulted in a large absolute growth to a total that could no longer be supported by finding new hunting grounds. There ensued a transition from migratory hunting and gathering to migratory slash-and-burn agriculture. The consequence was the rapid geographical spread of crops, with wheat and barley moving east and west from the Middle East across the whole of Eurasia within only 5,000 years.

Growth of migratory populations

About 10,000 years ago a new and more productive way of life, involving sedentary agriculture, became predominant. This allowed greater investment of labour and technology in crop production, resulting in a more substantial and securer food source, but sporadic migrations persisted.

The next pulse of migration, beginning around 4000 to 3000 BC, was stimulated by the development of seagoing sailing vessels and of pastoral nomadry. The Mediterranean Basin was the centre of the maritime culture, which involved the settlement of offshore islands and led to the development of deep-sea fishing and long-distance trade. Other favoured regions were those of the Indian Ocean and South China Sea. Meanwhile, pastoral nomadry involved biological adaptations both in humans (allowing them to digest milk) and in species of birds and mammals that were domesticated. Once completed, these adaptations allowed humans to consume the meat of most male newborn animals and the maternal milk thereby made available.

Both seafarers and pastoralists were intrinsically migratory. The former were able to colonize previously uninhabited lands or to impose their rule by force over less mobile populations. The pastoralists were able to populate the extensive grassland of the Eurasian Steppe and the African and Middle Eastern savannas, and their superior nutrition and mobility gave them clear military advantages over the sedentary agriculturalists with whom they came into contact. Even as agriculture continued to improve with innovations such as the plow, these mobile elements persisted and provided important networks by which technological innovations could be spread widely and rapidly.

That complex of human organization and behaviour commonly termed Western civilization arose out of such developments. Around 4000 BC seafaring migrants from the south overwhelmed the local inhabitants of the Tigris–Euphrates floodplain and began to develop a social organization based upon the division of labour into highly skilled occupations, technologies such as irrigation, bronze metallurgy, and wheeled vehicles, and the growth of cities

Urban cultures

of 20,000–50,000 persons. Political differentiation into ruling classes and ruled masses provided a basis for imposition of taxes and rents that financed the development of professional soldiers and artisans, whose specialized skills far surpassed those of pastoralists and agriculturalists. The military and economic superiority that accompanied such skills allowed advanced communities to expand both by direct conquest and by the adoption of this social form by neighbouring peoples. Thus migration patterns played an important role in creating the early empires and cultures of the ancient world.

By about 2000 BC such specialized human civilizations occupied much of the then-known world—the Middle East, the eastern Mediterranean, South Asia, and the Far East. Under these circumstances human migration was transformed from unstructured movements across unoccupied territories by nomads and seafarers into quite new forms of interaction among the settled civilizations.

These new forms of human migration produced disorder, suffering, and much mortality. As one population conquered or infiltrated another, the vanquished were usually destroyed, enslaved, or forcibly absorbed. Large numbers of people were captured and transported by slave traders. Constant turmoil accompanied the ebb and flow of populations across the regions of settled agriculture and the Eurasian and African grasslands. Important examples include the Dorian incursions in ancient Greece in the 11th century BC, the Germanic migrations southward from the Baltic to the Roman Empire in the 4th to 6th centuries AD, the Norman raids and conquests of Britain between the 8th and 12th centuries AD, and the Bantu migrations in Africa throughout the Christian Era.

Modern mass migrations. Mass migrations over long distances were among the new phenomena produced by the population increase and improved transportation that accompanied the Industrial Revolution. The largest of these was the so-called Great Atlantic Migration from Europe to North America, the first major wave of which began in the late 1840s with mass movements from Ireland and Germany. These were caused by the failure of the potato crop in Ireland and in the lower Rhineland, where millions had become dependent upon this single source of nutrition. These flows eventually subsided, but in the 1880s a second and even larger wave of mass migration developed from eastern and southern Europe, again stimulated in part by agricultural crises and facilitated by improvements in transportation and communication. Between 1880 and 1910 some 17,000,000 Europeans entered the United States; overall, the total amounted to 37,000,000 between 1820 and 1980.

European settlement of America

Since World War II equally large long-distance migrations have occurred. In most cases groups from developing countries have moved into the industrialized countries of the West. Some 13,000,000 migrants have become permanent residents of western Europe since the 1960s. More than 10,000,000 permanent immigrants have been admitted legally to the United States since the 1960s, and illegal immigration has almost surely added several millions more.

Forced migrations. Slave migrations and mass expulsions have been part of human history for millennia. The largest slave migrations were probably those compelled by European slave traders operating in Africa from the 16th to the 19th century. During that period perhaps 20,000,000 slaves were consigned to American markets, though substantial numbers died in the appalling conditions of the Atlantic passage.

The largest mass expulsion is probably that imposed by the Nazi government of Germany, which deported 7,000,000–8,000,000 persons, including some 5,000,000 Jews later exterminated in concentration camps. After World War II, 9,000,000–10,000,000 ethnic Germans were more or less forcibly transported into Germany, and perhaps 1,000,000 members of minority groups deemed politically unreliable by the Soviet government were forcibly exiled to Central Asia. Earlier deportations of this type included the movement of 150,000 British convicts to Australia between 1788 and 1867 and the 19th-century exile of 1,000,000 Russians to Siberia.

Political refugees

Forced migrations since World War II have been large indeed. Some 14,000,000 persons fled in one direction or the other at the partition of British India into India and Pakistan. Nearly 10,000,000 left East Pakistan (now Bangladesh) during the fighting in 1971; many of them stayed on in India. An estimated 3,000,000–4,000,000 persons fled from the war in Afghanistan during the early 1980s. More than 1,000,000 refugees have departed Vietnam, Cuba, Israel, and Ethiopia since World War II. Estimates during the 1980s suggested that approximately 10,000,000 refugees had not been resettled and were in need of assistance.

Internal migrations. The largest human migrations today are internal to nation-states; these can be sizable in rapidly increasing populations with large rural-to-urban migratory flows.

Early human movements toward urban areas were devastating in terms of mortality. Cities were loci of intense infection; indeed, many human viral diseases are not propagated unless the population density is far greater than that common under sedentary agriculture or pastoral nomadism. Moreover, cities had to import food and raw materials from the hinterlands, but transport and political disruptions led to erratic patterns of scarcity, famine, and epidemic. The result was that cities until quite recently (the mid-19th century) were demographic sinkholes, incapable of sustaining their own populations.

Urban growth since World War II has been very rapid in much of the world. In developing countries with high overall population growth rates the populations of some cities have been doubling every 10 years or less (see below *Population composition*).

Natural increase and population growth. *Natural increase.* Put simply, natural increase is the difference between the numbers of births and deaths in a population; the rate of natural increase is the difference between the birthrate and the death rate (see Figure 1). Given the fertility and mortality characteristics of the human species (excluding incidents of catastrophic mortality), the range of possible rates of natural increase is rather narrow. For a nation, it has rarely exceeded 4 percent per year; the highest known rate for a national population—arising from the conjunction of a very high birthrate and a quite low death rate—is that experienced in Kenya during the 1980s, in which the natural increase of the population approximated 4.1 percent per annum. Rates of natural increase in other developing countries generally are lower; these countries averaged about 2.5 percent per annum during the same period. Meanwhile the rates of natural increase in industrialized countries are very low: the highest is approximately 1 percent, most are in the neighbourhood of several tenths of 1 percent, and some are slightly negative (that is, their populations are slowly decreasing).

Population growth. The rate of population growth is the rate of natural increase combined with the effects of migration. Thus a high rate of natural increase can be offset by a large net out-migration, and a low rate of natural increase can be countered by a high level of net in-migration. Generally speaking, however, these migration effects on population growth rates are far smaller than the effects of changes in fertility and mortality.

Population "momentum." An important and often misunderstood characteristic of human populations is the tendency of a highly fertile population that has been increasing rapidly in size to continue to do so for decades after the onset of even a substantial decline in fertility. This results from the youthful age structure of such a population, as discussed below. These populations contain large numbers of children who have still to grow into adulthood and the years of reproduction. Thus even a dramatic decline in fertility, which affects only the numbers at age zero, cannot prevent the continuing growth of the number of adults of childbearing age for at least two or three decades.

Eventually, of course, as these large groups pass through the childbearing years to middle and older age, the smaller numbers of children resulting from the fertility decline lead to a moderation in the rate of population growth. But the delays are lengthy, allowing very substantial additional

population growth after fertility has declined. This phenomenon gives rise to the term population momentum, which is of great significance to developing countries with rapid population growth and limited natural resources. The nature of population growth means that the metaphor of a "population bomb" used by some lay analysts of population trends in the 1960s was really quite inaccurate. Bombs explode with tremendous force, but such force is rapidly spent. A more appropriate metaphor for rapid population growth is that of a glacier, since a glacier moves at a slow pace but with enormous effects wherever it goes and with a long-term momentum that is unstoppable.

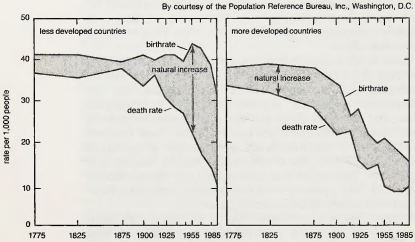


Figure 1: Population growth through natural increase in the less developed and more developed countries, 1775–1985.

POPULATION COMPOSITION

The most important characteristics of a population—in addition to its size and the rate at which it is expanding or contracting—are the ways in which its members are distributed according to age, sex, ethnic or racial category, and residential status (urban or rural).

Population pyramids

Age distribution. Perhaps the most fundamental of these characteristics is the age distribution of a population. Demographers commonly use population pyramids to describe both age and sex distributions of populations. A population pyramid is a bar chart or graph in which the length of each horizontal bar represents the number (or percentage) of persons in an age group; for example, the base of such a chart consists of a bar representing the youngest segment of the population, those persons less than, say, five years old. Each bar is divided into segments corresponding to the numbers (or proportions) of males and females. In most populations the proportion of older persons is much smaller than that of the younger, so the chart narrows toward the top and is more or less triangular, like the cross section of a pyramid; hence the name. Youthful populations are represented by pyramids with a broad base of young children and a narrow apex of older people, while older populations are characterized by more uniform numbers of people in the age categories. Figure 2 presents population pyramids for three nations that have markedly different characteristics: high fertility and rapid population growth (Mexico), low fertility and slow growth (United States), and very low fertility and negative growth (West Germany).

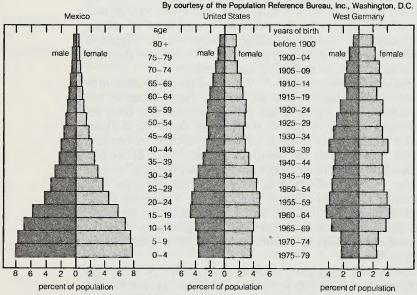


Figure 2: Population pyramids, showing age and sex distributions of the populations of Mexico, the United States, and West Germany in 1980.

Sex differences in natality and mortality

Sex ratio. A second important structural aspect of populations is the relative numbers of males and females who compose it. Generally, slightly more males are born than females (a typical ratio would be 105 or 106 males for every 100 females). On the other hand, it is quite common for males to experience higher mortality at virtually all ages after birth. This difference is apparently of biological origin. Exceptions occur in countries such as India, where the mortality of females may be higher than that of males in childhood and at the ages of childbearing because of unequal allocation of resources within the family and the poor quality of maternal health care.

The general rules that more males are born but that females experience lower mortality mean that during childhood males outnumber females of the same age, the difference decreases as the age increases, at some point in the adult life span the numbers of males and females become equal, and as higher ages are reached the number of females becomes disproportionately large. For example, in Europe and North America, among persons more than 70 years of age in 1985, the number of males for every 100 females was only about 61 to 63. (According to the Population Division of the United Nations, the figure for the Soviet Union was only 40, which may be attributable to high male mortality during World War II as well as to possible increases in male mortality during the 1980s.)

The sex ratio within a population has significant implications for marriage patterns. A scarcity of males of a given age depresses the marriage rates of females in the same age group or usually those somewhat younger, and this in turn is likely to reduce their fertility. In many countries, social convention dictates a pattern in which males at marriage are slightly older than their spouses. Thus if there is a dramatic rise in fertility, such as that called the "baby boom" in the period following World War II, a "marriage squeeze" can eventually result; that is, the number of males of the socially correct age for marriage is insufficient for the number of somewhat younger females. This may lead to deferment of marriage of these women, a contraction of the age differential of marrying couples, or both. Similarly, a dramatic fertility decline in such a society is likely to lead eventually to an insufficiency of eligible females for marriage, which may lead to earlier marriage of these women, an expansion of the age gap at marriage, or both. All of these effects are slow to develop; it takes at least 20 to 25 years for even a dramatic fall or rise in fertility to affect marriage patterns in this way.

Ethnic or racial composition. The populations of all nations of the world are more or less diverse with respect to ethnicity or race. (Ethnicity here includes national, cultural, religious, linguistic, or other attributes that are perceived as characteristic of distinct groups.) Such divisions in populations often are regarded as socially important, and statistics by race and ethnic group are therefore commonly available. The categories used for such groups differ from nation to nation, however; for example, a person

of Pakistani origin is considered "black" or "coloured" in the United Kingdom but would probably be classified as "white" or "Asian" in the United States. For this reason, international comparisons of ethnic and racial groups are imprecise, and this component of population structure is far less objective as a measure than are the categories of age and sex discussed above.

Geographical distribution and urbanization. It goes without saying that populations are scattered across space. The typical measure of population in relation to land area, that of population density, is often a meaningless one, since different areas vary considerably in their value for agricultural or other human purposes. Moreover, a high population density in an agrarian society, dependent upon agriculture for its sustenance, is likely to be a severer constraint upon human welfare than would the same density in a highly industrialized society, in which the bulk of national product is not of agricultural origin.

Also of significance in terms of geographical distribution is the division between rural and urban areas. For many decades there has been a nearly universal flow of populations from rural into urban areas. While definitions of urban areas differ from country to country and region to region, the most highly urbanized societies in the world are those of western and northern Europe, Australia, New Zealand, temperate South America, and North America; in all of these the fraction of the population living in urban areas exceeds 75 percent, and it has reached 85 percent in West Germany. An intermediate stage of urbanization exists in the countries making up much of tropical Latin America, where 50 to 65 percent of the population lives in cities. Finally, in many of the developing countries of Asia and Africa the urbanization process has only recently begun, and it is not uncommon to find less than one-third of the population living in urban areas.

Growth of cities

The rapidity of urbanization in some countries is quite astonishing. The population of Mexico City in 1960 was around 5,000,000; it was estimated to be about 17,000,000 in 1985 and was projected to reach 26,000,000 to 31,000,000 by 2000. A rule of thumb for much of the developing world is that the rate of growth of urban areas is twice that of the population as a whole. Thus in a population growing 3 percent annually (doubling in about 23.1 years), it is likely that the urban growth rate is at least 6 percent annually (doubling in about 11.6 years).

POPULATION THEORIES

Population size and change play such a fundamental role in human societies that they have been the subject of theorizing for millennia. Most religious traditions have had something to say on these matters, as did many of the leading figures of the ancient world.

In modern times the subject of demographic change has played a central role in the development of the politico-economic theory of mercantilism; the classical economics of Adam Smith, David Ricardo, and others; the cornucopian images of utopians such as the Marquis de Condorcet; the contrasting views of Malthus as to the natural limits imposed on human population; the sociopolitical theories of Marx, Engels, and their followers; the scientific revolutions engendered by Darwin and his followers; and so on through the pantheon of human thought. Most of these theoretical viewpoints have incorporated demographic components as elements of far grander schemes. Only in a few cases have demographic concepts played a central role, as in the case of the theory of the demographic transition that evolved during the 1930s as a counter to biological explanations of fertility declines that were then current.

Population theories in antiquity. The survival of ancient human societies despite high and unpredictable mortality implies that all societies that persisted were successful in maintaining high fertility. They did so in part by stressing the duties of marriage and procreation and by stigmatizing persons who failed to produce children. Many of these pronatalist motives were incorporated into religious dogma and mythology, as in the biblical injunction to "be fruitful and multiply, and populate the earth," the Hindu laws of Manu, and the writings of Zoroaster.

Greek and Roman views

The ancient Greeks were interested in population size, and Plato's *Republic* incorporated the concept of an optimal population size of 5,040 citizens, among whom fertility was restrained by conscious birth control. The leaders of imperial Rome, however, advocated maximizing population size in the interest of power, and explicitly pronatalist laws were adopted during the reign of Augustus to encourage marriage and fertility.

The traditions of Christianity on this topic are mixed. The pronatalism of the Old Testament and the Roman Empire was embraced with some ambivalence by a church that sanctified celibacy among the priesthood. Later, during the time of Thomas Aquinas, the church moved toward more forceful support of high fertility and opposition to birth control.

Islāmic writings on fertility were equally mixed. The 14th-century Arab historian Ibn Khaldūn incorporated demographic factors into his grand theory of the rise and fall of empires. According to his analysis, the decline of an empire's population necessitates the importation of foreign mercenaries to administer and defend its territories, resulting in rising taxes, political intrigue, and general decadence. The hold of the empire on its hinterland and on its own populace weakens, making it a tempting target for a vigorous challenger. Thus Ibn Khaldūn saw the growth of dense human populations as generally favourable to the maintenance and increase of imperial power.

On the other hand, contraception was acceptable practice in Islām from the days of the Prophet, and extensive attention was given to contraceptive methods by the great physicians of the Islāmic world during the Middle Ages. Moreover, under Islāmic law the fetus is not considered a human being until its form is distinctly human, and hence early abortion was not forbidden.

Mercantilism and the idea of progress. The wholesale mortality caused by the Black Death during the 14th century contributed in fundamental ways to the development of mercantilism, the school of thought that dominated Europe from the 16th through the 18th century. Mercantilists and the absolute rulers who dominated many states of Europe saw each nation's population as a form of national wealth: the larger the population, the richer the nation. Large populations provided a larger labour supply, larger markets, and larger (and hence more powerful) armies for defense and for foreign expansion. Moreover, since growth in the number of wage earners tended to depress wages, the wealth of the monarch could be increased by capturing this surplus. In the words of Frederick II the Great of Prussia, "the number of the people makes the wealth of states." Similar views were held by mercantilists in Germany, France, Italy, and Spain. For the mercantilists, accelerating the growth of the population by encouraging fertility and discouraging emigration was consistent with increasing the power of the nation or the king. Most mercantilists, confident that any number of people would be able to produce their own subsistence, had no worries about harmful effects of population growth. (To this day similar optimism continues to be expressed by diverse schools of thought, from traditional Marxists on the left to "cornucopians" on the right.)

Laissez-faire economics

Physiocrats and the origins of demography. By the 18th century the Physiocrats were challenging the intensive state intervention that characterized the mercantilist system, urging instead the policy of laissez-faire. Their targets included the pronatalist strategies of governments; Physiocrats such as François Quesnay argued that human multiplication should not be encouraged to a point beyond that sustainable without widespread poverty. For the Physiocrats, economic surplus was attributable to land, and population growth could therefore not increase wealth. In their analysis of this subject matter the Physiocrats drew upon the techniques developed in England by John Graunt, Edmond Halley, Sir William Petty, and Gregory King, which for the first time made possible the quantitative assessment of population size, the rate of growth, and rates of mortality.

The Physiocrats had broad and important effects upon the thinking of the classical economists such as Adam Smith, especially with respect to the role of free markets

unregulated by the state. As a group, however, the classical economists expressed little interest in the issue of population growth, and when they did they tended to see it as an effect rather than as a cause of economic prosperity.

Utopian views. In another 18th-century development, the optimism of mercantilists was incorporated into a very different set of ideas, those of the so-called utopians. Their views, based upon the idea of human progress and perfectibility, led to the conclusion that once perfected, mankind would have no need of coercive institutions such as police, criminal law, property ownership, and the family. In a properly organized society, in their view, progress was consistent with any level of population, since population size was the principal factor determining the amount of resources. Such resources should be held in common by all persons, and if there were any limits on population growth, they would be established automatically by the normal functioning of the perfected human society. Principal proponents of such views included Condorcet, William Godwin, and Daniel Malthus, the father of the Reverend Thomas Robert Malthus. Through his father the younger Malthus was introduced to such ideas relating human welfare to population dynamics, which stimulated him to undertake his own collection and analysis of data; these eventually made him the central figure in the population debates of the 19th and 20th centuries.

Malthus and his successors. In 1798 Malthus published *An Essay on the Principle of Population as It Affects the Future Improvement of Society, with Remarks on the Speculations of Mr. Godwin, M. Condorcet, and Other Writers.* This hastily written pamphlet had as its principal object the refutation of the views of the utopians. In Malthus' view, the perfection of a human society free of coercive restraints was a mirage, because the capacity for the threat of population growth would always be present. In this, Malthus echoed the much earlier arguments of Robert Wallace in his *Various Prospects of Mankind, Nature, and Providence* (1761), which posited that the perfection of society carried with it the seeds of its own destruction, in the stimulation of population growth such that "the earth would at last be overstocked, and become unable to support its numerous inhabitants."

Critics of Malthus

Not many copies of Malthus' essay, his first, were published, but it nonetheless became the subject of discussion and attack. The essay was cryptic and poorly supported by empirical evidence. Malthus' arguments were easy to misrepresent, and his critics did so routinely.

The criticism had the salutary effect of stimulating Malthus to pursue the data and other evidence lacking in his first essay. He collected information on one country that had plentiful land (the United States) and estimated that its population was doubling in less than 25 years. He attributed the far lower rates of European population growth to "preventive checks," giving special emphasis to the characteristic late marriage pattern of western Europe, which he called "moral restraint." The other preventive checks to which he alluded were birth control, abortion, adultery, and homosexuality, all of which as an Anglican minister he considered immoral.

In one sense, Malthus reversed the arguments of the mercantilists that the number of people determined the nation's resources, adopting the contrary argument of the Physiocrats that the resource base determined the numbers of people. From this he derived an entire theory of society and human history, leading inevitably to a set of provocative prescriptions for public policy. Those societies that ignored the imperative for moral restraint—delayed marriage and celibacy for adults until they were economically able to support their children—would suffer the deplorable "positive checks" of war, famine, and epidemic, the avoidance of which should be every society's goal. From this humane concern about the sufferings from positive checks arose Malthus' admonition that poor laws (*i.e.,* legal measures that provided relief to the poor) and charity must not cause their beneficiaries to relax their moral restraint or increase their fertility, lest such humanitarian gestures become perversely counterproductive.

Having stated his position, Malthus was denounced as a reactionary, although he favoured free medical assistance for the poor, universal education at a time that this was a radical idea, and democratic institutions at a time of elitist alarums about the French Revolution. Malthus was accused of blasphemy by the conventionally religious. The strongest denunciations of all came from Marx and his followers (see below). Meanwhile, the ideas of Malthus had important effects upon public policy (such as reforms in the English Poor Laws) and upon the ideas of the classical and neoclassical economists, demographers, and evolutionary biologists, led by Charles Darwin. Moreover, the evidence and analyses produced by Malthus dominated scientific discussion of population during his lifetime; indeed, he was the invited author of the article "Population" for the supplement (1824) to the fourth, fifth, and sixth editions of the *Encyclopædia Britannica.* Though many of Malthus' gloomy predictions have proved to be misdirected, that article introduced analytical methods that clearly anticipated demographic techniques developed more than 100 years later.

Neo-Malthusians

The latter-day followers of Malthusian analysis deviated significantly from the prescriptions offered by Malthus. While these "neo-Malthusians" accepted Malthus' core propositions regarding the links between unrestrained fertility and poverty, they rejected his advocacy of delayed marriage and his opposition to birth control. Moreover, leading neo-Malthusians such as Charles Bradlaugh and Annie Besant could hardly be described as reactionary defenders of the established church and social order. To the contrary, they were political and religious radicals who saw the extension of knowledge of birth control to the lower classes as an important instrument favouring social equality. Their efforts were opposed by the full force of the establishment, and both spent considerable time on trial and in jail for their efforts to publish materials—condemned as obscene—about contraception.

Marx, Lenin, and their followers. While both Karl Marx and Malthus accepted many of the views of the classical economists, Marx was harshly and implacably critical of Malthus and his ideas. The vehemence of the assault was remarkable. Marx reviled Malthus as a "miserable parson" guilty of spreading a "vile and infamous doctrine, this repulsive blasphemy against man and nature." For Marx, only under capitalism does Malthus' dilemma of resource limits arise. Though differing in many respects from the utopians who had provoked Malthus' rejoinder, Marx shared with them the view that any number of people could be supported by a properly organized society. Under the socialism favoured by Marx, the surplus product of labour, previously appropriated by the capitalists, would be returned to its rightful owners, the workers, thereby eliminating the cause of poverty. Thus Malthus and Marx shared a strong concern about the plight of the poor, but they differed sharply as to how it should be improved. For Malthus the solution was individual responsibility as to marriage and childbearing; for Marx the solution was a revolutionary assault upon the organization of society, leading to a collective structure called socialism.

The strident nature of Marx's attack upon Malthus' ideas may have arisen from his realization that they constituted a potentially fatal critique of his own analysis. "If [Malthus'] theory of population is correct," Marx wrote in 1875 in his *Critique of the Gotha Programme* (published by Engels in 1891), "then I can*not* abolish this [iron law of wages] even if I abolish wage-labor a hundred times, because this law is not only paramount over the system of wage-labor but also over every social system."

Stalin's pronatalism

The anti-Malthusian views of Marx were continued and extended by Marxians who followed him. For example, although in 1920 Lenin legalized abortion in the revolutionary Soviet Union as the right of every woman "to control her own body," he opposed the practice of contraception or abortion for purposes of regulating population growth. Lenin's successor, Joseph Stalin, adopted a pronatalist argument verging on the mercantilist, in which population growth was seen as a stimulant to economic progress. As the threat of war intensified in Europe in the 1930s, Stalin promulgated coercive measures to increase Soviet population growth, including the banning of abortion despite its status as a woman's basic right. Although contraception

is now accepted and practiced widely in most Marxist-Leninist states, some traditional ideologists continue to characterize its encouragement in Third-World countries as shabby Malthusianism.

The Darwinian tradition. Charles Darwin, whose scientific insights revolutionized 19th-century biology, acknowledged an important intellectual debt to Malthus in the development of his theory of natural selection. Darwin himself was not much involved in debates about human populations, but many who followed in his name as "social Darwinists" and "eugenicists" expressed a passionate if narrowly defined interest in the subject.

In Darwinian theory the engine of evolution is differential reproduction of different genetic stocks. The concern of many social Darwinists and eugenicists was that fertility among those they considered the superior human stocks was far lower than among the poorer—and, in their view, biologically inferior—groups, resulting in a gradual but inexorable decline in the quality of the overall population. While some attributed this lower fertility to deliberate efforts of people who needed to be informed of the dysgenic effects of their behaviour, others saw the fertility decline itself as evidence of biological deterioration of the superior stocks. Such simplistic biological explanations attracted attention to the socioeconomic and cultural factors that might explain the phenomenon and contributed to the development of the theory of the demographic transition.

Theory of the demographic transition. The classic explanation of European fertility declines arose in the period following World War I and came to be known as demographic transition theory. (Formally, transition theory is a historical generalization and not truly a scientific theory offering predictive and testable hypotheses.) The theory arose in part as a reaction to crude biological explanations of fertility declines; it rationalized them in solely socioeconomic terms, as consequences of widespread desire for fewer children caused by industrialization, urbanization, increased literacy, and declining infant mortality.

Changing role of the family

The factory system and urbanization led to a diminution in the role of the family in industrial production and a reduction of the economic value of children. Meanwhile, the costs of raising children rose, especially in urban settings, and universal primary education postponed their entry into the work force. Finally, the lessening of infant mortality reduced the number of births needed to achieve a given family size. In some versions of transition theory, a fertility decline is triggered when one or more of these socioeconomic factors reach certain threshold values.

Until the 1970s transition theory was widely accepted as an explanation of European fertility declines, although conclusions based on it had never been tested empirically. More recently careful research on the European historical experience has forced reappraisal and refinement of demographic transition theory. In particular, distinctions based upon cultural attributes such as language and religion, coupled with the spread of ideas such as those of the nuclear family and the social acceptability of deliberate fertility control, appear to have played more important roles than were recognized by transition theorists.

TRENDS IN WORLD POPULATION

Before considering modern population trends separately for developing and industrialized countries, it is useful to present an overview of older trends. It is generally agreed that only 5,000,000–10,000,000 humans (*i.e.*, one one-thousandth of the present world population) were supportable before the agricultural revolution of about 10,000 years ago. By the beginning of the Christian era, 8,000 years later, the human population approximated 300,000,000, and there was apparently little increase in the ensuing millennium up to the year AD 1000. Subsequent population growth was slow and fitful, especially given the plague epidemics and other catastrophes of the Middle Ages. By 1750, conventionally the beginning of the Industrial Revolution in Britain, world population may have been as high as 800,000,000. This means that in the 750 years from 1000 to 1750, the annual population growth rate averaged only about one-tenth of 1 percent.

The reasons for such slow growth are well known. In the absence of what is now considered basic knowledge of sanitation and health (the role of bacteria in disease, for example, was unknown until the 19th century), mortality rates were very high, especially for infants and children. Only about half of newborn babies survived to the age of five years. Fertility was also very high, as it had to be to sustain the existence of any population under such conditions of mortality. Modest population growth might occur for a time in these circumstances, but recurring famines, epidemics, and wars kept long-term growth close to zero.

Modern increase in world population

From 1750 onward population growth accelerated. In some measure this was a consequence of rising standards of living, coupled with improved transport and communication, which mitigated the effects of localized crop failures that previously would have resulted in catastrophic mortality. Occasional famines did occur, however, and it was not until the 19th century that a sustained decline in mortality took place, stimulated by the improving economic conditions of the Industrial Revolution and the growing understanding of the need for sanitation and public health measures.

The world population, which did not reach its first 1,000,000,000 until about 1800, added another 1,000,000,000 persons by 1930. (To anticipate further discussion below, the third was added by 1960, the fourth by 1974, and the fifth before 1990; see Figure 3.) The most rapid growth in the 19th century occurred in Europe and North America, which experienced gradual but eventually dramatic declines in mortality. Meanwhile, mortality and fertility remained high in Asia, Africa, and Latin America.

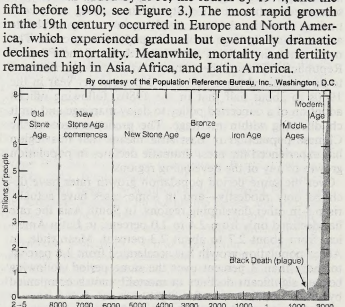


Figure 3: World population growth since 8000 BC.

Beginning in the 1930s and accelerating rapidly after World War II, mortality went into decline in much of Asia and Latin America, giving rise to a new spurt of population growth that reached rates far higher than any previously experienced in Europe. The rapidity of this growth, which some described as the "population explosion," was due to the sharpness in the falls in mortality that in turn were the result of improvements in public health, sanitation, and nutrition that were mostly imported from the developed countries. The external origins and the speed of the declines in mortality meant that there was little chance that they would be accompanied by the onset of a decline in fertility. In addition, the marriage patterns of Asia and Latin America were (and continue to be) quite different from those of Europe; marriage in Asia and Latin America is early and nearly universal, while that in Europe is usually late and significant percentages of people never marry.

These high growth rates occurred in populations already of very large size, meaning that global population growth became very rapid both in absolute and in relative terms. The peak rate of increase was reached in the early 1960s, when each year the world population grew by about 2 percent, or about 68,000,000 people. Since that time both mortality and fertility rates have decreased, and the annual growth rate has fallen moderately, to about 1.7 percent. But even this lower rate, because it applies to a larger population base, means that the number of people added each year has risen from about 68,000,000 to 80,000,000.

The developing countries since 1950. After World War

II there was a rapid decline in mortality in much of the developing world. In part this resulted from wartime efforts to maintain the health of armed forces from industrialized countries fighting in tropical areas. Since all people and governments welcome proven techniques to reduce the incidence of disease and death, these efforts were readily accepted in much of the developing world, but they were not accompanied by the kinds of social and cultural changes that had occurred earlier and had led to fertility declines in industrialized countries.

The reduction in mortality, unaccompanied by a reduction in fertility, had a simple and predictable outcome: accelerating population growth. By 1960 many developing countries had rates of increase as high as 3 percent a year, exceeding by two- or threefold the highest rates ever experienced by European populations. Since a population increasing at this rate will double in only 23 years, the populations of such countries expanded dramatically. In the 25 years between 1950 and 1975, the population of Mexico increased from 27,000,000 to 60,000,000; Iran from 14,000,000 to 33,000,000; Brazil from 53,000,000 to 108,000,000; and China from 554,000,000 to 933,000,000.

The greatest population growth rates were reached in Latin America and in Asia during the mid- to late 1960s. Since then, these regions have experienced variable but sometimes substantial fertility declines along with continuing mortality declines, resulting in usually moderate and occasionally large declines in population growth. The most dramatic declines have been those of the People's Republic of China, where the growth rate was estimated to have declined from well over 2 percent per year in the 1960s to about half that in the 1980s, following official adoption of a concerted policy to delay marriage and limit childbearing within marriage. The predominance of the Chinese population in East Asia means that this region has experienced the most dramatic declines in population growth of any of the developing regions.

Population control in China

Over the same period population growth rates have declined only modestly—and in some cases have actually risen—in other developing regions. In South Asia the rate has declined only from 2.4 to 2.0 percent; in Latin America, from about 2.7 to about 2.3 percent. Meanwhile, in Africa population growth has accelerated from 2.6 percent to more than 3 percent over the same period, following belated significant declines in mortality not accompanied by similar reductions in fertility.

The industrialized countries since 1950. For many industrialized countries, the period after World War II was marked by a "baby boom." One group of four countries in particular—the United States, Canada, Australia, and New Zealand—experienced sustained and substantial rises in fertility from the depressed levels of the prewar period. In the United States, for example, fertility rose by two-thirds, reaching levels in the 1950s not seen since 1910.

A second group of industrialized countries, including most of western Europe and some eastern European countries (notably Czechoslovakia and East Germany), experienced what might be termed "baby boomlets." For a few years after the war, fertility increased as a result of marriages and births deferred during wartime. These increases were modest and relatively short-lived, however, when compared with those of the true baby-boom countries mentioned above. In many of these European countries fertility had been very low in the 1930s; their postwar baby boomlets appeared as three- to four-year "spikes" in the graph of their fertility rates, followed by two full decades of stable fertility levels. Beginning in the mid-1960s, fertility levels in these countries began to move lower again and, in many cases, fell to levels comparable to or lower than those of the 1930s.

A third group of industrialized countries, consisting of most of eastern Europe along with Japan, showed quite different fertility patterns. Most did not register low fertility in the 1930s but underwent substantial declines in the 1950s after a short-lived baby boomlet. In many of these countries the decline persisted into the 1960s, but in some it was reversed in response to governmental incentives.

By the 1980s the fertility levels in most industrialized countries were very low, at or below those needed to maintain stable populations. There are two reasons for this phenomenon: the postponement of marriage and childbearing by many younger women who entered the labour force, and a reduction in the numbers of children born to married women.

POPULATION PROJECTIONS

Demographic change is inherently a long-term phenomenon. Unlike populations of insects, human populations have rarely been subject to "explosion" or "collapse" in numbers. Moreover, the powerful long-term momentum that is built into the human age structure means that the effects of fertility changes become apparent only in the far future. For these and other reasons, it is by now conventional practice to employ the technology of population projection as a means of better understanding the implications of trends.

Population projections represent simply the playing out into the future of a set of assumptions about future fertility, mortality, and migration rates. It cannot be stated too strongly that such projections are not predictions, though they are misinterpreted as such frequently enough. A projection is a "what-if" exercise based on explicit assumptions that may or may not themselves be correct. As long as the arithmetic of a projection is done correctly, its utility is determined by the plausibility of its central assumptions. If the assumptions embody plausible future trends, then the projection's outputs may be plausible and useful. If the assumptions are implausible, then so is the projection. Because the course of demographic trends is hard to anticipate very far into the future, most demographers calculate a set of alternative projections that, taken together, are expected to define a range of plausible futures, rather than to predict or forecast any single future. Because demographic trends sometimes change in unexpected ways, it is important that all demographic projections be updated on a regular basis to incorporate new trends and newly developed data.

A standard set of projections for the world and for its constituent countries is prepared every two years by the Population Division of the United Nations. These projections include a low, medium, and high variant for each country and region. To illustrate the nature of such projected futures, Figure 4 presents these alternatives for the world as a whole and for its developed and developing regions.

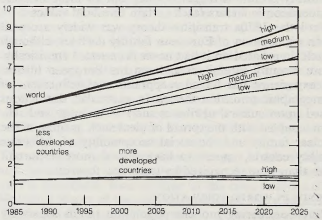


Figure 4: Projected population of the world, less developed countries, and more developed countries (high, medium, and low variants, AD 1985–2025).

BIBLIOGRAPHY

General works: ROLAND PRESSAT, *The Dictionary of Demography* (1985; originally published in French, 1979); JOHN A. ROSS (ed.), *International Encyclopedia of Population*, 2 vol. (1982), a comprehensive reference work that contains articles on topics ranging from classic demography to current problems and that provides coverage of world regions and countries as well as organizations and agencies active in the field; PETER R. COX, *Demography*, 5th ed. (1976), an examination of methods used in the study of population; DENNIS H. WRONG, *Population and Society*, 4th ed. (1977), an introduction to the main aspects of the population dilemma; and WARREN S. THOMPSON and DAVID T. LEWIS, *Population Problems*, 5th ed. (1965), a comprehensive sociological study.

Population change: ARTHUR A. CAMPBELL, *Manual of Fertility Analysis* (1983), a methodological study: GERRY E. HENDERSHOT and PAUL J. PLACEK (eds.), *Predicting Fertility: Demographic Studies of Birth Expectations* (1981), a survey of the concepts, knowledge, and methods of fertility; NORMAN E. HIMES, *Medical History of Contraception* (1936, reissued 1970), a classic historical treatise; JOHN BONGAARTS and ROBERT G. POTTER, *Fertility, Biology, and Behavior: An Analysis of the Proximate Determinants* (1983), a scholarly examination of such aspects of the population problem as family planning, family size, birth intervals, and fertility; ANSLEY J. COALE and SUSAN COTTS WATKINS (eds.), *The Decline of Fertility in Europe* (1985), a summary of studies of fertility transition in Europe during the 1970s; MICHAEL S. TEITELBAUM and JAY M. WINTER, *The Fear of Population Decline* (1985), the history and bases of past and current concerns about low fertility, with special emphasis on major Western countries; PRANAY GUPTE, *The Crowded Earth: People and the Politics of Population* (1984), a study of the effects of Western development programs on overpopulation; SAMUEL H. PRESTON, *Mortality Patterns in National Populations: With Special Reference to Recorded Causes of Death* (1976), a discussion of the determinants and consequences of national mortality patterns, with attention to the role of standards of living, sex differences, and major causes of death; ALAN A. BROWN and EGON NEUBERGER (eds.), *Internal Migration: A Comparative Perspective* (1977), a collection of scholarly articles; *International Migration Policies and Programmes: A World Survey* (1982), a study of immigration and refugee programs, one of the series of population studies conducted by the United Nations; DAVID GRIGG, *Population Growth and Agrarian Change: An Historical Perspective* (1980), a study of the relationship between demography and economics; ELI S. MARKS, WILLIAM SELTZER, and KAROL J. KRÓTKI, *Population Growth Estimation: A Handbook of Vital Statistics Measurement* (1974), an examination of methods for demographic estimates; and NATHAN KEYFITZ, *Population Change and Social Policy* (1982), a compendium of insightful essays. The 1984 issue of the annual *World Development Report* of the World Bank focuses on population changes in underdeveloped countries.

Population composition: COLIN MCEVEDY and RICHARD JONES, *Atlas of World Population History* (1978), comparisons of population statistics; NATHAN KEYFITZ and WILHELM FLIEGER, *World Population: An Analysis of Vital Data* (1968), a statistical analysis of demographic data for different countries; NATHAN KEYFITZ (ed.), *Population and Biology: Bridge Between Disciplines* (1984), a collection of papers on biological aspects of demography; DOROTHY L. NORTMAN, *Population and Family Planning Programs: Compendium of Data Through 1983*, 12th ed. (1985), including an analysis of demographic and social characteristics; DAVID M. HEER, *Society and Population*, 2nd ed. (1975), a brief study of the population of nation-states; PHILIP M. HAUSER *et al.*, *Population and the Urban Future* (1982), which examines various demographic topics, including the quality of life of urbanized populations; RAYMOND F. DASMANN, *Environmental Conservation*, 5th ed. (1984), a study of the human influence on nature and its consequences for human population; and COLIN CLARK, *Population Growth and Land Use*, 2nd ed. (1977), an analysis of specific problems.

Population theories: RONALD FREEDMAN (ed.), *Population: The Vital Revolution* (1964), is a collection of authoritative essays on major aspects of contemporary demographic analysis. The history of population theory is presented in CHARLES EMIL STANGELAND, *Pre-Malthusian Doctrines of Population: A Study in the History of Economic Theory* (1904, reprinted 1967); JOSEPH J. SPENGLER, *French Predecessors of Malthus: A Study in Eighteenth-Century Wage and Population Theory* (1942, reprinted 1980); WILLIAM PETERSEN, *Malthus* (1979); PATRICIA JAMES, *Population Malthus, His Life and Times* (1979); E.P. HUTCHINSON, *The Population Debate: The Development of Conflicting Theories up to 1900* (1967); and J. DUPÂQUIER, A. FAUVE-CHAMOUX, and E. GREBENIK (eds.), *Malthus Past and Present* (1983). Other works include ALFRED SAUVY, *General Theory of Population* (1969; originally published in French, 1952–54), a study of the relationship between the demographic-biological characteristics of societies and their economic and social circumstances; PHILIP M. HAUSER (ed.), *The Population Dilemma*, 2nd rev. ed. (1969), a collection of papers by leading theoreticians; GARRETT HARDIN, *Population, Evolution, and Birth Control: A Collage of Controversial Ideas*, 2nd ed. (1969), a survey of the history of views and opinions; JONAS SALK and JONATHAN SALK, *World Population and Human Values: A New Reality* (1981), which traces changes in attitudes to developments in population; and JULIAN L. SIMON, *The Ultimate Resource* (1981), an optimistic perspective for the world's population.

Development trends: The spatial distribution of population characteristics is studied in GLENN T. TREWARTHA, *The Less Developed Realm: A Geography of Its Population* (1972), and GLENN T. TREWARTHA (ed.), *The More Developed Realm: A Geography of Its Population* (1978). ALFRED SAUVY, *Fertility and Survival: Population Problems from Malthus to Mao Tse-Tung* (1961; originally published in French, 1959), presents an analysis of problems and proposes solutions. Other works include *The Determinants and Consequences of Population Trends: New Summary of Findings of Interaction of Demographic, Economic, and Social Factors*, 2 vol. (1973–78), a monumental compendium of theory and evidence on population trends compiled by the United Nations; the *Demographic Yearbook*, published by the Statistical Office of the United Nations, a basic source for more than 200 countries and territories; and BERNARD BERELSON (ed.), *Family Planning Programs: An International Survey* (1969), and *Family Planning and Population Programs: A Review of World Developments* (1966), comprehensive assessments of data. Later sources on trends in population development include ESTER BOSERUP, *Population and Technological Change* (1981); GAVIN W. JONES, *Population Growth and Educational Planning in Developing Nations* (1975); JOHN CLELAND and JOHN HOBCRAFT (eds.), *Reproductive Change in Developing Countries: Insights from the World Fertility Survey* (1985); GAVIN W. JONES (ed.), *Demographic Transition in Asia* (1984); C. ALISON MCINTOSH, *Population Policy in Western Europe: Responses to Low Fertility in France, Sweden, and West Germany* (1983); DONALD J. BOGUE, *The Population of the United States: Historical Trends and Future Projections* (1985); JOHN L. ANDRIOT (ed.), *Population Abstracts of the United States* (1980); RICHARD L. RUBENSTEIN, *The Age of Triage: Fear and Hope in an Overcrowded World* (1983); JANE MENKEN (ed.), *World Population & U.S. Policy: The Choices Ahead* (1986), papers by leading experts prepared for the second American Assembly on population issues; and NATIONAL RESEARCH COUNCIL, COMMITTEE ON POPULATION, *Population Growth and Economic Development: Policy Questions* (1986).

Population projections: World Population Prospects: Estimates and Projections as Assessed in 1982 (1985), and *Patterns of Urban and Rural Population Growth* (1980), are publications of the United Nations Department of International Economic and Social Affairs. See also JUST FAALAND (ed.), *Population and the World Economy in the 21st Century* (1982), a collection of papers with economic and demographic forecasts; and MICHAEL D. BAYLES, *Morality and Population Policy* (1980), and *Reproductive Ethics* (1984), which discuss the moral aspects of long-term population policies.

For current discussions of relevant topics and reports of recent research the following periodicals may be recommended: *Contraception* (monthly); *Demography* (quarterly); *Economic Development and Cultural Change* (quarterly); *Economic History Review* (quarterly); *Family Planning Perspectives* (bimonthly); *Human Biology* (quarterly); *International Labour Review* (bimonthly); *International Migration Review* (quarterly); *Journal of Biosocial Science* (quarterly); *Journal of Economic History* (quarterly); *Journal of Interdisciplinary History* (quarterly); *Monthly Labor Review*; *Population and Development Review* (quarterly); *Population Bulletin* (quarterly); *Population Index* (quarterly); *Population Studies* (three times a year); *Science* (weekly); and *Social Biology* (quarterly). Many non-English periodicals in the field appear with systematic summaries in English: *Annales: Économies, Sociétés, Civilisations* (France, bimonthly); *Demografía* (Hungary, quarterly); *Demografía y economía* (Mexico, quarterly); *Demografie* (Czechoslovakia, quarterly); *Genus* (Italy, weekly); *Journal of Population Problems* (Japan, quarterly); and *Population* (France, bimonthly).

(M.S.T.)

Portugal

One of the oldest countries in Europe, with established frontiers since the 13th century, Portugal lies on the extreme western rim of the European continent. It is bordered on the east and north by Spain and on the west and south by the Atlantic Ocean. Occupying about 16 percent of the Iberian Peninsula, Portugal has a total area of 35,672 square miles (92,389 square kilometres). Despite its small size, the country displays a great diversity of geographic features; the land rises to 6,532 feet (1,991 metres) at its highest point in the Serra da Estrela (Estrela Mountains). To the west and southwest lie the Atlantic islands of the Azores and Madeira, which are part of metropolitan Portugal. Until the 1970s, Portuguese overseas territories included the Cape Verde Islands, São Tomé and Príncipe, Portuguese Guinea, Angola, and Mozambique, all in Africa; Macau in Asia; and Timor in Oceania. Of these, by the late 20th century only Macau (Macao) remained a Portuguese dependency.

As a member of the European Community (EC) and the North Atlantic Treaty Organization (NATO), Portugal plays a greater role in both European and world affairs than its size would suggest. Nonetheless, it is one of the poorest countries in western Europe.

The article is divided into the following sections:

Physical and human geography 1044
- The land 1044
  - Relief
  - Drainage
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 1048
- The economy 1049
  - Agriculture, forestry, and fishing
  - Mining and industry
  - Banking and finance
  - Trade
  - Transportation
- Government and social conditions 1050
  - Government
  - Education
  - Welfare services
- Cultural life 1051
  - Daily life
  - Architecture
  - Visual and decorative arts
  - Music
  - Literature
  - Theatre and motion pictures
  - Museums, libraries, and art galleries
  - Recreation
  - Press and broadcasting

History 1052
- Pre-Roman, Roman, Germanic, and Muslim periods 1052
- The county and kingdom of Portugal to 1383 1053
  - The kingdom and the Reconquest
  - Disputes with Castile
- The house of Aviz, 1383–1580 1054
  - Independence assured
  - Consolidation of the monarchy
- Medieval social and economic development 1055
- The discoveries and the empire 1056
  - Conquest and exploration
  - Control of the sea trade
- Union of Spain and Portugal, 1580–1640 1056
- The house of Bragança, 1640–1910 1057
  - The 18th century
  - The French Revolutionary and Napoleonic Wars
  - Constitutionalism
  - Social and economic conditions
- The First Republic, 1910–26 1059
- The dictatorship, 1926–74 1060
  - The Salazar regime
  - The New State after Salazar
- Portugal since 1974 1061
  - The revolution
  - The 1976 constitution and subsequent reforms

Bibliography 1062

Physical and human geography

THE LAND

Portugal occupies one-sixth of the Iberian Peninsula along its western (Atlantic) side. Thus, even more than Spain, which accounts for the rest, Portugal lies much farther to the west of Europe than is usually remembered. With Spanish Galicia, northern Portugal comprises the mountainous border of the Meseta (the block of ancient rock that forms the core of the Iberian Peninsula), while southern Portugal also contains extensive areas of limestone and other sedimentary strata, mostly plateaus or plains. Other physical features link Portugal with Spain: the major rivers (Douro, Tagus, Guadiana) rise in the central Meseta before draining west (or, in the case of the Guadiana, south) to the Atlantic, while the proximity of the Meseta gives a continental nuance to the climate of the northern Portuguese interior and to that region's vegetation. Southern Portugal, however, is predominantly Mediterranean both in vegetation and climate. Despite Portugal's remarkable scenic diversity, the essence of its relief and underlying geology can be described under three headings: the north, the northern interior, and the south. The coastal provinces of Beira Litoral and Estremadura are transitional in cultural landscape, vegetation, and climate but southern in relief and geology.

Physiographic regions

Relief. Only 11.6 percent of Portugal rises above 2,300 feet. Most of the mountains are north of the Tagus River (Rio Tejo), which, flowing northeast to southwest, divides the country. North of the Tagus, more than 90 percent of the land rises above 1,300 feet; in the south, only one range is higher than 3,200 feet.

North. In the northwest the mountains of the Minho province, surmounted by the Serra do Larouco (5,003 feet; Larouco Mountains), form an amphitheatre facing the Atlantic Ocean. The area is composed of metamorphic rocks (crystalline schists, slates, and quartzites) and intrusive or granitic rocks, marked by fault lines that give rise to hot springs. Rolling plateaus and deeply entrenched streams identify the relatively fertile granitic regions; the schist outcrops offer poor soils. Between the ridges, rivers such as the Lima and Cávado flow through deep gorges or in flat-floored valleys. The narrow coastal strip is backed by hills that rise steeply to 1,000 to 2,000 feet.

Northern interior. Behind the mountains of Minho is the province of Trás-os-Montes e Alto Douro, bordered on two sides by Spain. South of the Douro, also the western extension of the Spanish Meseta, are the provinces of Beira Alta and Beira Baixa. In the north of this region are high plateaus (*terra fria*, or "cold country," at 2,000 to 2,600 feet), heavily faulted and composed largely of ancient Precambrian rock (570 million to 3.9 billion years old), such as granites, schists, and slates. They form a rolling topography. The rivers follow lines of structural weakness and run northwest-southeast. After rejuvenation in Pliocene-Quaternary times (5.3 million years ago to the present), they now flow in deep valleys and canyons that may be as much as 1,600 feet deep. The monadnock of

Serra de Nogueira (4,324 feet; Nogueira Mountains) rises above the general plateau level. A similar pattern of relief and geology is found south of the Douro, but the sheltered valleys of dark-coloured schists trap the heat of a Mediterranean climate, creating a *terra quente* ("hot country"). The high plain of Beira is bounded on the northwest by the Serra do Caramulo (3,514 feet; Caramulo Mountains) and the Serra de Montemuro (4,534 feet; Montemuro Mountains) and their foothills; it is flanked to the south and southeast by the granite and schist escarpment of Serra da Estrela (6,532 feet), which, together with the lower Serra da Guardunha (3,920 feet; Guardunha Mountains), represents the continuation in Portugal of the Central Sierras of Spain. Between the two ridges, the Guarda Plateau has an average altitude of 3,000 feet. The mountain chain extends southwestward to the Serra do Açor (4,632 feet; Açor Mountains) and the Serra da Lousã (3,950 feet; Lousã Mountains). Farther south the eroded plateau of Beira Baixa drops from an altitude of 1,600 feet to 700 feet toward the Tagus River. It merges almost imperceptibly into the highlands of Alto Alentejo.

South. Some 63 percent of Portugal's land below 1,300 feet is found in the south. From the lower Douro to Lisbon and the lower Tagus, the plains (mainly Triassic sandstones—208 to 245 million years old) of the western provinces of Beira Litoral and Estremadura give rise to low-lying, sometimes marshy, coasts. Shifting sands and dunes have been stabilized since the 19th century, but lagoons and salt marshes still characterize the river mouths (*e.g.*, Aveiro Lagoon at the mouth of the Vouga River). Inland in Estremadura, east and south of Leiria, Jurassic limestones (144 to 208 million years old), interspersed with large areas of Cretaceous sandstones and conglomerates (66.4 to 144 million years old), have been eroded into rolling sandy hills and steep calcareous escarpments. Ancient volcanic activity has left basalt plateaus. Between the Tagus and the Guadiana, the Alto Alentejo is a continuation of the Spanish tablelands—a series of plateaus, either crystalline (Cambrian and Silurian schists, 505 to 570 million and 408 to 438 million years old, respectively), at 600 to 1,300 feet with poor soils except where outcrops of diorite have weathered into rich black soils, or limestone, with piedmont springs at their foot. In northern Baixo Alentejo, ridges of quartz and marble oriented northwest-southeast account for a monotonously undulating relief between 300 and 600 feet. This terminates in the east with the schistose mountains of the Serra do Caldeirão (1,893 feet; Caldeirão Mountains). Sheltered by the mountains from northern climatic influences lie the more extensive scarps and hills of the Algarve. These are composed of limestones and sandstones of Mesozoic age (66.4 to 245 million years old). The Serra de Monchique (Monchique Mountains), a dissected massif of intrusive igneous rock (syenite), reaches 2,959 feet at Mount Foia.

The Atlantic islands

The Madeira Islands. The Madeira archipelago comprises eight volcanic islands in the Atlantic, 600 miles southwest of the mainland. Only Madeira (307 square miles) and Porto Santo are inhabited. Madeira forms an assymetrical mountainous hump in mid-ocean, rising to 6,109 feet in the interior, falling more steeply in the north. Thick layers of basalt alternating with beds of ash and scoriae create a stepped relief that has been deeply dissected into gorges and canyons. The cratorial Curral das Freiras is about 2,265 feet deep; Paúl da Serra is a bleak high-level plateau. The island of Porto Santo is low-lying and flat. The eastern tip of Madeira, like the uninhabited Desertas Islands, is sandy.

The Azores. Nine islands constitute the Azores, which extend, in three groups, over 400 miles in the mid-Atlantic. The easternmost island, Santa Maria, lies 875 miles from the Portuguese mainland; the westernmost, Flores, is 1,232 miles from Cape Race, Nfd., Can. The islands are volcanic, occasionally active. They have a varied scenery, with crater lakes, plateaus at various altitudes, mountainous massifs, flat-bottomed valleys, and rugged coastlines.

Drainage. All the main rivers of Portugal flow from Spain over the edge of the Meseta in a series of defiles (narrow gorges), and their usefulness, either for navigation or as routeways for roads and railways, is thus limited. The longest river crossing Portugal is the Douro (200 miles); in Alto Douro the riverbed drops 16 feet per mile in gorges 90 to 160 feet deep. The longest river wholly in the country is the Mondego (137 miles), which rises in the Serra da Estrela. Other mainly Portuguese rivers include the Vouga, Sado, and Zêzere (tributary of the Tagus). Like the Mondego, all are navigable for short distances. Rich silt land (campo) in the lower Tagus valley is the result of regular flooding, which is especially severe when strong southerly gales drive high seas up the estuary. The Guadiana, which flows south into the Gulf of Cadiz, forms part of the frontier with Spain, as does the Minho in the north.

Major harbours

Portugal has more than 500 miles of coastline, four-fifths of which faces westward. Except at the mouths of the larger rivers, there are few major indentations or natural harbours; the most important are those of Lisbon, on the Tagus, and Setúbal, on the Sado. The entrance to the Tagus is a long, narrow, deepwater channel opening out into a broad expanse of inland water. Other harbours depend on the protection of headlands (*e.g.*, the artificial harbours of Leixões and Sines).

Climate. Climate, through its effect on vegetation, divides Portugal. As in Spain, three sets of influences are involved: Atlantic, continental (Mesetan), and Mediterranean. The former predominates overall, putting most of the country into the humid zone of the Iberian Peninsula, but this is especially true in the northwest, where it gives a mild, rainy climate. Summer temperatures near sea level may average up to 76° F (24° C) but are rather lower at higher, exposed altitudes. Winter temperatures average 37° to 40° F (3° to 4° C) but tend to be milder south of the Douro. Annual rainfall averages more than 40 inches (1,000 millimetres).

In the extreme northwest, much of Minho receives 40 to 80 inches, with more than 100 inches falling on mountain slopes. Inland, lee slopes are arid, receiving at best 20 inches, much of which is lost through high evaporation rates. In the interior, continental influences increase the duration of the summer drought to more than a month and intensify winter severity. High-pressure conditions from the Spanish Meseta or from Siberian anticyclones bring very cold temperatures. In the highest areas of the Bragança and of the Serra da Estrela, temperatures drop to 32° F (0° C), and snow remains on the summits for several months. In the south, where the Azores high-pressure system prevails in summer, the period of drought lengthens to two months. Mean temperatures may reach 75° F (24° C) in summer and 52° F (11° C) in winter. Average annual rainfall is 20–23 inches along the coast but slightly higher in the mountains of the Algarve.

In the Madeiras, Mediterranean, Saharan, and oceanic influences produce seasonal rainfall, occasional dry winds, and thermal equability, respectively. Climate varies markedly with altitude. In the Azores, the anticyclone dominates, though conditions can be highly variable. Rainfall, for instance, is irregular both in annual total and in regime: Horta (on Faial Island) may have more than 40 inches a year or suffer severe drought.

Regional flora

Plant and animal life. *Vegetation.* The vegetation of Portugal is a mixture of Atlantic, or European, and Mediterranean (with some African) species. Overall, the former accounts for some two-thirds of species, but the regional distribution is revealing. North of the Mondego valley, 57 percent of the plants are European species (more than 86 percent in the northern interior), and only 26 percent Mediterranean; in the south the proportions are 29 and 46 percent, respectively. A third are foreign species, introduced in periods of colonization.

Millennia of human activity have left the country with only a quarter of its area under woodland; the remainder carries scrub, *matorral,* or steppe. Mixed deciduous wood is confined to the north and northern interior, where the landscapes of Minho are lush and green except for the heaths (*mato*) of the Cambrian schists. These carry erica, heather, cistus, and bracken. The original oak climax (with *Quercus robur* as the dominant species) has been largely replaced by maritime pine with some cork oak (*Q. suber*) and, since the 19th century, eucalyptus. Olives, extended into the north in Roman times, are limited to altitudes



MAP INDEX

Political subdivisions

Aveiro . . . . . 40 45 N 8 30 W
Azores (Açores) . . . . . 38 30 N 28 00 W
Beja . . . . . 37 58 N 7 50 W
Braga . . . . . 41 35 N 8 25 W
Bragança . . . . . 41 30 N 6 45 W
Castelo Branco . . . . . 40 00 N 7 30 W
Coimbra . . . . . 40 12 N 8 25 W
Évora . . . . . 38 35 N 7 50 W
Faro . . . . . 37 12 N 8 10 W
Guarda . . . . . 40 40 N 7 10 W
Leiria . . . . . 39 40 N 8 50 W
Lisboa . . . . . 39 00 N 9 08 W
Madeira . . . . . 32 44 N 17 00 W
Portalegre . . . . . 39 15 N 7 35 W
Porto . . . . . 41 15 N 8 20 W
Santarém . . . . . 39 15 N 8 35 W
Setúbal . . . . . 38 20 N 8 30 W
Viana do Castelo . . . . . 41 55 N 8 25 W
Vila Real . . . . . 41 35 N 7 35 W
Viseu . . . . . 40 45 N 7 50 W

Cities and towns

Abrantes . . . . . 39 28 N 8 12 W
Águas Santas . . . . . 41 12 N 8 35 W
Águeda . . . . . 40 34 N 8 27 W
Albufeira . . . . . 37 05 N 8 15 W
Alcains . . . . . 39 55 N 7 27 W
Alcobaça . . . . . 39 33 N 8 59 W
Aljustrel . . . . . 37 52 N 8 10 W
Almada . . . . . 38 41 N 9 09 W
Almeirim . . . . . 39 12 N 8 38 W
Alpiarça . . . . . 39 15 N 8 35 W
Alverca . . . . . 38 54 N 9 02 W
Amadora . . . . . 38 45 N 9 14 W
Angra do Heroísmo (Angra) . . . . . 38 39 N 27 13 W
Aveiro . . . . . 40 38 N 8 39 W
Barcelos . . . . . 41 32 N 8 37 W
Barreiro . . . . . 38 40 N 9 04 W
Batalha . . . . . 39 39 N 8 50 W
Beja . . . . . 38 01 N 7 52 W
Belém (Santa Maria de Belém) . . . . . 38 42 N 9 12 W
Borba . . . . . 38 48 N 7 27 W
Braga . . . . . 41 33 N 8 26 W
Bragança . . . . . 41 49 N 6 45 W
Buarcos . . . . . 40 10 N 8 53 W
Caldas da Rainha . . . . . 39 24 N 9 08 W
Campo Maior . . . . 39 01 N 7 04 W
Cantanhede . . . . . 40 21 N 8 36 W
Cartaxo . . . . . 39 09 N 8 47 W
Cascais (Cascaes) . . . . . 38 42 N 9 25 W
Castelo Branco . . . . 39 49 N 7 30 W
Chaves . . . . . 41 44 N 7 28 W
Cintra, see Sintra
Coimbra . . . . . 40 12 N 8 25 W
Covilhã . . . . . 40 17 N 7 30 W

Cuba ..... 38 10 N 7 53 W
Elvas ..... 38 53 N 7 10 W
Entroncamento ... 39 28 N 8 28 W
Espinho ..... 41 01 N 8 38 W
Estremoz ..... 38 51 N 7 35 W
Évora ..... 38 34 N 7 54 W
Fafe ..... 41 27 N 8 10 W
Faro ..... 37 01 N 7 56 W
Fátima ..... 39 37 N 8 39 W
Feira ..... 40 55 N 8 33 W
Ferreira do Alentejo ..... 38 03 N 8 07 W
Figueira da Foz ... 40 09 N 8 52 W
Freixo de Espada à Cinta .. 41 05 N 6 49 W
Funchal ..... 32 38 N 16 54 W
Fundão ..... 40 08 N 7 30 W
Gondomar ..... 41 09 N 8 32 W
Gouveia ..... 40 30 N 7 36 W
Guarda ..... 40 32 N 7 16 W
Guimarães ..... 41 27 N 8 18 W
Horta ..... 38 32 N 28 38 W
Ílhavo ..... 40 36 N 8 40 W
Lagoa ..... 37 44 N 25 34 W
Lagos ..... 37 06 N 8 40 W
Lajes ..... 38 23 N 28 16 W
Lajes ..... 38 46 N 27 06 W
Lamego ..... 41 06 N 7 49 W
Leiria ..... 39 45 N 8 48 W
Lisbon (Lisboa) .. 38 43 N 9 08 W
Loulé ..... 37 08 N 8 02 W
Lourosa ..... 40 59 N 8 33 W
Macedo de Cavaleiros ..... 41 32 N 6 58 W
Mafra ..... 38 56 N 9 20 W
Mangualde ..... 40 36 N 7 46 W
Matosinhos (Matozinhos) .... 41 11 N 8 42 W
Mirandela ..... 41 29 N 7 11 W
Mogadouro ..... 41 20 N 6 43 W
Montemor-o-Novo ..... 38 39 N 8 13 W
Montijo ..... 38 42 N 8 58 W
Moreira ..... 41 14 N 8 39 W
Moura ..... 38 08 N 7 27 W
Nazaré ..... 39 36 N 9 04 W
Odivelas ..... 38 47 N 9 11 W
Oeiras ..... 38 41 N 9 19 W
Olhão ..... 37 02 N 7 50 W
Oliveira de Azeméis ..... 40 50 N 8 29 W
Oliveira do Douro ..... 41 05 N 8 02 W
Oporto, see Porto
Ovar ..... 40 52 N 8 38 W
Penafiel ..... 41 12 N 8 17 W
Peniche ..... 39 21 N 9 23 W
Peso da Régua ..... 41 10 N 7 47 W
Pias ..... 38 01 N 7 29 W
Pombal ..... 39 55 N 8 38 W
Ponta Delgada ... 37 44 N 25 40 W
Ponte de Lima .... 41 46 N 8 35 W
Portalegre ..... 39 17 N 7 26 W
Portimão (Vila Nova de Portimão) ..... 37 08 N 8 32 W
Porto (Oporto) .... 41 09 N 8 37 W
Póvoa de Varzim .. 41 23 N 8 46 W
Praia da Vitória ... 38 44 N 27 04 W
Quarteira ..... 37 04 N 8 06 W
Queluz ..... 38 45 N 9 15 W
Redondo ..... 38 39 N 7 33 W
Regungos [de Monsaraz] .. 38 25 N 7 32 W
Ribeira Grande ... 37 49 N 25 31 W
Rio Maior ..... 39 20 N 8 56 W
Sacavém ..... 38 47 N 9 06 W
Santa Cruz [da Graciosa] ... 39 05 N 28 01 W
Santa Cruz [das Flores] .... 39 27 N 31 07 W
Santarém ..... 39 14 N 8 41 W
São João da Madeira ..... 40 54 N 8 30 W
São Mamede de Infesta ..... 41 12 N 8 36 W
Sesimbra ..... 38 26 N 9 06 W
Setúbal ..... 38 32 N 8 54 W
Sines ..... 37 57 N 8 52 W
Sintra (Cintra) .... 38 48 N 9 23 W
Tavira ..... 37 07 N 7 39 W
Tomar ..... 39 36 N 8 25 W
Torre de Moncorvo ..... 41 10 N 7 03 W
Torres Vedras .... 39 06 N 9 16 W
Urgelrfça ..... 40 30 N 7 53 W
Valença ..... 42 02 N 8 38 W
Valpaços ..... 41 36 N 7 19 W
Velas ..... 38 41 N 28 13 W
Vendas Novas .... 38 41 N 8 27 W
Viana do Castelo .. 41 42 N 8 50 W
Vieira de Leiria ... 39 52 N 8 56 W
Vila do Conde .... 41 21 N 8 45 W
Vila do Porto .... 36 56 N 25 09 W
Vila Franca de Xira ..... 38 57 N 8 59 W
Vila Nova de Gaia ..... 41 08 N 8 37 W
Vila Nova de Portimão, see Portimão
Vila Real ..... 41 18 N 7 45 W
Vila Real de Santo Antônio .. 37 12 N 7 25 W
Vila Viçosa ..... 38 47 N 7 25 W
Viseu ..... 40 39 N 7 55 W

Physical features and points of interest

Agueiera Reservoir ..... 40 22 N 8 08 W
Alqueva Reservoir ..... 38 19 N 7 25 W
Alvão Nature Park ..... 41 21 N 7 50 W
Arrábida Nature Park ..... 38 28 N 9 00 W
Atlantic Ocean ... 40 00 N 10 00 W
Ave, *river* ..... 41 20 N 8 45 W
Aveiro Lagoon ..... 40 38 N 8 44 W
Azores (Açores), *islands* ..... 38 30 N 28 00 W
Cabril Reservoir ..... 39 58 N 8 00 W
Cádiz, Gulf of .... 36 50 N 7 10 W
Caldeirão Mountains ..... 37 20 N 8 20 W
Caramulo Mountains ..... 40 34 N 8 11 W
Carvoeiro, Cape ..... 39 21 N 9 24 W
Castelo do Bode Reservoir ..... 39 40 N 8 16 W
Cávado, *river* ..... 41 32 N 8 48 W
Côa, *river* ..... 41 05 N 7 06 W
Coroa, Mount .... 41 55 N 7 00 W
Corvo Island .... 39 42 N 31 06 W
Deserta Islands .. 32 30 N 16 30 W
Douro (Duero), *river* ..... 41 08 N 8 40 W
Espichel, Cape ... 38 25 N. 9 13 W
Estrela, *mountain* ..... 40 19 N 7 37 W
Estrela Mountains ..... 40 20 N 7 38 W
Faial (Fayal) Island ..... 38 34 N 28 42 W
Flores Island ..... 39 26 N 31 13 W
Foia, *peak* ..... 37 19 N 8 36 W
Formosa River Nature Park .... 37 00 N 7 55 W
Gerez Mountains ..... 41 48 N 8 00 W
Graciosa Island ... 39 04 N 28 00 W
Grândola Mountains ..... 38 06 N 8 38 W
Guadiana, *river* ... 37 14 N 7 22 W
Lima (Limia), *river* ..... 41 41 N 8 50 W
Lousã Mountains .. 40 04 N 8 13 W
Madeira Island .... 32 44 N 17 00 W
Madeira Islands ... 32 40 N 16 45 W
Maranhão Reservoir ..... 39 05 N 7 55 W
Minho (Miño), *river* ..... 41 52 N 8 51 W
Mira, *river* ..... 37 43 N 8 47 W
Mogadouro Mountains ..... 41 19 N 6 40 W
Monchique Mountains ..... 37 19 N 8 36 W
Mondego, *river* ..... 40 09 N 8 52 W
Mondego, Cape .. 40 11 N 8 55 W
Montargil Reservoir ..... 39 08 N 8 07 W
Montemuro Mountains ..... 40 58 N 8 01 W
Montezinho Nature Park .... 41 53 N 6 45 W
Nogueira Mountains ..... 41 42 N 6 52 W
Peneda-Gerês National Park .... 41 50 N 8 10 W
Pico, see Ponta do Pico, Mount
Pico Island ..... 38 28 N 28 20 W
Ponta do Pico, Mount ..... 38 28 N 28 25 W
Porto Santo Island ..... 33 04 N 16 20 W
Roca, Cape ..... 38 47 N 9 30 W
Ruivo de Santana, Mount .. 32 45 N 16 56 W
Sabor, *river* ..... 41 10 N 7 07 W
Sado, *river* ..... 38 29 N 8 55 W
Sado Estuary Nature Reserve ..... 38 28 N 8 45 W
Santa Clara Reservoir ..... 37 30 N 8 28 W
Santa Maria, Cape ..... 36 58 N 7 54 W
Santa Maria Island ..... 36 58 N 25 06 W
São Jorge Island .. 38 38 N 28 03 W
São Miguel Island ..... 37 47 N 25 30 W
São Vicente, Cape ..... 37 01 N 9 00 W
Selvagens Islands ..... 30 05 N 15 55 W
Serra da Estrela Nature Park .... 40 25 N 7 30 W
Serra da Malcata Nature Reserve ..... 40 15 N 7 05 W
Serra de São Mamede Nature Park ..... 39 22 N 7 25 W
Serras de Aire e Candeeiros Nature Park ..... 39 30 N 8 47 W
Setúbal Bay ..... 38 27 N 8 53 W
Sintra-Cascais Protected Landscapes .... 38 50 N 9 27 W
Sorraira, *river* .... 38 56 N 8 53 W
Sudoeste Alentejano e Costa Vicentina Protected Landscapes .... 37 30 N 8 48 W
Tagus (Tejo), *river* ..... 38 40 N 9 24 W
Tagus (Tejo) Estuary Nature Reserve ..... 38 50 N 9 00 W
Tâmega (Támega), *river* .. 41 05 N 8 21 W
Terceira Island ... 38 43 N 27 13 W
Tua, *river* ..... 41 13 N 7 26 W
Vouga, *river* ..... 40 41 N 8 40 W
Zêzere, *river* .... 39 28 N 8 20 W

below 1,200 feet, but inland, where the summer drought is longer, they can be found in some areas with elevations as high as 2,200 feet. In the Douro valley, juniper scrub (*Juniperus oxycedrus*) has been replaced by vineyards.

The thickest forests in Portugal are found in the province of Beira Alta, where cultivation is limited to less than a quarter of the area. Pines prevail in the north, chestnut (*Castanea*) groves on the granites, and ericas on a dense maquis. The altitudinal succession of vegetation is strongly marked on the Serra da Estrela, where a zone of Pyrenean oak (*Q. pyrenaica*) extends above the pedunculated oaks, chestnuts, and pines to an elevation of 5,500 feet, while southern slopes are covered with a maquis of cistus (*Cistus crispus* and *C. ladanifer*) and tree heath. The central ridge marks the southern limit of deciduous oak; these are replaced by *Q. lusitanica*. Heat-loving stone pines (*Pinus pinea*) mix with maritime pines.

In the south, Alentejo still has extensive *matorral* and *charneca*, wastes dominated by cistus or groves of cork oak, often managed in estates, and holm oak (*Q. ilex*). Most of the Algarve landscape is dominated by the trees typical of Mediterranean arboriculture (olive, vine, fig, almond, and carob).

The rich vegetation of both Madeira and the Azores has been Europeanized. About 100 plants are peculiar to Madeira, either as indigenous or as highly individualized varieties. A wealth of ferns, mosses, heaths (especially tree heaths), and junipers reflect the influence of grazing and human activities in both groups of islands. Only Madeira carries much woodland, most of it the result of reforestation (*e.g.*, of poplar, pine, and eucalyptus).

Wildlife. The fauna of Portugal is again a mixture of European and North African types. As in Spain, the wild goat, wild pig, and deer are found in the Portuguese countryside. The wolf survives in the remote parts of the Serra da Estrela, and the lynx in Alentejo. The fox, rabbit, and Iberian hare are ubiquitous. Birdlife is rich because the peninsula lies on the winter migration route of western and central European species.

In the Azores, only the smaller mammals, such as the rabbit, weasel, ferret, rat (brown and black), mouse, and bat, are found. Game birds include woodcock, red partridge, quail, and snipe. Monk seals are native to the Madeira Islands. Forty species of birds breed there, but only the wren is endemic. The variety of beetles (695 species, many indigenous) and of moths (more than 100 species, about one-quarter of which are peculiar to the Madeiras) is remarkable.



Terraced vineyards near the town of Peso da Régua in the upper Douro River valley, northern Portugal.

© Robert Frerck/Odyssey Productions

Fish are plentiful in the Atlantic waters of mainland Portugal, especially the European sardine. Crustaceans are common on the northern rocky coasts. There are oyster beds in the Aveiro Lagoon and in the estuaries of the Tagus and Sado. Larger fish (tunny, bonito) as well as mullet are caught off the Azores and the Madeira Islands.

Settlement patterns. The landscapes of mainland Portugal are the result of human activity since prehistoric times. Inhabited caves and rock shelters, some with rock art (*e.g.*, in Escoural), indicate occupation in the Upper Paleolithic Period. Most of the later Neolithic megalithic monuments and rock-cut tombs are found in western central Portugal or south of the Tagus. Known sites of early metal-using (Copper Age [Chalcolithic] and Bronze Age) people are concentrated in the drier portions of Portugal. Only from the Early Iron Age onward does the whole of Portugal appear to have been equally densely occupied, although the area singularly rich in Roman remains is that lying between Braga and the Gerez valley.

Age of urban centres

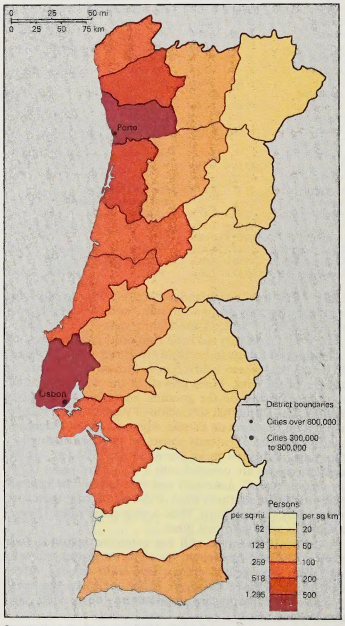
Many of Portugal's urban centres date from Roman times. Settlements developed on lower ground around native fortified hilltop *castros* in northern Portugal. The harbour at Lisbon had been used by the Carthaginians, but it was the Romans who enlarged the site into a strategically located administrative centre for the province of Lusitania. Lisbon continued as a Visigothic stronghold. Indeed, fortification is the keynote for most of Portugal's settlement history. Throughout the country, the Middle Ages and the Reconquest left fortified, usually hilltop, towns, especially toward the Spanish frontier in the south (*e.g.*, Santarém, Tomar, Évora, Portalegre, Estremoz, Beja, Castelo Branco, Abrantes, and Monsanto). Other small towns grew from Cistercian colonization of Estremaduran wastes (*e.g.*, the abbey at Alcobaça, granges at Alvominha, Cos, Turquel, Maiorga, and Salir do Marto). Later new towns were created for a variety of reasons, including proximity to mineral springs (Caldas da Rainha) or important fortresses (Leiria, Viana do Castelo). Around the coasts, the fortunes of port settlements were as unstable as the shifting sands that blocked their harbours, and most modern ports are of relatively recent origin (*e.g.*, Faro, in the south, and Olhão, an 18th-century fishing settlement).

The pattern of rural settlement also reflects both historical and physical factors. The bocage of Minho is associated with a dense distribution of individual holdings on granite that drops to a thin scatter in areas of schist. Most buildings are of two stories with an outside staircase. Nucleated settlement, associated with communal farming systems, is characteristic of Trás-os-Montes and the pastoral districts of Beira Alta. In Estremadura, farmsteads are composed of a number of single-storied buildings. In the formerly feudal south, estate labourers and tenants were housed centrally, in long, barracklike buildings grouped around courtyards (montes), whose origins go back to Arab and even Roman times. Another Arab legacy are the water-wheels and fig-drying floors associated with the dispersed farmsteads of arboriculture districts in the Algarve.

Neither the Madeiras nor the Azores were occupied before the start of colonization by the Portuguese in the 16th century. (I.M.P. do A./C.D.S.)

THE PEOPLE

Although western Iberia has been occupied for a long time, relatively few human remains of the Paleolithic Period have been found. Neolithic and Bronze Age discoveries



Population density of Portugal.

Early settlers

are more common, among them many dolmens. Some of the earliest permanent settlements were the northern *castros*, hill villages first built by Neolithic farmers who began clearing the forests. Incoming peoples—Phoenicians, Greeks, and Celts—intermingled with the settled inhabitants, and Celticized natives occupied the fortified *castros*. For 200 years these were centres of resistance to the Roman legions. Subsequently the Romans, Suebi, Visigoths, Moors, and Jews exerted influence on the territory. Portugal's situation at the western extremity of Europe made it a gathering place for invaders by land, and its long coastline invited settlement by seafarers.

Although formed of such different elements, the population of Portugal is one of the most homogeneous in Europe, having physical characteristics common to circum-Mediterranean peoples. For example, the Portuguese in general have brown eyes, dark wavy hair, and pallid or brunet skin.

The vast majority of Portuguese (some 95 percent) are Roman Catholic. Regular attendance at mass, however, has declined in the cities and larger towns, particularly those in the south. Less than 1 percent of the population is Protestant, with Anglicans, Methodists, Brethren, and Congregationalists comprising the oldest and largest denominations. Fundamentalist and Evangelical churches have grown in recent years. The Jewish population of Portugal has remained small since the late 15th-century Inquisition, which forced Jews to convert or emigrate.

Population distributions within Portugal reveal striking contrasts between the more densely populated north and the more sparsely populated south. A number of rural areas have suffered considerable population losses, resulting in economic and social depression, particularly in parts of the north, the central east, and the southern coastal areas. The coastal zones between Braga and Setúbal, with their low-lying plains and urban development, have attracted a large proportion of the population. Only the industrial areas of Lisbon and Porto are able to absorb their own working populations. Areas such as Minho, the central plains, and the coastal areas of the Algarve are seriously overpopulated.

In the main, rural settlement is dispersed, the inhabitants living in small villages under a system of open-field farming. The Beira Litoral and Estremadura have settlements varying between dispersed and clustered farmsteads. In the region of Aveiro, clusters of farmsteads and other dwellings are strung along roads in strips, often of considerable length and density. Fishing, one of the earliest enterprises of the Portuguese, still plays an important role in coastal settlements. Owing in part to the rigours and hazards of this and certain other traditionally male occupations, as well as emigration, women have substantially outnumbered men in the Portuguese population since the first census, in 1864.

The decolonization process that took place after the revolution of April 25, 1974, inevitably had demographic repercussions on metropolitan Portugal because of the large number of people (mostly Portuguese) who left the former overseas provinces. According to official figures, incoming refugees totaled about one million, most of whom came from Angola because of the civil war between the liberation movements. The majority of the repatriates crowded into Portuguese cities and towns, the effect of which was a high unemployment rate that continued into the late 1980s.

Incoming refugees

Portugal has one of the highest rates of emigration in Europe. Before 1960 most of its émigrés went to Brazil and a few other Latin-American countries. The population underwent its only decline in the modern period during the 1960s, when two external developments coincided: severe labour shortages in industrialized western Europe induced an outflow of Portuguese workers, and Portugal's efforts to put down the liberation movements in its African colonies prompted thousands of young men to emigrate illegally in order to avoid conscription.

THE ECONOMY

At the height of its colonial endeavours in Asia, Africa, and South America, Portugal was the richest nation in the world. Because this wealth was not used to develop domestic industrial infrastructure, however, the country gradually became, over the 19th and 20th centuries, one of the poorest in western Europe. After the 1974 revolution, the Portuguese economy was disconnected from Portugal's remaining overseas possessions in Africa and reoriented toward Europe. Portugal joined the European Community (EC) on Jan. 1, 1986, since which time the economy has grown steadily.

Agriculture, forestry, and fishing. Crop yields and animal productivity are well below the EC average because of low agricultural investment, minimal mechanization, little use of fertilizers, and the fragmented land-tenure system. The main crops grown in Portugal are cereals (wheat, barley, corn [maize], and rice), potatoes, grapes (for wine), olives, and tomatoes. Portugal is the world's largest exporter of tomato paste and a leading exporter of wines. These exports help offset the cost of imported wheat and meat.

Types of farms

Small farms predominate, particularly in the north, where they are too small and made up of too many dispersed holdings to allow integrated farming and rational crop rotation. In the south (except Algarve) the system of *latifúndios*, or large estates, before 1975 prevented intensive cultivation because the *latifúndios* were owned by absentee landlords who had no interest in making capital investment in machinery, fertilizers, and other items that would increase productivity. After the revolution (from March 1975) an agrarian reform got under way south of the Tagus, where about 3.2 million acres (1.3 million hectares) of land in large holdings was expropriated (with compensation) and nationalized. The policy adopted was aimed at, among other things, destroying the *latifúndio* system, changing the tenancy system, handing uncultivated land back to the people, abolishing quitrent, increasing the irrigated area, introducing new crops, intensifying the output of fodder and cereals, and developing livestock. A large part of the nationalized land was turned over to collective and cooperative production units. The hasty transition, however, precipitated political tension and a decline in the agricultural output of the Alentejo. The land redistribution policy was reversed after 1976, as the succeeding socialist governments sought to encourage modernization and higher productivity by a return to private ownership.

Forestry

One-third of Portugal is wooded. Most of the mountainous areas are well suited to forestry, and forest products (cork, resins, and pine and eucalyptus timber), the market value of which has increased, find a growing demand in industry. For this reason, considerable reforestation has been undertaken since the last quarter of the 19th century in areas where crop yields are low and where erosion tends to be severe.

Portugal's long coastline and the abundance of fish in the surrounding waters have favoured the development of the fishing industry. Sardines, anchovies, and tuna caught near the coasts, together with other species such as codfish from the North Atlantic, make a large contribution to food supplies. The fishing industry has prospered, and its products are exported all over the world. Matosinhos in the north, Setúbal, and Portimão and Olhão and other ports in the Algarve are among the main fishery centres. However, even the fishing industry is not able to feed the population; a quarter of the fish consumed in Portugal must be imported, mainly from Iceland (stockfish), Norway (dried cod), and Russia (sardines).

Mining and industry. Portugal depends heavily on the importation of petroleum and petroleum products. Coal, the production of which has increased since the mid-1980s, accounts for about 5 percent of the country's energy use. One-third of Portugal's electricity is provided by hydropower. Tungsten, tin, chromium, and other alloy metals are mined in commercial quantities, and most of the tungsten is exported. Coal mined at Moncorvo supplies the national steelworks. Copper is extracted at Neves Corvo.

Two regions account for three-fourths of Portugal's industrial capacity: Lisbon-Setúbal in the south and Porto-Braga-Aveiro in the north. The Lisbon-Setúbal region supports oil refining, chemical industries, cement processing,

steelmaking, shipbuilding, automobile assembly, electronics manufacture, wood-pulp and cork production, and fish and beverage processing. Light industry prevails in the north: Porto produces textiles, footwear, furniture, wine, and processed food, and Aveiro is a centre for wood pulp and wood products, while Braga specializes in clothing, cutlery, and electronics. A third concentration of industry, particularly fuel and energy production, has developed at Sines, a deepwater port about 90 miles south of Lisbon.

Following the revolution, deep alterations occurred that were not significantly changed until Portugal's entry into the EC in 1986. After the political crisis of March 1975, heavy industry, basic industries (*e.g.,* cement and petrochemical processing, shipbuilding, generation of electricity), and even some light industries were nationalized, a policy that was reversed 15 years later. Under the new system, industrial ownership takes three forms: public enterprises still control mining and the basic and heavy industries; private domestic firms handle the traditional labour-intensive light industries and construction; and subsidiaries of multinational corporations dominate the more technologically advanced industries, such as electronics manufacture, automobile assembly, and pharmaceutical production.

New industrial policies

Banking and finance. Like the manufacturing sector, the banking and insurance industries underwent nationalization in 1975, with gradual reprivatization beginning in the 1980s. By the early 1990s the nationalized banks and insurance firms continued to account for 80 percent and 60 percent of their markets, respectively. State-owned banks suffered from undercapitalization, partly as a result of the legal requirement that they extend low-interest loans to public enterprises and to the government in order to finance large budget deficits. Financial markets have developed since Portugal joined the EC. Deposit and lending rates have been freed, and money-market instruments are available. The Lisbon stock exchange (Bolsa) has been modernized.

Trade. For a country its size, Portugal has a large foreign trade. Total imports (primarily food and beverages, wheat, crude oil, machinery, automobiles, and raw materials) have been greater than total exports (of which the most important are textiles, clothing, footwear, paper pulp, wine, cork, and tomato paste). Britain, Germany, France, and other EC countries are Portugal's principal trading partners. Trade with the country's former possessions in Africa has declined to almost nil.

The trade deficit has traditionally been financed by emigrant worker remittances and tourism income. Since Portugal's accession to the EC, large inflows of private capital, direct investment, and repatriated flight capital have increased the country's foreign reserves, which had been drastically depleted from 1974 to 1984 to pay for a high trade imbalance.

Transportation. Although transport and communications were seriously neglected in the past, there has been a concerted effort to remedy the situation. A four-lane *autoestrada* (superhighway) now connects Lisbon to Porto, the capital of the north, and similar routes are being constructed from Porto and Lisbon to Spain. Highways connect the largest towns and extend to the border and ports. Secondary roads link the towns with almost every part of the interior, and the total road network is being extended. Several of the country's main roads date to ancient times.

Plans under way in the early 1990s to improve the national railroad system included the possible introduction of high-speed train service between Lisbon and Madrid. The Portela Airport at Lisbon, an important junction for intercontinental flights, is the base of the Portuguese airline, Transportes Aéreos Portugueses (TAP), which provides regular services. A large new international airport was scheduled for construction across the Tagus River from Lisbon. Important secondary airports are located at Porto and Faro. (I.M.P. do A./W.C.O.)

GOVERNMENT AND SOCIAL CONDITIONS

Government. Portugal has been a republic since the overthrow of the monarchy in 1910. Its status as a democratic state began with the so-called Revolution of the Carnations on April 25, 1974, which brought down the authoritarian, corporative republic established by António de Oliveira Salazar in 1932. According to the postrevolutionary constitution of 1976, Portugal's chief of state is the president of the republic, who is directly elected by universal suffrage for a five-year term. The president represents the nation; serves as chief commander of the armed forces; appoints the prime minister, with due consideration to the results of parliamentary elections; appoints and dismisses other members of the government at the proposal of the prime minister; sends messages to parliament and convenes or dissolves it as necessary; and sets dates of elections after consultation with the Council of State.

Constitution of 1976

The constitution designates the Council of Ministers (Conselho de Ministros), or the Cabinet, as Portugal's chief policy-making body. The Cabinet consists of the prime minister, who presides over its meetings; the ministers of government departments; and some secretaries of state (ministers without portfolios). The prime minister is simultaneously responsible to the president (regarding the overall functioning of governmental institutions) and to parliament (regarding the content of public policy). The prime minister directs, coordinates, and implements government policy. By tradition, the prime minister is the head of the civil service.

The parliament is the unicameral Assembly of the Republic (Assembleia da República). It has 250 deputies, elected for four-year terms under a system of proportional representation. Among its duties are discussing and voting upon legislation, authorizing the government to raise revenues, and approving the laws passed by the legislatures of the autonomous regions.

The role of the military as the watchdog of the 1974 revolution and the subsequent transition to democracy was enshrined by the 1976 constitution in the Council of the Revolution (Conselho da Revolução). A constitutional committee operated in conjunction with the Council of the Revolution, which pronounced on the constitutionality of laws submitted to it. Revisions made to the constitution in 1982 abolished the Council of the Revolution and the constitutional committee and replaced them with a Council of State (Conselho de Estado) and a Constitutional Tribunal (Tribunal Constitucional). Members of the Council of State are the president of the republic, who presides; the president of the parliament; the prime minister; the president of the Constitutional Tribunal; the attorney general; the presidents of the governments of the autonomous regions; certain former presidents of the republic; five persons appointed by the president; and five persons selected by the parliament. The Constitutional Tribunal has 13 justices appointed by parliament.

Portugal has three tiers of government below the national level. The lowest tier comprises the parishes (*freguesias*), of which there are about 4,000. Each has a directly elected parish assembly (*assembleia de freguesia*), which appoints its own executive body, the parish board (*junta de freguesia*). The second tier consists of the municipalities (*concelhos*), which number 305. Municipalities include rural and urban areas within their territorial limits. Each municipality has a municipal assembly (*assembleia municipal*), made up of the presidents of the boards of the constituent parishes and an equal number plus one of directly elected members; a municipal chamber (*câmara municipal*), which is the executive of the municipality; and a municipal council (*conselho municipal*), a consultative organ through which the views of social, cultural, professional, and economic organizations within the municipality are transmitted to the municipal chamber. The tier above the municipalities—that of the administrative regions—had not yet been fully established by the early 1990s but was to have regional assemblies and executives and was intended to decentralize the national government. The archipelagoes of Madeira and the Azores have been granted the special status of autonomous regions in recognition of their geographic, economic, social, and cultural uniqueness, as well as their historical aspirations for greater independence. Each autonomous region has its own government (cabinet and president), legislature (regional assembly), and administration.

Local government

The Portuguese police are divided into three categories. Under the control of the Ministry of Internal Administration are the Public Security Police (Polícia de Segurança Pública, or PSP) and the Republican National Guard (Guarda Nacional Republicana, or GNR). The GNR includes the road police and has jurisdiction over rural areas. The PSP patrols urban areas and directs city traffic. Under the supervision of the Ministry of Finance is the Fiscal Guard (Guarda Fiscal), which is stationed at frontier crossings and points of entry and is responsible for collecting import duties and investigating smuggling and other violations of border regulations.

Education. Private schools supplement the state schools that provide free education for the majority of people. There are five long-established universities: the University of Coimbra (founded under another name in Lisbon in 1290, relocated to Coimbra and renamed in 1537), the University of Lisbon and the Technical University of Lisbon, the University of Porto, and the Portuguese Catholic University (founded in 1968 in Lisbon).

Welfare services. The Portuguese welfare system is composed of several types of institutions that insure workers against sickness, disability, and old age and provide for the payment of pensions and family allowances. Compulsory insurance by employers is in force in most sectors of business and industry, the employee paying into the fund as well. Trade-union provident funds and welfare funds for other employees take care of most categories of workers, and there are voluntary mutual assistance associations and provident institutions for the military forces and civil servants. Many large companies conduct their own welfare and sick benefit schemes and pensions for their employees.

Workers have the choice of belonging to one of two trade-union federations. One is the Intersindical, which is communist-controlled. The other is the União Geral dos Trabalhadores (UGT; General Union of Workers), which is democratic.

Hospitals

Portugal has both public and private hospitals. Regional hospitals are located in the main provincial centres and subregional hospitals in smaller centres. Charity hospitals (the *santas casas da misericórdia*) give free treatment to the underprivileged. The civil hospitals, mainly run by the municipalities with state support and funds from welfare and social organizations, charge according to the means of the patient. All over the country there are dispensaries and clinics serving mothers and infants. Special institutions include a cancer hospital and research unit in Lisbon, a school of tropical medicine, a sanatorium in the northern mountains, and a modern rehabilitation centre for the disabled near Lisbon.

CULTURAL LIFE

Portuguese culture is based on a past that dates from prehistoric times into the eras of Roman and Moorish invasion. All have left their traces in a rich legacy of archaeological remains, including prehistoric cave paintings at Escoural, the Roman township of Conimbriga, the Temple of Diana in Évora, and the typical Moorish architecture of such southern towns as Olhão and Tavira. Throughout the centuries, Portugal's arts have been enriched by foreign influences, including Flemish, French, and Italian. The voyages of the Portuguese discoverers opened the country to Oriental influences, and the revelation of Brazil's wealth of gold and jewels fed the Baroque flame in decoration.

Daily life. In spite of certain affinities with the neighbouring Spaniards, the Portuguese have their own distinctive way of life. The geographic variety of the country has evoked different responses, but there is less regionalism than in Spain. Dancing and singing play a prominent part in the life of the people. Almost every village has its own *terreiro,* or dance floor of beaten earth. Each region has its own style of dances and songs; most traditional songs are of a slower rhythm than those in Spain. Small accordions are often used to accompany dances, and Portuguese guitars accompany the fado, a song form that epitomizes *saudade,* the yearning, romantic aspect of the Portuguese character. Some of the best examples of the regional dances are the *vira, chula, corridinho, tirana,* and fandango, many of which reflect the courting and matrimonial traditions of the area. Much has been done to preserve these and other folk expressions as tourist attractions.

Traditional dances

National dress is still seen in the northern Minho province at weddings and other festivals. Traditional garments such as the red and green stocking cap of the Alentejo cattleman still exist, and the *samarra* (a short jacket with a collar of fox fur) and *cifões* ("chaps") survive. In Trás-os-Montes e Alto Douro shepherds wear straw cloaks. The wearing of black for protracted periods of mourning is common especially in the villages.

The staple diet is one of fish, vegetables, and fruit. Although Portugal's waters abound with fresh fish, the dried, salted codfish known as *bacalhau,* now often imported, is considered the national dish. In many areas, meat is seldom eaten, although the Alentejo region is known for its pork and Trás-os-Montes for cured meats. Breads and sweets—the latter a legacy of Moorish occupation—take a variety of forms, with many regional specialties. Wine is the ubiquitous table beverage. Perhaps the most famous Portuguese export is the fortified red wine called port after the town of Porto, where it has been bottled for centuries; a similar wine is produced on and named after the island of Madeira.

Architecture. Romanesque and Gothic influences have given the country some of its greatest cathedrals, and in the late 16th century a national style (Arte Manuelina) was synthesized by adapting several forms into a luxuriantly ornamented whole. Outstanding examples of Portuguese architecture include the Jerónimos monastery in Lisbon, in ornate Manueline style; the Sé (cathedral) of Lisbon, in part of the facade of which the remains of Roman construction may still be seen; the Palace of Justice in Lisbon, a fine, soaring example of austere modern architecture; the castle and church of the Convent of Christ (Convento de Cristo) in Tomar; the late Portuguese Gothic abbey of Santa Maria da Vitória in Batalha; the granite Torre dos Clérigos (Tower of the Clerics) in Porto; and Braga's Romanesque cathedral.

Visual and decorative arts. Sculpture found rich expression in the magnificent tombs of the 12th and 13th centuries, and late 18th-century Baroque wood sculptures, of which the crèches of Joachim Machado de Castro are the finest, also are outstanding. The Classical and Romantic traditions of Italy and France influenced Machado de Castro in the late 18th century and António Soares dos Reis a century later. A school of primitive painters headed by Nuno Gonçalves was prominent in the 15th century, and subsequently Flemish artists interpreted the native style, decorating palaces and convents and leaving a rich heritage of religious art. The 19th century saw another rebirth of national art with a late Romantic period. An era of naturalist realism that followed gave way to rather diffuse experimentation in the 20th century. Maria Helena Vieria da Silva was the country's finest abstract painter, and Carlos Botelho was noted for his street scenes of Lisbon.

Among the decorative arts, the Portuguese glazed tiles (*azulejos*) are outstanding. Many 16th- and 17th-century buildings are faced with tiles, and the rooms and halls of palaces and mansions exhibit blue and white tiled panels or motifs in other soft colours. Exceptionally fine examples are found in the Pátio da Carranca (*pátio,* "courtyard") of the Paço do Sintra (*paço,* "palace") at Sintra, São Roque church in Lisbon, and the Quinta da Bacalhoa at Vila Fresca de Azeitão near Setúbal.

Music. Liturgical forms such as plainsong dominated early Portuguese music, but the secular tradition of troubador singing became popular in the Middle Ages. Polyphonic music, employing multiple vocal parts in harmony, was developed in the 15th century. The Renaissance fostered a rich output of composition for solo instruments and ensembles as well as the voice. The modern revival of so-called academic music in Portugal was primarily the work of Luís de Freitas Branco, whose Neoclassic tradition has been perpetuated by Joly Braga Santos. The Calouste Gulbenkian Foundation (founded by the oil magnate) continues to inspire the country's musical life. A group of young composers is acquiring prestige both at home and

The Gulbenkian Foundation

in international competitions, among them Filipe Pires, A. Vitorino de Almeida, Alvaro Cassuto, and Jorge Peixinho. There is an active National Broadcasting Symphony Orchestra, and Porto has had its own symphony orchestra since 1962, the year in which the Chamber Orchestra was set up by the Gulbenkian Foundation. In 1971 the Lisbon Municipal Orchestra was created, and the Teatro São Carlos also has its own orchestra and ballet company. Folk music and dancing and the traditional fado remain the country's fundamental forms of musical expression.

Literature. The Portuguese language became synthesized in the 12th century, when a lyrical quality was outstanding in both poetry and prose. A soldier-poet, Luis de Camões, gave expression to the nation's epic genius in the 15th century, as the poet Fernando Pessoa expressed its decadence in the 19th. Lyric poetry still flourishes. The tendency of fiction has been away from the romanticism of the 19th and early 20th centuries and toward realism. José Maria Eça de Queirós was the outstanding realist novelist, his works including *Os Maias* (1888; *The Maias*) and *A Cidade e as Serras* (1901; *The City and the Mountains*). In the 20th century the outstanding regional novelist was Aquilino Ribeiro, whose works include *Jardim dos Tormentos* (1913; "Garden of Torments") and *O Homem que Matou o Diabo* (1930; "The Man Who Killed the Devil"), while José Maria Ferreira de Castro was the outstanding realist, with *A Selva* (1930; *The Jungle*) and *Os Emigrantes* (1928; "The Emigrants"). The novelist, essayist, and poet Vitorino Nemésio received acclaim for his novel *Mau Tempo no Canal* (1944; "Bad Weather in the Channel"). Censorship under the Salazar regime considerably stifled meaningful literary production. Among those whose work flourished under the new democracy were José Saramago, Miguel Torga, and Almeida Faria.

Theatre and motion pictures. While the theatre is rapidly picking up after a period of stagnation, the Portuguese film industry is still inactive. With the rebuilding of the classic drama theatre, the National Theatre of Dona Maria II in Lisbon (gutted by fire), there is a revival of the serious play. State and Gulbenkian Foundation aid has given new life to the "little theatre" not only in Lisbon but also in the provinces. International opera, theatre, and ballet are prominent in regular seasons in the National Theatre of São Carlos, in Lisbon, and in other theatres.

Museums, libraries, and art galleries. In Lisbon, in addition to the Gulbenkian Foundation, which houses Gulbenkian's collections, the most outstanding are the National Art Museum, the National Museum of Coaches, with Europe's finest collection of ancient vehicles, the National Museum of Contemporary Art, and the Museum School of Decorative Arts, which trains craftsmen in furniture restoration, bookbinding, repair of ancient tapestries, and other fine handicrafts. Provincial museums include the Soares dos Reis National Museum in Porto, Machado de Castro National Museum in Coimbra, and the Regional Museum of Aveiro. The National Library and Ajuda Library in Lisbon have fine collections, while the National Archives of Torre do Tombo contain valuable national documents. Outside Lisbon, the library of the Convent of Mafra in Mafra and that of the University of Coimbra have historical importance.

Libraries

Recreation. Regional fairs occur throughout the year, many of them combined with religious festivals. Religious customs in this Roman Catholic country include the burning of the yule log in the atrium of the village church at Christmas, so that the poor may warm themselves, and reenactments of the fight between St. George and the Dragon in the streets of Monção at Corpus Christi.

Bullfighting is a popular sport and varies markedly from its Spanish counterpart. The Portuguese bullfighter rides a horse and does not seek to kill the bull, the horns of which may be sheathed to protect the horse.

Soccer attracts both spectators and participants in great numbers. Basketball and roller hockey are other popular team sports. Portugal's long seacoast and mild climate make beach-going a favourite pastime.

Press and broadcasting. Before the revolution of 1974, all media in Portugal were censored. The 1976 constitution guarantees freedom of the press and broadcasting. Portuguese newspapers cater to a comparatively small, largely urban reading public. The nationalization of industry that began in 1974 encompassed the leading Lisbon newspapers, which had been owned by banks. Gradual reprivatization began in 1979. The daily *Diário de Notícias* (founded 1864) was long Portugal's most prestigious newspaper. Since the onset of privatization, however, the weekly *Expresso* has assumed this position. Other newspapers founded in the capital since the revolution are the dailies *Correio da Manhã, Jornal de O Dia,* and *Público* (founded 1990) and the weekly *O Jornal. Anglo-Portuguese News,* an English-language weekly, has been published in Lisbon since 1937. The northern city of Porto has three principal dailies: *O Comércio do Porto, O Primeiro de Janeiro,* and *Jornal de Notícias.* Magazines of national and international news and review include the weekly *Sábado* and the monthly *A Revista.*

The broadcast media reach a much larger portion of the Portuguese population than do the print media. In 1975 all private radio broadcasting, except the church-owned Rádio Renascença, was nationalized. The reprivatization process has paralleled that of other industries. Ownership reform came much more slowly to television broadcasting, which since its inception had been limited to the state-owned Radiotelevisão Portuguesa (RTP). In 1991 two private concerns, one financed by a publishing group and the other by the Roman Catholic church, received television broadcasting licenses. (J.Sh./W.C.O.)

For statistical data on the land and people of Portugal, see the *Britannica World Data* section in the BRITANNICA BOOK OF THE YEAR.

History

Portugal became an independent monarchy in 1139. Its name derives from Portus Cale, a pre-Roman or Roman settlement near the mouth of the Douro River. The southern part of the Roman province of Gallaecia (now Galicia) was occupied by the Germanic Suebi in AD 411, and Portucale (Porto) was held by them. They were subdued by the Visigoths, whose state was overthrown by Muslim invasions in the 8th century. A Christian territory of Portugal was constituted in 868; it later became a county and was extended to Coimbra. Afonso Henriques assumed the title of king (1139) and annexed Lisbon. His successors took Alentejo and Algarve from the Muslims (by 1252). Portugal now includes the Madeira Islands and the Azores, first settled in the 15th century.

PRE-ROMAN, ROMAN, GERMANIC, AND MUSLIM PERIODS

The earliest human remains found in Portugal are Neanderthal-type bones from Furninhas. A distinct culture first emerges in the Mesolithic middens of the lower Tagus valley, dated about 5500 BC. Neolithic cultures entered from Andalusia. There are varied types of beehive huts and passage graves: agriculture, pottery, and the working of soft metals followed by the same route. In the 1st millennium BC, Celtic peoples entered the peninsula by the Pyrenees, and many groups were projected westward by natural pressure. Phoenician and later Carthaginian influence reached southern Portugal in the same period. By 500 BC, Iron Age cultures predominated in the north. Celtic hilltop settlements (*castros*) retained their vitality after the Roman conquest.

Celtic tribes

After the Second Punic War (218–201 BC), Rome dominated the eastern and southern seaboards of the peninsula, and Celtic peoples who had partially absorbed the indigenous population occupied the west. A Celtic federation, the Lusitani, resisted Roman penetration under the brilliant leadership of Viriathus; but after his assassination (*c.* 140 BC) Decius Junius Brutus was able to march northward through central Portugal, cross the Douro River, and subdue the Gallaeci. Julius Caesar governed the territory for a time. In 25 BC Augustus founded Augusta Emerita (Mérida) as the capital of Lusitania, which incorporated the present central Portugal. Gallaecia, to the north of the Douro, became a separate province under the Antonines. In Roman times the region of Beja and Évora was a wheat belt. The valley of the Tagus was famous for its horses

and farms, and there were important mines in Alentejo. Notable Roman remains include the Temple of Diana at Évora and the site of Conimbriga (Condeixa). Christianity reached Lusitania in the 3rd century and Galicia in the 4th.

Suebi, Visigoths, and Muslims

After AD 406, barbarian peoples forced their way into Gaul and crossed the Pyrenees. A Germanic tribe, the Suebi, was settled in southern Galicia, their rulers residing at or near Bracara Augusta (Braga) and Portucale. They annexed Lusitania and for a time overran the rest of the peninsula, but the Visigoths were sent to subdue them and extinguished their monarchy (AD 469). Records are silent until about 550, when the Suebic monarchy had been restored and was reconverted to Catholicism by St. Martin of Braga. With the Muslim invasion of 711, the only serious Gothic resistance was made at Mérida; on its fall the northwest submitted. Berber troops were placed in central Portugal and Galicia. When 'Abd ar-Raḥmān I set up the Umayyad monarchy at Córdoba (756), there was some resistance in the west. Lisbon was independent for a few years (*c.* 805). The restoration of the Christian sees of Galicia, the discovery of the supposed tomb of St. James, and the erection of his shrine at Santiago de Compostela were followed by the organization of the frontier territory of Portucale (868) by Vimara Peres; Coimbra was annexed by the Christians but later lost again.

THE COUNTY AND KINGDOM OF PORTUGAL TO 1383

By the 10th century the county of Portugal (north of the Douro) was held by Mumadona Dias and her husband Hermenegildo Gonçalves and their descendants, one of whom was tutor and father-in-law to the Leonese ruler, Alfonso V. But, when this dynasty was overthrown by the Navarrese-Castilian house of Sancho III the Great, the western county lost its autonomy. Sancho's son, Ferdinand I of Castile, reconquered Coimbra in 1064 but entrusted it to a Mozarabic governor. When the African Almoravids annexed Muslim Spain, Alfonso VI of León (1065–1109) and Castile (1072–1109) provided for the defense of the west by calling on Henry, brother of Duke Eudes (Odo) of Burgundy, whom he married to his illegitimate daughter Teresa and made count of Portugal. From 1095, therefore, Henry and Teresa (who used the title of queen) ruled Portugal and Coimbra. On the death of Alfonso VI, his realms passed to his legitimate daughter Urraca and her little son Alfonso (VII). Henry of Portugal sought power but had achieved little when he died in 1112, leaving Teresa with an infant son, Afonso Henriques. Teresa's intrigues with her Galician favourite, Fernando Peres of Trava, lost her the support of the Portuguese barons, and in 1128 followers of her son defeated and drove her into exile. Thus

Afonso Henriques

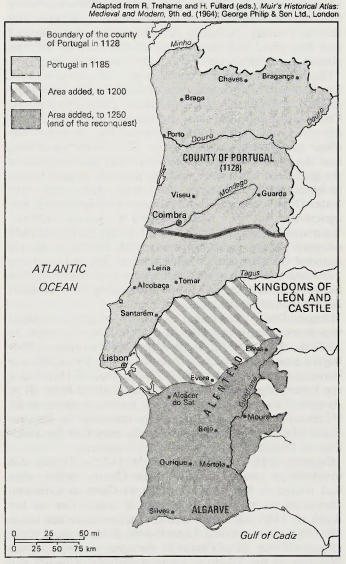
Afonso Henriques became count of Portugal. Though at first obliged to submit to his cousin Alfonso VII, he began to use the title of king, traditionally because of a victory over the Muslims at Ourique (July 25, 1139; but perhaps a symbolic legend). In 1143 his cousin accepted his autonomy, but the title of king was formally conceded only in 1179, when Afonso Henriques placed Portugal under the direct protection of the Holy See, promising an annual tribute. He had captured Santarém (March 1147) and Lisbon (October 1147), the latter with the aid of English, French, German, and Flemish crusaders bound for Palestine. An English priest, Gilbert of Hastings, became first bishop of the restored see of Lisbon. Although the new Moroccan dynasty of the Almohads struck back (1179–84), when Afonso I died (Dec. 6, 1185) the Portuguese frontier was firmly established on the Tagus. The new military orders, the Templars—including those of Calatrava (from *c.* 1156) and of Santiago (from *c.* 1170)—governed castles and territory on the frontier, and the Cistercians were responsible for the introduction of agriculture and architecture in central Portugal (Alcobaça).

The kingdom and the Reconquest. Although Afonso Henriques granted charters to new settlements, it was his son Sancho I (1185–1211) who enfranchised many municipalities (*concelhos*), especially in eastern and central Portugal. The privileges of these communities were embodied in charters (*forais*), which attracted settlers from the more feudal north. Even Muslims were enfranchised, though on the other hand many of them were enslaved. Assisted by passing crusaders, Sancho captured Silves in Algarve (1189); but in 1190 an Almohad army from Africa advanced to the Tagus, and, although Lisbon, Santarém, and Tomar stood firm, the Muslims recovered Silves in 1191, together with most of the land south of the Tagus. But peace was made before Sancho's death, and it was left to his son Afonso II the Fat (1211–23) to endeavour to strengthen the power of the throne at the expense of the church.

Though Afonso II was an unwarlike king, his followers were beside the Castilians at the great Christian victory of Las Navas de Tolosa in 1212 and, again assisted by crusaders, recovered Alcácer do Sal in 1217. Meanwhile Afonso repudiated the bequests of large estates made by his father to his brothers and accepted those to his sisters only after a war with León and in a form, settled by the pope, that recognized Afonso's sovereignty. In the first year of his reign, Afonso called a meeting of the Cortes at Coimbra, to which the nobility and prelates were summoned (representatives of the commoners were not to appear until 1254). Both estates obtained important concessions; in fact, the position of the church and the orders was now so strong that Afonso II and his successors were involved in recurrent conflicts with Rome. Afonso himself instituted (from 1220) *inquirições*, or royal commissions, to investigate the nature of holdings and recover whatever had been illegally taken from the crown.

The Cortes

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Medieval and Modern*, 9th ed. (1964); George Philip & Son Ltd., London



Portugal in the 12th and 13th centuries.

Little is known of the reign of his son Sancho II (1223–*c.* 1246), but the reconquest of Alentejo was now completed, and much of Algarve was reduced. On his accession, Sancho found the church in full ascendancy as a result of the agreement made before his father's death. At all events, his younger brother Afonso, who had become count of Boulogne by his marriage (1238) with Matilde, daughter of Count Raynald I of Dammartin, was granted a papal commission (1245) to take over the government, and Sancho was ordered to be deposed by papal bull.

When Afonso reached Lisbon (late 1245 or early 1246), he received the support of the church and of the inhabitants of Lisbon and other towns. After a civil war lasting two years, Sancho II retired to Toledo, dying there in January 1248.

On his arrival, the count of Boulogne had already declared himself king as Afonso III, and the death of Sancho without issue gave his usurpation the mantle of legality. He brought together the divided kingdom, completed the reconquest of Algarve, transferred the capital from Coimbra to Lisbon, and, fortified by the support of the towns, summoned the Cortes at Leiria at which for the first time commoners representing the municipalities made their appearance (1254). His conquest of Algarve aroused the jealousy of Castile. Two campaigns were fought (1250 and 1252), Afonso probably being worsted, and peace was made by means of a marriage pact. Although still the husband of Matilde of Boulogne, Afonso married Beatriz, illegitimate daughter of Alfonso X of Castile, holding the disputed territory of Algarve as a fief of Castile until such time as the eldest son of the marriage should reach the age of seven, when Algarve was to return to Portugal. This marriage led to a dispute with the Holy See, in which Afonso was placed under an interdict. Despite his early connection with Rome, Afonso refused to give way; in 1263 the bigamous marriage was legalized and his eldest son, Dinis, legitimized. Shortly afterward, Afonso launched *inquirições*, as a result of which the church was deprived of much property. The prelates protested these actions of the royal commissions, and most of them subsequently left the country. Although Afonso was excommunicated and threatened with deposition, he continued to defy the church until shortly before his death early in 1279.

First representation of commons

The achievements of Afonso's reign—the completion of the Reconquest, the assertion of royal power before the church, and the incorporation of the commoners in the Cortes—indicate important institutional advances. Under his son Dinis (1279–1325), Portugal was to come into closer touch with western Europe. The chartering of fairs and the increased use of minted money bear witness to the growth of commerce, and the planting of pine forests to hold back the advancing sand dunes near Leiria illustrates Dinis' concern for furtherance of shipbuilding and agriculture. In 1317 Dinis engaged a Genoese admiral, Emmanuele Pessagno, to build up his navy, having already adopted various measures to stimulate foreign trade. He founded the University of Coimbra (at first in Lisbon) in 1290 and was both a poet and a patron of literature. Yet he was especially famed as the "farmer king" (*rei lavrador*) for his interest in the land.

Advances under Dinis

Despite Dinis' attachment to the arts of peace, Portugal was several times involved in strife during his reign. In 1297 the Treaty of Alcañices with Castile confirmed Portugal's possession of Algarve and provided for an alliance between Portugal and Castile. The mother of Dinis' son, the future Afonso IV, was Isabel, daughter of Peter III of Aragon. This remarkable woman, later canonized as St. Elizabeth of Portugal and popularly known as the Rainha-Santa (the Holy Queen), successfully exercised her influence in favour of peace on several occasions.

Disputes with Castile. Afonso IV (1325–57) was also involved in various disputes with Castile. Isabel, who had retired to the Convent of Santa Clara at Coimbra, continued to intervene in favour of peace; but on her death in 1336 war broke out, and terms were not made till 1340, when Afonso himself with a Portuguese army joined Alfonso XI of Castile in the great victory over the Muslims on the Salado River in Andalusia. Afonso's son Peter (Pedro) was married (1336) to Constança (d. 1345), daughter of the Castilian infante Juan Manuel; but soon after the marriage he fell in love with one of her ladies, Inês de Castro, by whom he had several children. Afonso was persuaded to countenance the assassination of Inês in 1355, and one of Peter I's earliest acts on his accession was to take vengeance on her murderers. During his short reign (1357–67), Peter devoted himself to the dispensation of justice; his judgments, which he executed himself, were severe and often violent, and his iron rule was tempered only by fits of reveling.

Peter's son by Constança, Ferdinand (1367–83), inherited a wealthy throne almost free of external entanglements; but the dispute between Peter the Cruel and Henry of Trastamara (later Henry II) for the Castilian throne was raging; and, on the murder (1369) of the former, several Castilian towns offered Ferdinand their allegiance, which he was unwise enough to accept. Henry II duly invaded Portugal in 1369, and by the Peace of Alcoutim (1371) Ferdinand was constrained to renounce his claim and to promise to marry Henry's daughter; however, he instead took a Portuguese, Leonor Teles, although she was already married and in spite of the protests of the commoners of Lisbon. He also made alliance (1372) with the English in the person of John of Gaunt, duke of Lancaster, who had married the elder daughter of Peter the Cruel and claimed the Castilian throne. In 1372 Ferdinand provoked Henry II, who invaded Portugal and besieged Lisbon. Unable to resist, Ferdinand was forced to repudiate his alliance with John of Gaunt and to act as an ally of Castile, surrendering various castles and persons as hostages. It was only on the death (1379) of Henry that Ferdinand dared openly to challenge Castile again. In 1380 the English connection was resumed, and in the following year John of Gaunt's brother, Edmund of Langley, earl of Cambridge (afterward duke of York), took a force to Portugal for the invasion of Castile and betrothed his son Edward to Ferdinand's only legitimate child, Beatriz. In mid-campaign, Ferdinand came to terms with the enemy (August 1382), agreeing to marry Beatriz to a Castilian prince. She did, in effect, become the wife of John I of Castile; and when Ferdinand died, prematurely decrepit, Leonor Teles became regent and Castile claimed the Portuguese crown.

Alliance with the English

Leonor had long been the paramour of the Galician João Fernandes Andeiro, count of Ourém, who had intrigued with both England and Castile and whose influence was much resented by Portuguese patriots. Opponents of Castile chose as their leader an illegitimate son of Peter I, John, master of Aviz, who killed Ourém (December 1383) and, being assured of the support of the populace of Lisbon, assumed the title of defender of the realm. The regent fled to Alenquer and thence to Santarém, and the king of Castile came to her aid, soon however, relegating her to a Spanish convent. Lisbon was besieged for five months (1384), but an outbreak of plague obliged the Castilians to retire.

THE HOUSE OF AVIZ, 1383–1580

The legitimate male line of Henry of Burgundy ended at Ferdinand's death, and, when the Cortes met at Coimbra (March–April 1385), John was declared king and became the founder of a new dynasty. This result was not unopposed, for many of the nobility and clergy still considered the queen of Castile the rightful heiress; but popular feeling was strong, and John I had valuable and trusted allies in Nuno Álvares Pereira, "the Holy Constable," his military champion, and in João das Regras, his chancellor and jurist.

Independence assured. A number of towns and castles still held out for Castile when in August 1385 John I of Castile and a considerable army made their appearance in central Portugal. Although much outnumbered, the Portuguese won the great Battle of Aljubarrota (Aug. 14, 1385), in which the Castilian chivalry was dispersed and John of Castile himself barely escaped. The victory assured John I of his kingdom and made him a desirable ally. A small force of English archers had been present at Aljubarrota. Now the Treaty of Windsor (May 9, 1386) raised the Anglo-Portuguese connection to the status of a firm, binding, and permanent alliance between the two crowns. John of Gaunt duly went to the Peninsula (July 1386) and attempted an invasion of Castile in conjunction with John I. This was not successful, but the Portuguese king married Gaunt's daughter Philippa of Lancaster (1387), who introduced various English usages into Portugal. The truce arranged with Castile in 1387 was prolonged at intervals until peace was finally concluded in 1411.

The victory of John of Aviz may be regarded as a victory of the national spirit against the feudal attachment to established order. As much of the older nobility had

adhered to Castile, John rewarded his followers at their and at the crown's expense. Meanwhile, commerce prospered, and the marriage of John's daughter Isabella to Philip the Good of Burgundy was to be followed by the growth of close trading relations between Portugal and Philip's county of Flanders. With the conclusion of peace with Castile, John found an outlet for the activities of his frontiersmen and of his own sons in the conquest of Ceuta (1415), from which may be dated the great age of Portuguese expansion.

Trade with Flanders

During the short reign of John's eldest son, Edward (Duarte; 1433–38), an unsuccessful attempt was made to conquer Tangier in 1437 by John's third son, Prince Henry the Navigator, and his younger brother Ferdinand, who was captured by the Moors and died, still unransomed, in 1443. On Edward's death, his son Afonso V was still a child, and his brother Pedro, duke of Coimbra, had himself made regent (1440) instead of the widow, Leonor of Aragon. But Pedro's own regency was later challenged by the powerful Bragança family, descended from Afonso, illegitimate son of John of Aviz, and Beatriz, daughter of Nuno Álvares Pereira. This family continued to set the young king against his uncle, who was forced to resign the regency, driven to take up arms, and killed at Alfarrobeira (May 1449). Afonso V (1438–81) proved unable to resist the demands of the Braganças, who now became the wealthiest family in Portugal. Having married Joan, daughter of Henry IV of Castile, Afonso laid claim to the Castilian throne and became involved in a lengthy struggle with Ferdinand and Isabella in the region of Zamora and Toro, where he was defeated in 1476. He then sailed to France to entreat the aid of Louis XI, in which he failed, and on his return concluded with Castile the Treaty of Alcáçovas (1479), abandoning the claims of his wife Joan. Afonso never recovered from his reverse, and during his last years his son John administered the kingdom.

Consolidation of the monarchy. John II (1481–95) was as cautious, firm, and jealous of the royal power as his father had been openhanded and negligent. At the first Cortes of his reign, he exacted a detailed oath of homage that displeased his greatest vassals. A suspicion of conspiracy enabled him to arrest Fernando II, duke of Bragança, and many of his followers; the duke was sentenced to death and executed (1484) at Évora. As well as attacking the power of the nobility, John lessened the effects of the unfavourable treaty with Castile. Calculating and resolute, he later received the epithet "the Perfect Prince."

John II was predeceased by his legitimate son and therefore succeeded by his cousin, the duke of Beja, as Manuel I (1495–1521), known as "the Fortunate." Manuel, who assumed the title of "lord of the conquest, navigation, and commerce of India, Ethiopia, Arabia, and Persia," inherited, because of the work of John II, a firmly established autocratic monarchy and a rapidly expanding overseas empire. Drawn toward Spain by the common need to defend their overseas interests as defined by the Treaty of Tordesillas (1494), Manuel nourished the hope of joining the whole peninsula under the house of Aviz; he married Isabella, eldest daughter of Ferdinand and Isabella, who, however, died (1498) in giving birth to a son, Miguel da Paz. This child was recognized as heir to Portugal, Castile, and Aragon but died in infancy. Manuel then married Isabella's sister Maria (d. 1517) and thirdly Eleanor, sister of the emperor Charles V.

As a condition of the marriage with Isabella, Manuel was required to "purify" Portugal of Jews. John II had admitted many Jewish refugees expelled from Spain (1492) and had taxed them heavily but was also to supply ships for them to leave Portugal. This was not done, and Manuel now ordered all Jews to leave within 10 months, by October 1497. On their assembly in Lisbon, every effort was made to secure their conversion by promises or by force. Some who resisted were allowed to go, but the rest were "converted" under promise that no inquiry should be made into their beliefs for 20 years. As "Christians" they could not be forced to emigrate, and they were indeed prohibited from leaving Portugal. In April 1506 a large number of these "new Christians," or *marranos,* were massacred in Lisbon during a riot, but Manuel afterward protected them and allowed many to emigrate to Holland, where their experience of Portuguese trade was put at the service of the Dutch.

Forced conversion of Jews

If Manuel failed to realize his dream of ruling Spain, his son John III (1521–57) lacked the power to resist Castilian influence. A pious, retiring man, he was ruled by his wife, Catherine, sister of the emperor Charles V, and encouraged the installation of the Inquisition (1536); the first auto-da-fé (public sentencing ceremony) was held in 1540. The Society of Jesus, established in 1540, soon controlled education in Portugal. In 1529 the settlement by the Treaty of Saragossa (Zaragoza) of a dispute over the possession of the Moluccas removed an obstacle to Portuguese-Spanish understanding, and the line of Tordesillas was matched by a similar line in the Pacific; all the territories of the New World were in theory divided between Spain and Portugal, while the Reformation (as well as the discoveries) had come between the latter and its English ally.

John III was succeeded by his grandson Sebastian (1557–78), then a child of three. As a boy, Sebastian became obsessed with the idea of a crusade against Morocco. Fanatically religious, he had no doubts of his own powers and listened only to flatterers. He visited Ceuta and Tangier in 1574 and began in 1576 to prepare a large expedition against Larache: this departed in June 1578 and on August 4 was utterly destroyed by the Moors in the Battle of the Three Kings near Alcazarquivir (Ksar el-Kebir). Sebastian himself was killed, 8,000 of his men are said to have died, and 15,000 were captured. Only a handful escaped.

The disaster of Larache

Sebastian was succeeded by his great-uncle, Cardinal Henry (1578–80), a brother of John III. His age and celibacy made it certain that the Portuguese throne would soon pass from the direct line of Aviz. Philip II of Spain, nephew of John III and husband (by his first marriage) of John's daughter Maria, had already made his preparations and, on the death of the cardinal-king (Jan. 31, 1580), summoned the authorities to obey him. An army under the great duke of Alba entered Portugal in 1580; the resistance of António the prior of Crato (illegitimate son of John III's brother Luís), acclaimed António I at Santarém, collapsed; and Philip II of Spain became Philip I of Portugal (1580–98).

MEDIEVAL SOCIAL AND ECONOMIC DEVELOPMENT

Medieval Portugal comprised regions of considerable diversity. In the north, the old aristocracy of Leonese descent owned large estates worked mainly by serfs. In the southern territory that had been reconquered from Muslim rule, there were many towns, often separated by districts almost barren and depopulated. The initiative in settling these areas was taken by Cistercian monks, who had reached Portugal by 1143; later kings such as Sancho I and Afonso III established *concelhos* (municipalities), granting to them by the issue of *forais* (charters) many privileges designed to attract population. Tax concessions were often given, and freedom was promised to serfs or to Christian captives after a year's residence. But in the south the *concelhos* were burdened with defense duties, the *cavaleiros-vilãos* (villein knights) being obliged to horse and arm themselves, and the *peões,* or less substantial men, being obliged to serve as foot soldiers in defense of the country and perhaps also on a *fossado* (raid) into Muslim territory.

At court the king was advised by his *curia regis* (court or council), comprising the *majordomus curiae,* the head of the administration, the military chief or *signifer,* the *dapifer curiae* (steward of the household), the chancellor (an official whose origins in Portugal were Burgundian rather than Visigothic), and any members of the greater aristocracy, the *ricos-homens,* who might be at court. This class also comprised the bishops and abbots and masters of the orders of knighthood; many held private civil or military authority. The lesser nobility were without such rights. Below them came various classes of free commoners, such as the *cavaleiros-vilãos* and the *malados,* men who had commended themselves to protectors. There were numerous serfs and slaves.

Court officials

By the end of the medieval period several shifts in the social structure had occurred. Many of the old aristocracy

lost their position at the advent of the house of Aviz, and the new nobility, exemplified in the house of Bragança, was often of bureaucratic or ministerial origin. Representatives of the commoners, first attending the Cortes in 1254 on behalf of the *concelhos*, took an increasing part in politics. The Cortes were very frequently called in the reigns of John I, Edward, and Afonso V; but the avenues of power had become wider by the 16th century, and John III's proposal (1525) to call them only every 10 years aroused no opposition. Although the trade guilds were slow in developing, they took some part in determining local taxation in the 13th century. Trade increased, Portuguese merchants having had connections with the Low Countries from the time of Afonso Henriques and with England from the early 13th century. The political crisis of 1385 was followed by inflation and debasements: there was thereafter no national gold currency until 1435, when West African sources began to be tapped.

THE DISCOVERIES AND THE EMPIRE

Conquest and exploration. The idea of expansion into Africa was a logical result of the completion of the Reconquest in the peninsula, and the conquest of Ceuta in North Africa (1415) probably provided the impulse toward further expansion. The simple idea of fighting the Muslims on their own soil was linked with more complicated motives: the desire to explore in a scientific sense, the hope of finding a way to the rich spice trade of the Indies, and the impulse to spread the Christian faith. These purposes were gradually molded together into a national enterprise, though at first they represented the hopes and aspirations of one man, Prince Henry. The third son of John I and Philippa of Lancaster, known rather inaccurately as "the Navigator" (he himself never went farther afield than Tangier), Henry became (1420) master of the Order of Christ, which King Dinis had founded (1319). The resources of the order were used to draw together skilled geographers and navigators and to equip a series of expeditions that only gradually began to bear fruit.

Henry the Navigator

The date of Prince Henry's earliest expedition was about 1418, when the island of Porto Santo was visited; the first call at Madeira probably dates from 1419. An attempt was made to settle in the Canaries, and between 1427 and 1431 the Azores were visited by Portuguese seamen. Both the Azores and Madeira were then uninhabited, and their colonization proceeded fairly rapidly from about 1445. Sugar was exported to Europe and gave the islands great economic importance. Meanwhile, Prince Henry's ships were probing the African coast, passing Cape Bojador in 1434 and Rio de Oro in 1436. The unsuccessful expedition against Tangier (1437) was followed by a break in the discoveries, but in 1439 Prince Henry was authorized to colonize the Azores; from 1440 further expeditions equipped with a new and lighter ship, the caravel, reached the Bay of Arguin (1443) and Cape Verde (1444) and by Henry's death (1460) had explored the coast as far south as Sierra Leone.

Under Afonso V, three military expeditions were sent against Morocco (1458, 1463, and 1471); by the last of them, Tangier and Arzila were captured. The African explorations were not entirely neglected, but it remained for John II, with his sharp sense of the national interest, to found a fortress and trading post in the Gulf of Guinea at Elmina (São Jorge da Mina, 1481–82). Diogo Cão discovered the mouth of the Congo in 1482 and then advanced to Cape Cross, 200 leagues southward (1486). An African monarchy, the Kongo, sought conversion and alliance with the Portuguese: its first Christian king, Afonso I (*c.* 1506–43), made Mbanza a centre of Portuguese influence, but the Kongo kingdom fell into internal strife, and Portuguese interests were transferred to the neighbouring kingdom of Angola. Paulo Dias de Novais founded Luanda, the first city in West Africa south of the Equator, in 1576. In 1488 Bartolomeu Dias rounded the Cape of Good Hope and reached the East African coast, and the seaway to India lay open. His return was followed in 1493 with the news that Christopher Columbus had, as he thought, discovered the "Indies" by sailing across the Atlantic. Much as this news must have perturbed the Portuguese, Columbus brought no news of the spiceries or the cities of the East; and John II ordered the preparation of an expedition to India by way of the Cape of Good Hope, though this sailed only after his death. John also contested the Spanish claim to all lands discovered west of the Atlantic, and by the Treaty of Tordesillas (June 7, 1494) Spain's rights were limited to what lay more than 370 leagues west of the Cape Verde Islands. Thus the territory that was to become Brazil was reserved for Portugal.

Bartolomeu Dias rounds the Cape of Good Hope

The Treaty of Tordesillas had confirmed Portugal's rights to the exploration of Africa and the seaway to India. In July 1497 Vasco da Gama set sail with four ships on the first expedition to India. They reached Calicut (Kozhikode) the following spring, and the survivors put into Lisbon in the autumn of 1499 with specimens of Oriental merchandise. A second fleet was prepared under Pedro Álvares Cabral, who touched the Brazilian coast (April 22, 1500) and claimed it for Portugal.

Control of the sea trade. In 1505 Francisco de Almeida arrived as viceroy of India and supported the ruler of Cochin against the *zamorin* of Calicut. The control of sea trade, the chief source of Portuguese wealth in the East, was assured by the defeat of Muslim naval forces off Diu in 1509. Almeida's successor, Afonso de Albuquerque, conquered Goa (1510), which he made the seat of Portuguese power, and Malacca (1511); sent two expeditions to the Moluccas (1512 and 1514); and captured Hormuz in the Persian Gulf (1515). Soon after, Fernão Peres de Andrade reached Canton in China; in 1542 Portuguese merchants were permitted to settle at Liampo (Ning-po), and in 1557 they founded the colony of Macau (Macao).

Albuquerque was responsible for this conception of a system of strong points that secured the trade of the Orient to Portugal for nearly a century. Goa soon became the chief port of western India; Hormuz controlled the Persian Gulf, and Malacca the gateway from the Indian Ocean to the South China Sea, while a string of fortified trading posts secured the coast of East Africa and the gulf and shores of India and Ceylon. Farther east, less fortified settlements were set up with the consent of the native rulers from Bengal to China, and the trade of the principal spice islands was in Portuguese hands. The preservation of the whole system was entrusted to a governor, who sometimes held the rank of viceroy, at Goa; and, although Portuguese arms had both triumphs and reverses, their control of the Oriental trade remained substantial, if never complete, until the 17th century, when the Dutch, at war with the joint crown of Portugal and Spain and deprived of their traditional trade with Lisbon, began to seek spices from their source and effectively demolished the Portuguese monopoly.

UNION OF SPAIN AND PORTUGAL, 1580–1640

After Philip II of Spain had occupied Portugal in 1580, the island of Terceira in the Azores held out for António of Crato, who himself sought alliances in England and France. In 1582 a French expedition to establish him in the Azores was defeated, and in 1589 an English attempt upon Lisbon, led by Sir Francis Drake and Sir John Norris, failed dismally. But, although António died in Paris in 1595, the true symbol of Portuguese independence was not the prior of Crato but King Sebastian himself. The Portuguese people refused to believe that he was dead and nourished a messianic faith in his reappearance, of which four pretenders sought to avail themselves, the last as late as 1600 and as far afield as Venice.

Meanwhile, Philip arrived in Portugal and was accepted as King Philip I (1580–98) by the Cortes held at Tomar (1581). He undertook to preserve Portuguese autonomy, to consider the union as a personal one like that of Aragon and Castile under Ferdinand and Isabella, to appoint only Portuguese to the administration, to summon Cortes frequently, and to be accompanied by a Portuguese council in Madrid. These undertakings were, however, neglected by Philip II (III of Spain, 1598–1621) and completely violated by Philip III (IV of Spain, 1621–40).

Exploitation by the Spanish

Portuguese resentment against Spanish rule was increased by the failure of these kings to visit Portugal, the appointment of Spaniards to Portuguese offices, the loss of trade

consequent on Spain's foreign wars, and the levying of taxation to sustain these wars. In 1624 the Dutch seized Bahia in Brazil, only to be expelled by a joint Spanish and Portuguese expedition (1625). But in 1630 they occupied Pernambuco and the adjoining sugar estates, which they held for a generation. The final straw was the Count-Duke de Olivares' plan (1640) of using Portuguese troops against the equally discontented Catalans. Two Portuguese insurrections, in 1634 and 1637, had failed to attain dangerous proportions, but in 1640 Spain's powers were extended to the utmost by war with France and revolt in Catalonia. The French minister, Cardinal de Richelieu, already had agents in Lisbon, and a leader was found in John, duke of Bragança, a grandson of the duchess Catherine (niece of John III) whose claims had been overridden in 1580 by Philip II of Spain. Taking advantage of the unpopularity of the governor, Margaret of Savoy, duchess of Mantua, and her secretary of state Miguel de Vasconcelos, the leaders of the party of independence carried through a nationalist revolution on Dec. 1, 1640. Vasconcelos was almost the only victim; the Spanish garrisons were driven out; and on December 15 the duke of Bragança was crowned as John IV (1640–56).

THE HOUSE OF BRAGANÇA, 1640–1910

The success of the new regime was not finally assured until 1668, when Spain at last recognized Portuguese independence. Faced with the threat of a Spanish invasion, John had sent missions to the courts of Europe in quest of alliances. France now refused a formal treaty. The Dutch, having seized northern Brazil, accepted a truce in Europe and proceeded to capture Angola from Portugal. John made a treaty (1642) with Charles I of England, which was made void by Charles's execution (1649). Meanwhile, the Portuguese defeated the Spaniards at Montijo (May 26, 1644) and warded off several invasions. In 1654 they negotiated a treaty with the English Commonwealth, obtaining aid in return for commercial concessions. The Dutch were finally expelled from Pernambuco in northern Brazil. By a secret article of the Treaty of the Pyrenees (1659), France promised Spain to give no further aid to Portugal, but in 1661 Portugal signed a treaty of alliance with the restored English monarchy. In 1662 Charles II of England married John's daughter Catherine of Bragança and, in return for a large dowry, including the cession of Bombay and Tangier, provided arms and men for the war with Spain. Portuguese defense was organized by the German soldier Friedrich Hermann von Schönberg (later duke of Schomberg); in June 1663 Sancho Manuel, count of Vila Flor, defeated Don John of Austria at Ameixial, and in June 1665 von Schönberg won the important victory of Montes Claros. Peace was finally made by the Treaty of Lisbon early in 1668.

On the death of John IV, his second son Afonso VI (1656–83) was 13 years of age. Afonso's mother, Luísa de Gusmão, acted as regent until, in June 1662, he began to rule. Afonso himself was feebleminded, but the country was capably governed by Luiz de Vasconcelos e Sousa, Count de Castelo Melhor, until 1667. The French princess, Maria Francesca of Savoy, who married Afonso in the previous year, now entered into an intrigue with his more personable brother Pedro (afterward Peter, or Pedro, II). They contrived to dismiss Castelo Melhor and to have Maria Francesca's marriage annulled. She at once married Pedro (1668), who was declared regent. Afonso was imprisoned until his death. During the reign of Peter II (1683–1706), Portugal recovered from the strain of the Spanish wars and began to benefit from the discovery of gold and precious stones in Brazil. The first strike of gold in Minas Gerais took place in 1693, and, in the last years of the 17th century, considerable wealth was being extracted; however, it was not until 1728 that diamonds were discovered, the mineral wealth of Brazil then forming a very substantial part of the revenue of the Portuguese crown.

Mineral wealth from Brazil

The 18th century. The War of the Spanish Succession (1701–13) saw Portugal's recent friends, England and France, on opposing sides; and, although Pedro sought at first to remain neutral, Portugal joined the Anglo-Austrian Grand Alliance in 1703, by which it afforded a base for the archduke Charles (later the emperor Charles VI) to conduct his war for the Spanish throne. In the same year (December 27), the English envoy, John Methuen, also concluded the treaty that bears his name, by which the exchange of port wine for English woolens became the basis for Anglo-Portuguese trade. Although the treaty of 1654 had secured great privileges for English merchants in Lisbon, neither it nor the treaties of 1642 and 1661, by which the traditional alliance was restored, had created trade. This was now done, and, by reason of the wealth that soon poured into Lisbon from Brazil, the English merchants gained a commanding position in the trade of Portugal. The political treaties of 1703 proved less fruitful. The Portuguese general António Luís de Sousa, marquess of Minas, entered Madrid in 1706; but French and Spanish forces were victorious at Almansa (1707), and in 1711 the French admiral René Duguay-Trouin sacked Rio de Janeiro. At the conclusion of the war, Portugal negotiated a peace treaty with France (April 1713) but did not conclude peace with Spain until 1715.

Under Peter's son John V (1706–50), Portugal attained a degree of prosperity unknown since the restoration. The tax of a royal fifth levied on the precious metals and stones of Brazil gave the monarchy an independent source of wealth. The Cortes, which had met irregularly since 1640, was no longer summoned, and government was carried out by ministers appointed by the king. John V desired the absolute authority enjoyed by Louis XIV. He did not fail to convert his wealth into papal and other dignities: the archbishop of Lisbon became a patriarch (1716), and Pope Benedict XIV in 1749 gave John the title "his most faithful majesty"; and royal academies, palaces, and libraries were inaugurated. But in his later years his ministers proved inadequate, and the kingdom sank into stagnation. On John's death, his son Joseph (1750–77) appointed as minister Sebastião José de Carvalho e Melo, later Count de Oeiras and Marquês de Pombal, who soon gained a complete ascendancy over the king and endeavoured to replace the stagnant absolutism with a more active type of despotism that, with some qualifications, deserves the epithet enlightened. His full powers date from his efficient handling of the crisis caused by the disastrous Lisbon earthquake of November 1755; but even before this he had reformed the sugar and diamond trades, set up a national silk industry (1750), and formed one chartered company to control the sardine- and tunny-fishing industry of Algarve and another to trade with northern Brazil. In 1756 he founded a board of trade with powers to limit the privileges enjoyed by the English merchants under the treaties of 1654 and 1661 and set up the General Company for Wines of Alto Douro to control the port-wine trade. Industries for the manufacture of hats (1759), cutlery (1764), and other articles were set up with varying success.

Enlightened despotism of Pombal

Pombal's methods were arbitrary and his enemies numerous. His reform of the wine industry provoked a riot in Porto (1757) that was savagely repressed; but his principal victims were the Jesuits, expelled in 1759 from all the Portuguese dominions, and the nobility, in particular José Mascarenhas, Duke de Aveiro, and the Távora family, who were accused of an attack on the king (Sept. 3, 1758), condemned, and executed (Jan. 12, 1759). Having eliminated the Jesuits from the educational system, Pombal applied regalist principles in the reform of the University of Coimbra (1772) and the royal board of censorship (1768), which supervised the system of lower education from 1771.

While Pombal succeeded in modifying the ascendancy of the British merchants in Portugal, he invoked the English alliance in 1762 when Spain, prompted by the renewal of the Bourbon Family Compact alliance with France, invaded Portugal. The Portuguese army was reformed by Wilhelm von Schaumburg-Lippe, and an English force was led by James O'Hara, 2nd Baron Tyrawley, and John Campbell, 4th Earl of Loudoun. Peace was signed in February 1763 at Fontainebleau. On the death of Joseph (Feb. 24, 1777), his daughter Maria I (1777–1816), who had married his brother, her uncle (Peter III), acceded;

Pombal was dismissed (1777) and eventually found guilty on several charges. His successors made peace with Spain by the Treaty of San Ildefonso (1777).

The French Revolutionary and Napoleonic Wars. Maria I suffered from melancholia after the loss of her consort (1786) and her eldest son John (1788). In 1792 her mental instability increased following news of the French Revolution, and she ceased to reign. Her son, who on her death became John VI (1816–26), then ruled in her name and in 1799 became prince regent. In 1793 Portugal joined England and Spain against France, sending a naval division to assist the English Mediterranean fleet and an army to the Catalan front. The Peace of Basel (July 1795), by which Spain abandoned its allies, left Portugal still at war. Although subjected to pressure from the French Directory and from the Spanish minister, Manuel de Godoy, Portugal remained unmolested until 1801, when Godoy sent an ultimatum and invaded Alentejo. By the Peace of Badajoz (June 1801), Portugal lost the town of Olivenza and paid an indemnity.

From the Peace of Amiens (1802) until 1807, Portugal was once more immune from attack, though subjected to continuous pressure to break off the English connection. Napoleon sought to close all continental ports to British ships, but Portugal endeavoured to maintain neutrality. The secret Franco-Spanish Treaty of Fontainebleau (October 1807) provided for its eventual dismemberment by Napoleon I and Godoy. Already General Andoche Junot was hastening across Spain with a French army, and on November 27 the prince regent and the royal family and court embarked on a fleet lying in the Tagus River and were escorted by British vessels to Brazil; the court remained 14 years at Rio de Janeiro. Junot declared the Braganças deposed, but his occupation of Portugal was challenged in August 1808 by the arrival of Sir Arthur Wellesley (later Duke of Wellington) and 13,500 British troops in Mondego Bay. Winning the victories of Roliça (August 17) and Vimeiro (August 21), Wellesley enabled his superiors to negotiate the Convention of Sintra (August 31), by which Junot was allowed to evacuate Portugal with his army. A second French invasion (1808–09) led to Sir John Moore's death at La Coruña (January 1809) and the reembarkation of the British forces. In February William Carr (later Viscount) Beresford was placed in command of the Portuguese army, and in March Marshal Nicolas-Jean de Dieu Soult advanced from Galicia and occupied Porto. Wellesley returned to Portugal in April, drove Soult from the north, and, after his victory of Talavera de la Reina in Spain (July), withdrew to Portugal. The third French invasion followed in August 1810 when Marshal André Masséna with Marshal Michel Ney and Junot entered Beira province. Defeated by Wellington at Bussaco (September 27) near Coimbra, the French found themselves facing the prepared lines of Torres Vedras, north of Lisbon, where they wintered amid great privations. By the spring of 1811 they could only retreat and, on March 5, began the evacuation of Portugal, harassed all the way by English and Portuguese attacks and crossing the frontier after a defeat at Sabugal (April 3).

Flight of the royal family to Brazil

Portugal made peace with France on May 30, 1814. It was represented at the Congress of Vienna but played little part in the settlement. The series of Anglo-Portuguese treaties concluded between the years 1809 and 1817, however, was important insofar as it extended many of the conditions of the Anglo-Portuguese alliance to Brazil and had an influence on the future of Africa. England's efforts to get Portuguese collaboration in suppressing the slave trade resulted in the treaty of Jan. 22, 1815, and in the additional convention of 1817, by reason of which Portugal's claims to a considerable part of Africa were formally recognized.

Constitutionalism. The Napoleonic campaigns caused great devastation in Portugal, and the absence of the royal family and the presence of a foreign commander (Beresford) combined with revolutionary agitation and the influence of Spanish liberalism to produce an atmosphere of discontent. On Dec. 16, 1815, Brazil was raised to the rank of a kingdom united with Portugal, and John VI, who succeeded in March 1816, showed no desire to return to Portugal. In 1817 Beresford suppressed a conspiracy in Lisbon, and the Masonic leader General Gomes Freire de Andrade was executed. Unrest increased, and, when Beresford himself went to Brazil (March 1820) to press John to return, a constitutionalist revolution began in Porto (Aug. 24, 1820), spread over the country, and led to the formation of a junta in Lisbon (October 4). On Beresford's return (October 10) he was not allowed to land, and British officers were expelled from the army. A constituent assembly was summoned that drew up a very liberal constitution, thus confronting John VI with an accomplished fact.

John's reluctance to return was at last overcome, and he left his elder son Pedro to govern Brazil, landing at Lisbon on July 3, 1821. He swore to uphold the constitution, but his wife, Carlota Joaquina, and their second son, Dom Miguel, refused to take the oath and were sentenced to banishment, though this was not carried out. The Portuguese constitutionalists, not appreciating the determination of Brazil not to yield up its status as a kingdom, sought to compel Pedro to return; but he, rather than sacrifice the rule of the Braganças in Brazil, declared for Brazilian independence (Sept. 7, 1822) and became emperor of Brazil as Pedro I. This enabled his brother Miguel (Michael) to appeal to absolutist forces in Portugal to overthrow the constitutionalists, and an insurrection led by Miguel almost succeeded (April 30, 1824); but, through the action of the foreign ministers, John VI was restored and Miguel went into exile in Vienna (June 1824).

Independence of Brazil

The War of the Two Brothers. John VI acknowledged the independence of Brazil in 1825, assuming pro forma the imperial title and then yielding it to Pedro; but when John died (March 10, 1826) no provision had been made for the succession except that his daughter Maria Isabel was named regent. Pedro, as Pedro, or Peter, IV of Portugal, issued from Brazil a charter providing for a parliamentary regime by the authorization of the monarchy and not based on the sovereignty of the people. He then made a conditional abdication (May 1826) of the Portuguese throne in favour of his daughter Maria da Glória, aged seven, provided that she should marry her uncle Miguel and that he should swear to accept the charter. This compromise could not be effective. The absolutists had hoped that Pedro would resign all rights to the Portuguese crown, and the council of regency hesitated to publish the charter until General João Carlos de Saldanha (later Duke de Saldanha) forced their hand. In 1827 Miguel took the oath and was appointed regent, landing in Lisbon in February 1828. His supporters at once began to persecute the liberals. A form of Cortes met in Lisbon and in July 1828 repudiated Pedro's claims and declared Miguel rightful king.

Only the island of Terceira in the Azores sustained the liberal cause. In June 1829, however, a regency on behalf of Maria da Glória was set up in Terceira; and in 1831 Pedro, having abdicated the Brazilian throne, went to Europe and began to raise money and men for the conquest of Portugal. In February 1832 the expedition sailed to Terceira, and in July the liberals, led by Pedro, disembarked at Mindelo near Porto, which city they soon occupied. The rest of the country, however, stood by Miguel, who besieged the liberals in Porto for a year (July 1832–July 1833). By then, Miguelite enthusiasm had waned; and António José de Sousa Manuel, Duke de Terceira, and Captain (later Sir) Charles Napier, who had taken command of the liberal navy, made a successful landing in Algarve (June 1833). Terceira advanced on Lisbon, which fell in July 1833, and Miguel capitulated at Evora-Monte in May 1834.

Further political strife. The War of the Two Brothers ended with the exile of Miguel (June) and the death of Pedro (Sept. 24, 1834). Maria da Glória became queen as Maria II (1834–53) at the age of 15. While Maria necessarily came under the influence of the successful generals of the civil war, her principal aim was to defend her father's charter (which had been granted by the crown) from those who demanded a "democratic" constitution like that of 1822, asserting the ultimate sovereignty of the nation.

In September 1836 the latter, thenceforth called Septem-

Septembrists and chartists

brists, seized power. The chartist leaders rebelled and were exiled, but by 1842 the Septembrist front was no longer united, and António Bernardo da Costa Cabral restored the charter. In 1846 the movement of Maria da Fonte, a popular rising against higher taxation to improve roads and reforms in public health in which almost all parties joined, put an end to Costa Cabral's government but left Portugal divided between the Septembrists, who held Porto, and Saldanha, now in Queen Maria's confidence, in Lisbon. Saldanha negotiated for the intervention of other members of the Quadruple Alliance (formed in April 1834 by England, France, Spain, and Portugal), and a combined British and Spanish force received the surrender of the Porto junta in June 1847 and ended the war with the Convention of Gramido (June 29, 1847). Saldanha governed until 1849, when Costa Cabral resumed office only to be overthrown in April 1851. Saldanha then held office for five years (1851–56), and the period of peace at length permitted the country to settle down. The "Regeneration" ended civil strife and established party government.

Maria II was succeeded by her eldest son by her second husband, Ferdinand of Saxe-Coburg, Peter, or Pedro, V (1853–61), who married Stephanie of Hohenzollern-Sigmaringen in 1858. He gave promise of being a capable monarch but died of typhoid fever on Nov. 11, 1861. His brother Louis (Luís; 1861–89) seemed to have inherited a country that had recovered from the Napoleonic invasions and from civil wars, political strife, and *pronunciamentos*. But, although the main parties were now defined as Historicals (*i.e.*, radicals) and Regenerators (or moderates), the alternation of governments gradually ceased to reflect public feeling, and in the last years of Luís' reign republicanism began to gain ground. (H.V.L.)

African colonies

Overseas empire. The loss of Brazil to independence in 1822 left Portugal's overseas empire a largely African one, with scattered small Asian holdings in western India (mainly Goa, Daman, and Diu), Indonesia (East Timor), and South China (Macau). Beginning in 1836, Portugal pursued a policy of African territorial expansion and economic enhancement, concentrating first on Angola and later on Mozambique and Portuguese Guinea. Portugal's role in the European partition of Africa that took place in the fourth quarter of the 19th century was limited by its long-standing economic dependence on Great Britain. A colonial movement gained momentum in Lisbon, and a Portuguese scheme known as the "Rose-Coloured Map," which laid claim to a colony stretching across Africa from Angola to Mozambique, was recognized by France and Germany in 1886. But Britain challenged Portugal's claim to territory in southeastern Africa in what are now Malawi and Zimbabwe and issued an ultimatum, dated Jan. 11, 1890, demanding the immediate withdrawal of Portuguese forces from the disputed regions. The ultimatum implicitly threatened, in the short run, the use of British naval force and, in the long run, an end to the venerable Anglo-Portuguese alliance, which dated from the late 14th century. The Lisbon government, obliged to comply, subsequently resigned. The ultimatum crisis rocked Portuguese public opinion, heightening imperial fever in the capital and principal towns, compromising the monarchy of King Charles I, and bolstering the oppositionist republicans.

Despite the failure of the central-corridor plan, Portugal retained a large African empire (about 8 percent of the continent), third in size after those of Britain and France. Two Anglo-Portuguese agreements—the 1891 boundary treaty and the so-called Windsor Treaty of Oct. 14, 1899, respectively—safeguarded Portugal's sovereignty over its existing colonies and reaffirmed the ancient alliance.

The rise of republicanism. During the period 1890–1910 the relatively stable politics of rotating governments under the constitutional monarchy disintegrated. Feuding monarchist parties and politicians agreed as to the severity of economic, financial, and social problems but quarreled over solutions. The republicans increased their hold on public opinion in Lisbon and the larger towns as well as in the rural south. In 1906 João Franco, a prominent politician with reformist plans for saving the monarchy, won appointment as premier. Unable to unite the factious monarchists, he began to govern by decree. Franco boldly undertook to reform finance and administration but was accused of illegal money transfers to Charles. These scandals were followed by rumours of plots, and on Feb. 1, 1908, Charles and his heir, Luís Filipe, were assassinated in an open carriage in the streets of Lisbon. Whether the regicides were isolated fanatics or agents of a hidden organization such as the Carbonária, a republican secret society, the deed was applauded by the republicans, who immediately began their preparations for a final attack on the monarchy.

Only 18 years of age at his accession, Charles's younger son, as King Manuel II (1908–10), was ill-equipped to solidify the crumbling monarchist factions. In the general elections of August 1910, the key cities of Lisbon and Porto voted in favour of a republic. On October 3 the murder of a leading republican figure, the distinguished psychiatrist Miguel Bombarda, by a mad patient offered the pretext for a rising that was already organized. Armed civilians, soldiers, and the men aboard some ships in the Tagus, under the leadership of a key Carbonária figure named António Machado Santos, began the republican revolution on October 4; the next morning, Portugal's first republic was declared from the balcony of the Lisbon City Hall. Manuel escaped via his yacht to Gibraltar and thence to England, where he remained in exile until his death in 1932.

Social and economic conditions. From the time of the Portuguese overseas conquests, the flow of wealth from the new territories and trading posts, if it sustained the court and capital, did little to improve the domestic economy, still largely rural. The favourable financial position of the late 15th century, derived from trade in slaves, gold, and spices, did not long survive into the 16th century, when the expenses of maintaining far-flung and unproductive foreign stations and the depredations of pirates quickly absorbed any surpluses. There were few native industries. Not only were manufactured goods such as cloth, tapestry, and metalware imported but also basic foodstuffs, salt meat, cured fish, and dairy produce. Agriculture was little regarded, and insufficient land was available for small holdings. During the years of Spanish domination, the ports were closed to English merchants; and by the time they were reopened after 1640 the flow of trade had found new channels, and the Dutch and English had outstripped Portugal as colonial powers. The discovery of gold in Brazil at the end of the 17th century brought about a revival in the country's economy, but gold production was in decline by 1750, while the diamond market was saturated. The later 18th century saw a series of protectionist measures, many introduced by Pombal. The Methuen Treaty with England (1703) had strengthened the port wine trade at the expense of Portuguese cloth; later further attempts were made to improve the export value of port wine. Support was also given for the production of woolen goods, linen, paper, porcelain, and cutlery and to the tunny and sardine fisheries. Pombal attempted to create an educated bourgeoisie, but the Portuguese textile industry could not withstand mechanized competition. After 1850 public works, railways, and ports were given priority, but Portugal had to wait until the 20th century for any sustained attack upon its economic difficulties.

Economic difficulties

THE FIRST REPUBLIC, 1910–26

The new regime formed a provisional government under the presidency of Teófilo Braga, a well-known writer. A new electoral law was issued giving the vote only to a restricted number of adult males. The provisional government presided over the election of a constituent assembly, which opened on June 19, 1911. The constitution was passed by the assembly on August 20, and the provisional government surrendered its authority a few days later (August 24) to the new president, Manuel José de Arriaga. Despite initial hopes that it would solve the massive problems inherited from the monarchy, the republic soon became western Europe's most turbulent, unstable parliamentary regime.

Although a monarchist invasion was unsuccessfully attempted by Henrique de Paiva Couceiro in October 1911, the main danger to the new regime came from its in-

Patterns of division

ternal divisions. For the moment, it was fairly united in the actions of abolishing the monarchy and disestablishing the Roman Catholic church. The religious orders were expelled (Oct. 8, 1910) and their property confiscated. New legislation banned the teaching of religion in schools and universities and annulled many religious holidays. Persecution of Catholics in the early years of the republic attracted international attention and brought the new political system into conflict with foreign diplomats, humanitarian organizations, and journalists. While the government initiated advances in education, health, civic freedoms, and colonial development, positive results were overwhelmed by administrative instability, labour unrest, public violence, and military intervention in politics.

By 1912 the republicans were divided into Evolutionists (moderates), led by António José de Almeida; Unionists (centre party), led by Manuel de Brito Camacho; and Democrats (the leftist core of the original party), led by Afonso Augusto da Costa. A number of prominent republicans had no specific party. The whirligig of republican political life offered little improvement on the monarchist regime, and in 1915 the army showed signs of restlessness. General Pimenta de Castro formed a military government and permitted the monarchists to reorganize, but a Democratic coup in May led to his arrest and consignment to the Azores, along with Machado Santos. Dominated by Afonso Costa's oratory, partisan press, and political machine, the Democrats' regime was in turn overthrown by another bloody military coup (December 1917), led by the former minister to Germany, Major Sidónio Pais.

The authoritarian, unstable "New Republic" of charismatic President Pais failed to pacify the feuding factions, and its collapse precipitated a brief civil war. Following Pais's assassination in Lisbon (Dec. 14, 1918), republicans and monarchists fought a civil war (January 1919) in which the monarchists' final armed effort to restore the monarchy failed, and political power was restored to the chastened Democrats. Four key tensions characterized the republic's troubled political system: (1) excessive factionalism, (2) the tendency of the factionaries to bear allegiance to personalities rather than to ideas, institutions, and the public interest, (3) disparity between the landholding patterns of the north (typified by *minifúndias*—small subsistence farms) and the south (typified by *latifúndias*—large estates worked by landless peasants), and (4) the concentration of economic development in Lisbon, at the expense of the provinces.

On the outbreak of World War I, Portugal had proclaimed its adhesion to the English alliance (Aug. 7, 1914) and on November 23 committed itself to military operations against Germany. On September 11 the first expedition left to reinforce the African colonies, and there was fighting in northern Mozambique, on the Tanganyika frontier, and in southern Angola, on the frontier of German South West Africa. In compliance with a request from Britain, in February 1916 Portugal seized German ships lying in Portuguese ports, and Germany declared war on Portugal (March 9). A Portuguese expeditionary force went to Flanders in 1917, under General Fernando Tamagnini de Abreu, and on April 9, 1918, the Germans mounted a major attack in the Battle of the Lys. Although the Allies won the war, and Portugal's colonies were safeguarded, the 0.75 percent of the war indemnity paid by Germany to Portugal was scant compensation for the heavy costs incurred, both in the field and at home: the casualties of the African campaigns and the western front; the alienation of a portion of the army officer corps; crippling war debts to Britain; intense inflation; and a scarcity of food and fuel.

Former Evolutionist Almeida became the only president to complete his term during the first republic, but the cycles of bankruptcy, corruption, public violence, and military insurrectionism continued. Finally, on May 28, 1926, the parliamentary republic was overthrown in a bloodless military coup that instituted what was to become western Europe's most long-lived authoritarian system.

THE DICTATORSHIP, 1926–74

The Salazar regime. The provisional military government was shortly taken over by General António Oscar de Fragoso Carmona, who favoured sweeping changes. In the face of financial crisis, Carmona appointed António de Oliveira Salazar minister of finance with full powers over expenditure. A prominent professor of economics at the University of Coimbra, Salazar assembled a civilian elite of intellectuals and bureaucrats to steer the course of recovery. Budgetary surpluses became the hallmark of his regime, making possible large expenditures for social programs, rearmament, and infrastructure development. This speedy progress, coupled with personal austerity and hard work, won Salazar the grudging collaboration of diverse parties and interest groups that included monarchists, conservative republicans, fascists, pseudofascists, nationalists, the church, business leaders, land barons, and the military establishment. As minister of colonies in 1930, he prepared the Colonial Act, assimilating the administration of the overseas territories to his system. In July 1932 he became prime minister, a post he was to hold (along with other key ministries during crises) until 1968. He served a brief interim term as president in 1951.

The Estado Novo

The new constitution of 1933 declared Portugal a "unitary, corporatist republic." Salazar's New State (Estado Novo) provided for a National Assembly, with deputies elected quadrennially as a bloc, and a Corporative Chamber comprising representatives of occupations. All seats in the assembly went to government supporters; the Corporative Chamber did not come into being until employers' and workers' syndicates were formed. The government regulated labour-management relations, banned strikes and lockouts, and monitored social welfare planning. Political parties were banned, and all eligible voters were encouraged to join the National Union, an approved loyalist movement. Ever mindful of the confusion that preceded it, the New State emphasized order over freedom and attempted to "neutralize" society through the use of censorship, propaganda, and political imprisonment. On the other hand, it partially restored the pre-1910 privileges of the church in law, society, and education.

During the Spanish Civil War (1936–39), Salazar backed General Franco's Nationalists, who triumphed and controlled all of Spain by the spring of 1939. In World War II, Portugal maintained official neutrality (while quietly favouring Britain) until Britain invoked the ancient Anglo-Portuguese alliance to obtain bases in the Azores. Portugal joined the North Atlantic Treaty Organization (NATO) as a founding member in 1949 but did not gain admission to the United Nations until 1955.

Portugal's foreign and colonial policies met with increasing difficulty both at home and abroad beginning in the 1950s. In the presidential election of 1958, oppositionist General Humberto Delgado generated political heat despite the futility of challenging the regime's candidate, Admiral Américo Tomás. On the world scene, the tensions of the Cold War gave Portugal's largely undeveloped overseas empire a new significance. The determination of the Indian government to annex Portuguese India led to a severing of diplomatic relations (August 1955) and to mass invasions of the Portuguese possessions by Indian passive resisters. Portugal effectively lost Dādra and Nagar Haveli to India (despite a ruling by the International Court of Justice in April 1960 favouring Portugal), and on Dec. 19, 1961, India took over Goa, Diu, and Daman.

Salazar had made it plain that decolonization was not part of his plan for Portugal, and, when Angola was the scene of disturbances in which a number of lives were lost, he reinforced the troops in the African territories and took over the Ministry of Defense. In 1964 there was a recrudescence of independence movements.

Economic development

Perhaps the most important contribution of the New State was in the economy, despite the failure to rejuvenate agriculture and a reluctance to industrialize. Development plans, closely monitored by the demanding Salazar, were conservative but consistent: The government significantly reduced its debt; diminished its economic dependence on British investment; tightly controlled foreign investment and did not openly encourage it until 1964–65, when expensive wars in three African "overseas provinces" (colonies)—Angola, Mozambique, and Portuguese Guinea—prompted a revision of the investment

code; supported industry but not massively; and emphasized infrastructure development over health, education, and welfare. From about 1960 until the inflation surge and energy crisis of 1973, Portugal experienced economic growth at an annual rate of 5 to 7 percent, which constituted a boom for western Europe's poorest nation.

The New State after Salazar. In September 1968 Salazar was incapacitated by a stroke. President Tomás invited Marcelo Caetano, one of the architects of the New State, to form a government, but Salazar was never informed of this transition. On July 27, 1970, Salazar died. Although Caetano was more worldly and less reserved than his predecessor, he proved unable to reverse the tide of Portugal's African wars, to alleviate the economic woes of 1973–74, or to avert revolution.

PORTUGAL SINCE 1974

The revolution. Two developments galvanized the movement that was shortly to topple the dictatorship. The first occurred in mid-1973, when career army officers became alienated by a government measure commissioning militia officers for service in the colonial wars. The second incitement was the publication in February 1974 of the book *Portugal e o Futuro* ("Portugal and the Future") by the colonial war hero General Antônio de Spínola, who argued that the wars in Africa could not be settled by force of arms and advocated negotiated autonomy for the colonies and an alternative to Caetano's leadership. Some 200–300 officers, calling themselves the Armed Forces Movement (Movimento das Forças Armadas, or MFA), led by Francisco da Costa Gomes and other officers, planned and implemented the coup of April 25, 1974, which came to be known as the Revolution of the Carnations because it occurred during the week that red carnations appeared in the flower shops.

The overthrow met with little resistance from the dictatorship's remaining loyalists and won initial support from an urban middle class vexed by economic and political uncertainty. The transition to a functioning, consolidating, pluralist Portuguese democracy mirrored, in a nonviolent way, the political course of the French Revolution: an early moderate-conservative phase (May 1974–March 1975) followed by a middle radical-leftist phase (March–late November 1975) and a final moderate reaction (late November 1975–June 1976).

After exiling Caetano and Tomás, a subgroup of the MFA calling itself the Junta of National Salvation filled the political vacuum, installing Spínola as president and commencing negotiation with the African nationalist movements. Independence was granted to Portuguese Guinea (as Guinea-Bissau) almost immediately after the revolution. The new regime abolished such instruments of repression as censorship, the paramilitary forces, and the secret police. Provisional President Spínola, who opposed rapid independence for the colonies without free referendums, resigned his office in September 1974, launched a countercoup attempt that failed (March 1975), and fled to exile.

By this time, radical MFA elements and their leftist civilian allies in the Portuguese Communist Party and other Marxist-Leninist groups had won virtual control over the government in Lisbon, sections of the armed forces, and the media. The MFA itself was restructured and a Council of the Revolution installed with the support of six political parties. A general election for the Constituent Assembly in April 1975 drew 92 percent of eligible voters, a record in western European history. The decolonization of the Cape Verde Islands and Mozambique was effected in July 1975. Portugal's remaining African territories achieved independence later the same year, thus ending a colonial involvement in that continent that had begun in 1415. Full-scale, internationalized civil war, however, followed Portugal's departure from Angola, and Indonesia forcibly annexed briefly independent East Timor.

The end of the colonies

Political and social instability prevailed through most of 1975: From the former African colonies, nearly a million people fled to Portugal, adding a refugee problem to the already volatile domestic situation; some 30 persons died in incidents of public violence; new political parties proliferated; and strikes were widespread. Also in that year, the government nationalized banking, transport, heavy industries, and the media. In Alentejo province in southern Portugal, farm workers expropriated *latifúndias* and established communal farming. On Nov. 25, 1975, moderate military elements crushed a radical leftist coup in the army and restored order.

The 1976 constitution and subsequent reforms. In April 1976 the Constituent Assembly approved a new constitution, which committed Portugal to making a transition to socialism. The parliamentary elections of April 25 produced no majority party; the Socialists, the Popular Democrats (centre-right), the Social Democratic Centre Party (conservative), and the Communist Party (founded 1921) made the strongest showings, and the Socialist leader, Mário Soares, formed a minority government. In June, General António Ramalho Eanes, who had been instrumental in preventing a radical leftist military coup in November 1975, won 61.5 percent of valid votes in the presidential election.

Soares' minority government resigned in December 1977, primarily because it was unable to enact an effective austerity program. A number of volatile coalition governments followed, until in 1980, in the general election scheduled by the constitution, a centre-right alliance, the Democratic Alliance (Alianca Democrática), swept into power. It was this government that succeeded in revising the character of the 1976 constitution. The Assembly of the Republic approved a series of reforms that included abolishing the Council of the Revolution and reducing the powers of the president. These constitutional reforms completed Portugal's transition to full civilian rule. Both government policy and public sentiment, as monitored by numerous elections and polls, favoured reprivatization of the largely nationalized economy, a de-emphasis on communal agriculture, and entry into the European Community (EC) as soon as possible.

The alliance faltered in 1982, propelling the country into yet another crisis. President Eanes called an early general election for April 1983, in which the Socialists, headed by Soares, scored an inconclusive victory. Because Portugal urgently needed a stable, broadly based government to tackle its severe economic problems, Soares formed a coalition government with the Social Democrats (formerly the Popular Democrats). It successfully implemented an 18-month emergency program and a four-year modernization program in its quest for admission to the EC.

The coalition, though precarious, lasted until June 13, 1985. It survived several internal crises caused predominantly by a division within the Social Democrats between a left wing favouring the coalition and a right wing disagreeing with the coalition's economic policies. In May 1985 Aníbal Cavaco Silva, leader of the right wing, became head of the party. He soon called into question the viability of the coalition, voicing doubts especially on subjects of labour and agrarian reform.

This crisis, which ended the coalition in June, had been intensified by nationwide strikes in the industrial and transport sectors, led by Communist unions, and by demonstrations from parties on both the left and the right of the political spectrum calling for an end to the coalition government. Soares resigned, and in the elections held in October 1985 the Social Democrats, campaigning on a platform advocating a free-market economy, became the largest single party, able to form a minority government with Cavaco Silva as prime minister. Portugal was admitted to the EC on Jan. 1, 1986, and on February 16, Soares became the country's first civilian president in 60 years. The parliamentary elections of 1987 marked another milestone as the Social Democrats of Cavaco Silva won the first clear majority of seats in the Assembly since the 1974 revolution. A renewal of this mandate four years later provided, at least and at last, the continuity necessary for carrying out reforms.

Election of a Social Democratic government

As the 20th anniversary of the revolution approached, there was cause in Portugal for both celebration and concern. Once the poorest and least-developed of EC members, the country now moved ahead of Greece on the strength of a 4.5 to 5 percent annual economic growth

rate. Promising improvements in health, welfare, and education have accompanied this expansion, and a new political climate has brought unprecedented civil liberties. On the other side of the ledger, Portugal remained dependent on oil imports, and food self-sufficiency was still a distant prospect. As trade barriers within the EC (now renamed the European Union; EU) fell in 1993, external competition threatened to drive many small and middle-sized companies out of the market.

During the 1990s Portugal's economy experienced significant growth, though unemployment continued to be a problem. The protection of Portugal's historic heritage also remained a serious issue. Portugal's economy was partly dependent on tourism, but its fragile environment was endangered by tourism, urban sprawl, and pollution. The country also attempted to address rural economic problems precipitated by migration to urban areas.

By the end of the 20th century, Portugal had solidified its democracy. With the military's withdrawal from politics and several revisions of the constitution, Portugal adopted what could be called a semipresidential system, which limited the president's powers by investing significant authority with the prime minister. Portugal developed a multiparty system, in which two major parties (the Socialists and Social Democrats) and several minor parties emerged. In 1995 Silva left office, replaced by Socialist António Guterres; the following year, Soares was succeeded as president by Socialist Jorge Sampaio, Lisbon's former mayor. In 1999 Portugal adopted the euro, the EU's single currency, and also returned Macau, its last overseas territory, to Chinese rule. Sampaio was reelected in 2001, but in 2002 Guterres's government was ousted by the Social Democrats, who formed a centre-right coalition government and who promised to reduce taxes and spending and privatize some public services. (D.L.W./Ed.)

For later developments in the political history of Portugal, see the BRITANNICA BOOK OF THE YEAR.

BIBLIOGRAPHY

General works. MARION KAPLAN, *The Portuguese: The Land and Its People*, new rev. ed. (1998), is a popular general work. WILLIAM CHISLETT, *Portugal—Investment and Growth* (1997), reviews major changes and Portugal's economy. Dated but still useful surveys include SARAH BRADFORD, *Portugal* (1973); and JOHN EPPSTEIN, *Portugal: The Country and Its People* (1967). WALTER C. OPELLO, JR., *Portugal: From Monarchy to Pluralist Democracy* (1991), is a comprehensive overview. ERIC SOLSTEN (ed.), *Portugal: A Country Study*, 2nd ed. (1993), contains chapters on the country's history, society, government and politics, economy, and national security. Still useful despite the lacunae, errors, and old entries is the long standard *Grande Enciclopédia Portuguesa e Brasileira* (more than 40 volumes originally, with updated additional volumes, 1935–60 and later supplements from the 1980s and '90s).

Physical and human geography. An overview of the country's features is presented in CLIFFORD EMBLETON (ed.), *Geomorphology of Europe* (1984). The development of the cultural landscape, including a summary of the physical landscape, is provided in JOSE MATTOSÉ and FRANCISCO FARIA PAULINO, *Portugal, Building up a Country* (1992); Orlando Ribeiro, *Portugal, o Mediterrâneo e o Atlântico*, 7th ed., rev. and enlarged (1998); and ORLANDO RIBEIRO, HERMANN LAUTENSACH, and SUZANNE DAVEAU, *Geografia de Portugal*, 4 vol. (1987–91). Pictorial representations include ARISTIDES DE AMORIM GIRÃO, *Atlas de Portugal*, 2nd ed. (1958); and VICTORIA ZALACAIN *et al.*, *Atlas de España y Portugal* (1982). A classic discussion useful for students of anthropology, social history, and demography on villages in northwest Portugal is CAROLINE B. BRETTELL, *Men Who Migrate, Women Who Wait: Population and History in a Portuguese Parish* (1986).

RODNEY J. MORRISON, *Portugal: Revolutionary Change in an Open Economy* (1981), assesses the impact of the 1974 revolution on the economy. ERIC N. BAKLANOFF, JORGE BRAGA DE MACEDO, and SIMON SERFATY (eds.), *Portugal Since the Revolution: Economic and Political Perspectives* (1981), analyzes the political economy of the early postrevolutionary period. DAVID CORKILL, *The Development of the Portuguese Economy, A Case of Europeanization* (1999), charts Portugal's development within mainstream Europe. THOMAS C. BRUNEAU, *Politics and Nationhood: Post-Revolutionary Portugal* (1984), is an excellent treatment of the politics of the revolutionary period. TOM GALLAGHER, *Portugal: A Twentieth-Century Interpretation* (1983); and HOWARD J. WIARDA, *Corporatism and Development: The Portuguese Experience* (1977), analyze Salazar's dictatorship. Politics and government are explored in THOMAS C. BRUNEAU and ALEX MACLEOD, *Politics in Contemporary Portugal: Parties and the Consolidation of Democracy* (1986). LAWRENCE S. GRAHAM AND HARRY M. MAKLER (eds.), *Contemporary Portugal: The Revolution and Its Antecedents* (1979), provides chapters on the prerevolutionary corporatist system. KENNETH MAXWELL (ed.), *Portugal in the 1980's: Dilemmas of Democratic Consolidation* (1986), examines shifting international involvements, economy, society, and state through the 1980s. (C.D.S./W.C.O./M.Kap.)

History. A selection of general reference works on Portugal, with an emphasis on history, includes DOUGLAS L. WHEELER, *Historical Dictionary of Portugal* (1993); JOEL SERRÃO (ed.), *Dicionário de história de Portugal*, 6 vol. (1963–71); ANTÓNIO BARRETO and MARIA FILOMENA MONICA (eds.), *Suplemento. Dicionário de história de Portugal*, 3 vol. (1999–2000); and IÉDA WIARDA (ed.), *The Handbook of Portuguese Studies* (1999).

General historical works include JOEL SERRÃO and A.H. DE OLIVEIRA MARQUES (eds.), *Nova história de Portugal* (1900–); and JOSÉ MATTOSO (ed.), *História de Portugal*, 8 vol. (1993–). Useful but dated are H.V. LIVERMORE, *Portugal: A Short History* (1973), and *A New History of Portugal*, 2nd ed. (1976); STANLEY G. PAYNE, *History of Spain and Portugal*, 2 vol. (1972); and A.H. DE OLIVEIRA MARQUES, *History of Portugal*, 2nd ed., 2 vol. (1976).

General works on Portugal's overseas empire after 1415 include C.R. BOXER, *The Portuguese Seaborne Empire, 1415–1825*, 2nd ed. (1991); A.J.R. RUSSELL-WOOD, *A World On the Move: The Portuguese in Africa, Asia, and America 1415–1808* (1992); GEORGE D. WINIUS (ed.), *Portugal, The Pathfinder* (1995); DOUGLAS L. WHEELER and RENÉ PÉLISSIER, *Angola* (1971, reprinted 1978); and JAMES DUFFY, *Portuguese Africa* (1959).

Discussions of modern Portugal include H.V. LIVERMORE (ed.), *Portugal and Brazil* (1953); SARAH BRADFORD, *Portugal* (1973); J.B. TREND, *Portugal* (1957); CARL A. HANSON, *Economy and Society in Baroque Portugal, 1668–1703* (1981); KENNETH MAXWELL, *Pombal, Paradox of the Enlightenment* (1995); MARIA FILOMENA MONICA, *Fontes Pereira de Melo* (1999); DOUGLAS L. WHEELER, *Republican Portugal: A Political History (1910–1926)* (1978); ANTÓNIO DE FIGUEIREDO, *Portugal: Fifty Years of Dictatorship* (1975); HUGH KAY, *Salazar and Modern Portugal* (1970); and JOHN SYKES, *Portugal and Africa. The People and the War* (1971). The 1974–75 revolution in Portugal and its aftermath are explored in LAWRENCE S. GRAHAM and HARRY M. MAKLER (eds.), *Contemporary Portugal: The Revolution and Its Antecedents* (1979); LAWRENCE S. GRAHAM and DOUGLAS L. WHEELER (eds.), *In Search of Modern Portugal: The Revolution and Its Consequences* (1983); KENNETH MAXWELL, *The Making of Portuguese Democracy* (1995); NORRIE MACQUEEN, *The Decolonization of Portuguese Africa: Metropolitan Revolution and the Dissolution of Empire* (1997); MÁRIO SOARES, *Portugal's Struggle for Liberty* (1975); DOUGLAS PORCH, *The Portuguese Armed Forces and The Revolution* (1977); NANCY GINA BERMEO, *The Revolution within the Revolution: Workers' Control in Rural Portugal* (1986); TOM GALLAGHER, *Portugal: A Twentieth-Century Interpretation* (1983); LESTER SOBEL, *The Portuguese Revolution* (1976); and ANTÓNIO COSTA PINTO (ed.), *Modern Portugal* (1998). (H.V.L./D.L.W.)

Portuguese Literature

The literature of Portugal is distinguished by a wealth and variety of lyric poetry, which has characterized it from the beginning; by its medieval lack of and later achievement in the national epic; by its wealth of historical writing; and by its relative slightness in drama, biography, and the essay. The early *cancioneiros* ("song-books") evidence a school of love poetry that spread, with the language, to Spain at a time when Spanish was as yet undeveloped for lyrical purposes. The *romanceiro,* or balladry, on the other hand, was much influenced by that of Spain, though not sharing the latter's predilection for the heroic. In its primitive version *Amadís de Gaula* (14th century; *Amadis of Gaul*), prototype of the romance of chivalry, probably was written in Portuguese, as was, later, *La Diana* (1559) of Jorge de Montemor (Montemayor), the masterpiece of the pastoral novel. *Os Lusíadas* (1572; *The Lusiads*), a history of the Portuguese (the name Lusíads deriving from the ancient Lusitania), by Luís de Camões, may be at once the most successful of the many Renaissance epics cast in the classical mold and the most national of great poems in any modern literature, and many works of history and travel of the 16th and 17th centuries are outstanding. Though Gil Vicente, in the early 16th century, was a dramatist of great gifts, no other appeared until João Baptista de Almeida Garrett in the 19th, and Portugal never developed a national drama.

This literature, which until the 19th century lay largely unstudied and unknown, has from the beginning been exposed to foreign influences. The earliest was Provençal, and Provençal taste ruled for more than a century. Then came Castilian, with a court poetry that provided models until the Renaissance saw the triumph of that form in Italy and in the classics of antiquity. In the 17th century with political domination Spain again imposed its literary standards, followed in the 18th century by France. The Romantic movement reached Portugal from both France and England, two countries whose influence, joined in a lesser degree by that of Germany, persisted long after. The closeness of contacts with Spain, reinforced by dynastic marriages that often brought to the court at Lisbon a predominantly Spanish atmosphere, explains why for two centuries and more after 1450 nearly every Portuguese writer of note was bilingual and wrote also in Spanish, so that some, like Montemor and Francisco Manuel de Melo, are numbered among the classics too of Spanish letters. Portuguese literature retains, nonetheless, a distinct individuality that contrasts strikingly with that of Spain. The medieval lyric, the plays of Gil Vicente, the bucolic verse and prose of the 16th century, and, above all, *Os Lusíadas,* are expressions of a clearly defined national temperament.

This article is divided into the following sections:

Writings of the early period 1063
Poetry
Prose
Developments of the 15th and 16th centuries 1063
Literature of the 15th century
Gil Vicente and early drama
The Renaissance in Portugal
The Italianate school of poetry and drama
The 16th-century history
The novel and other prose
The 17th century 1065
The 18th century 1065
The 19th century 1065
Romanticism and after
Drama, the novel, and history
The 20th century 1066
Bibliography 1066

WRITINGS OF THE EARLY PERIOD

Poetry. Though no literary documents belonging to the first century of Portugal's history as a nation have survived, there is evidence of the existence of an indigenous popular poetry. A few compositions from before 1200 survive; one, attributed to Sancho I, is the earliest extant parallelistic song, a brief, repetitive lyrical poem marked by a wistful sadness that is never wholly absent from Portuguese literature. Of the many later poems that survive, most belong to the major categories of *cantigas de amor* ("songs of love"; sung by a man to the woman he loves in vain), *cantigas de amigo* ("songs of the lover"; sung by a woman to express her yearning for her lover), and *cantigas de escárnio e maldizer* (satirical songs). This body of lyrics, represented in three great *cancioneiros,* shows the vitality of a school of poetry in Galician-Portuguese (the dialect of northern Portugal) that, while essentially inspired by the sophisticated French and Provençal songs of the troubadours, is also anchored in the popular tradition.

This poetry reached the peak of its creativity in the first half of the 13th century, coinciding with the reign of Afonso III (1248–79). His son, Dinis, had a deep interest in literature and was considered to be the best poet of his age in the Iberian Peninsula. Dinis founded his country's first university at Lisbon in 1290 (it later was moved to Coimbra) and encouraged translation into Portuguese of outstanding works from Spanish, Latin, and Arabic. To his court came troubadours from Leon, Castile, and Aragon to enjoy the last of a cult dying elsewhere, and about 2,000 poems by its 200 poets were preserved in the three great repositories of verse, *Cancioneiros da Ajuda, Cancioneiros da Vaticana,* and *Cancioneiros Colocci-Brancuti* (or *da Biblioteca Nacional de Lisboa*).

Dinis' encouragement of literature

In contrast with the restricted horizons of courtly verse, themes of adventure, war, and chivalry mingled with love, religion, and the sea in a collection of ballad poetry known as the *romanceiro.* Few of these ballads can be dated earlier than the 15th century; they belonged to an anonymous poetry kept alive by oral transmission, with a late artificial flowering from known poets in the 16th and 17th centuries.

Prose. Prose literature in Portugal took much longer than verse to perfect. Religious writings, brief annals of the early kings, and books of descent formed the earliest texts. The *Livro de Linhagens* ("Book of Genealogy") of Pedro Afonso, count of Barcelos (the natural son of Dinis), constituted a landmark by going beyond genealogy to history and legend. The work contains short epic narratives, romances, and tales of adventure and fantasy. Pedro was also responsible for the compilation in 1344 of the *Crónica Geral de Espanha* ("General Chronicle of Spain"), interesting, within the peninsular tradition of the general chronicle genre, for its original version of well-known episodes.

The early popularity of subject matter based on Celtic tradition is attested in the five songs based on Breton lays with which the *Cancioneiro Colocci-Brancuti* opens. The ideals of chivalry and the spirit of sentimental adventure associated with the knights of the Round Table made strong appeal to the Portuguese imagination: a *História dos Cavaleiros da Távola Redonda* ("History of the Knights of the Round Table") and the *Demanda do Santo Graal* ("Search for the Holy Grail"), adapted from the French, are the chief relics of considerable activity in this field.

DEVELOPMENTS OF THE 15TH AND 16TH CENTURIES

Literature of the 15th century. Under King John (João) I, founder of the new dynasty of Avis, the Portuguese court became once again a literary centre. The King himself wrote a treatise on hunting. His son Duarte (Edward) collected a rich library of the ancients and of medieval poems

The influence of the Avis dynasty

and histories and composed a moral treatise, *Leal Conselheiro* (1437/38; "Loyal Counselor"), which revealed a conscious stylist. But the historical chronicle distinguished the age, with credit to King Edward, who in 1434 created the office of *cronista mor do reino,* or "chief chronicler of the realm," and appointed Fernão Lopes, author of chronicles of the first 10 kings of Portugal. The *Crónica dos Sete Primeiros Reis do Portugal,* a chronicle of the first seven kings, discovered in 1947 and published in 1952–53, is almost certainly the first part of that work. Until its discovery only the chronicles of Pedro I, Ferdinand I, and John I were known. Vividness of style combined with serious documentation to produce in Lopes the finest writer of medieval Portuguese prose.

His successor in office, Gomes Eanes de Zurara, continued the chronicle on a lower level of artistry. His chief works are the *Crónica da Tomada de Ceuta* (*Conquests and Discoveries of Henry the Navigator*) and the *Crónica do Descobrimento e Conquista da Guiné* (*The Chronicle of the Discovery and Conquest of Guinea*).

Poetry was cultivated in the mid-15th century after a long eclipse, but much had changed. The dominant influence came now from Spain, and Portuguese poets initiated the long chapter of allegiance to Spain. Apart from the ballads, popular poetry had disappeared along with that of the troubadours. The constable Dom Pedro de Portugal initiated the fashion of writing in Castilian. As one of the first to adopt the new Spanish trend toward allegory and the cult of classical antiquity derived from Italy, his influence on his compatriots was doubly important. His own poems were inspired by deep feeling and much reflection on life, and he was one of almost 200 poets represented in an anthology of poetry, the *Cancioneiro Geral* (1516; "General Songbook"), of the chronicler Garcia de Resende, covering the preceding three-quarters of a century. The main subjects of these 1,000-odd poems, in Portuguese and Castilian, were love, satire, and epigram. Resende was a better poet than most of his contributors.

The *Cancioneiro Geral*

Gil Vicente and early drama. The emergence of the modern play may be traced in the works of the court dramatist Gil Vicente. Eleven of his 44 plays were written wholly, and another 17 partly, in Spanish. The *Barcas* (1517–19), a group of *autos,* or religious plays, revealed his dramatic power and a fondness for comic relief; in this lay his strength, and in construction lay his weakness. The phenomenon of a potential national theatre, however, died with its founder, and his real influence was felt in Spain. The Inquisition, introduced into Portugal in 1536, early declared war on the popular theatre on the charge of grossness. Vicente's own plays, which figured on the Spanish Index of 1559, were reduced in number to 35 and were sadly mutilated in the second edition of 1586.

The Renaissance in Portugal. The Renaissance reached Portugal both indirectly through Spain and directly from Italy, with which country there had been close cultural relations through the 15th century. In the following century many famous humanists took up residence in Portugal. In 1532 the historian and humanist João de Barros published the *Ropica Pnefma* ("Spiritual Merchandise"), the most important philosophical dialogue of the time in Portugal.

In 1547 John III reformed the University of Coimbra, and distinguished Portuguese teachers returned from abroad to assist the King in his task. At home Portugal produced scholars of note, including André de Resende, author of *De antiquitatibus Lusitaniae* (1593; "Of the Antiquities of Portugal"), and Francisco de Holanda, painter, architect, and author of the *Diálogos da Pintura Antiga* ("Dialogues on Ancient Painting").

The Italianate school of poetry and drama. The return in 1526 of the poet Francisco de Sá de Miranda after a six years' stay in Italy initiated a literary reform of far-reaching effect. Like his contemporary Garcilaso de la Vega in Spain, he introduced the new poetic forms of sonnet, canzone, ode, and epistle, and he gave fresh vigour to the national verse forms, mainly through his *Satires.* His chief disciple, Antônio Ferreira, a convinced classicist, wrote sonnets superior in form and style. Other poets continued the erudite school, which triumphed with Luís de Camões, author of *Os Lusíadas.* In Camões, a profound classical education combined with perfect mastery of his instrument and a lifetime of varied experience to produce in sonnets, eclogues, odes, elegies, and *canções* the greatest poetry in the language.

Luís de Camões

In the drama Sá de Miranda and his followers substituted prose for verse. Taking the Roman dramatist Terence as their model, they produced not Portuguese characters but Romano-Italian types. This revived classical comedy was to be short-lived. Sá de Miranda, avowedly to combat the school of Vicente, wrote *Os Estrangeiros* (*c.* 1528; "The Strangers"), the first prose comedy, and *Os Vilhalpandos* (*c.* 1528), both actions being set in Italy. Ferreira, a greater dramatist, likewise attempted both kinds: *O Cioso,* Italian even to the names of the personages, came nearer to being a comedy of character, but his fame rests chiefly on *Castro* (*c.* 1557; *Ignez de Castro*), which treated one of the most moving tragic themes in the nation's history—the execution of Inês de Castro—by reference to the ancient Greek dramatists Sophocles and Euripides. From the comic playwright Jorge Ferreira de Vasconcelos came a "new invention" of another kind with *Comédia Eufrosina* (1555), written under the influence of the Spanish novel *La Celestina.* This and his other plays, *Comédia Ulissipo* (1618) and *Comédia Aulegrafia* (1619; "Life at Court"), resembled novels in dialogue and contained a treasury of popular lore and wise and witty sayings introduced with a moral purpose.

The 16th-century history. Discovery and conquest in Africa, Asia, America, and on the ocean inspired historians as well as poets. Their records gained in vividness what they might have lost in scientific detachment. In the three "Decades" of his *Décadas da Ásia* (1552–63), João de Barros told in vigorous language the deeds of compatriots overseas down to 1526. His first "Decade" undoubtedly influenced Camões, and together, one by his prose and the other by his verse, these two fixed the written language. This work, continued by the more critical and inclusive Diogo do Couto, can rank as the noblest historical monument of the century. In his *Soldado Prático* (written before 1578, published in 1790; "Experienced Soldier"), Diogo do Couto added acute observations on the causes of Portuguese decadence in the East. Ten years of investigation in India underlay the *História do Descobrimento e Conquista da Índia Pelos Portugueses* (1551–61; "History of the Discovery and Conquest of India by the Portuguese") of the chronicler and notary Fernão Lopes de Castanheda, a work that ranks close to those of Barros and do Couto.

Literature of the Portuguese Empire

From this spate of writing on expansion overseas, attention returned, by way of chronicles of the monarchs who presided over the work, to the history of Portugal itself. Damião de Góis, diplomat, traveler, humanist, and intimate friend of Erasmus, possessed an encyclopaedic mind and was one of the most critical spirits of the age: his *Chronica do Felicissimo Rey Dom Emanuel* (1566–67; "Chronicle of the Most Happy King, Dom Emanuel") was most valuable where the author's own feelings of experience came into play.

Works of travel abounded, and their authors were often the first Europeans to visit the countries in question. Among the more noteworthy was a much-translated classic, *História da Vida do Padre Francisco Xavier* (1600; "History of the Life of Father Francis Xavier"), by Padre João de Lucena. Important both as history and as human documents were the *cartas,* or "letters," written home by Jesuits in China and Japan. An anonymous *Descobrimento da Florida* ("Discovery of Florida") and a *Tratado Descritivo do Brasil em 1587* ("Descriptive Treatise of Brazil in 1587") of Gabriel Soares de Sousa were reminders that Portugal was also present and active in the New World to the west. A work notable for its curious interest is the *Peregrinacam* (1614; *The Voyages and Adventures of Ferdinand Mendez Pinto, of Portugal*), which Fernão Mendes Pinto, prince of adventurers throughout the East, composed in his old age, and another, known for tragic pathos, is the *História Trágico-Marítima* ("Tragical Maritime History"), a collection of contemporary narratives, told by survivors or based on their accounts, of the more notable disasters that befell Portuguese ships between 1552 and 1604.

The novel and other prose. The poet Bernardim Ribeiro, whose five eclogues introduced pastoral poetry to Portugal, was equally an innovator in the pastoral novel with his *Hystoria de Menina e Moça* (1554; "The Story of Menina and Moça"). This tale of rustic love and melancholy with chivalric elements adopted themes and emotions previously found only in poetry. From it a musician and poet, Jorge de Montemor, drew some part of his inspiration for *La Diana*, which, written in Spanish, started a fashion subscribed to by Cervantes and Lope de Vega among many others and represented one of the outstanding contributions of Portugal to the neighbouring literature. The first romance of chivalry in Portuguese was João de Barros' *Crónica do Imperador Clarimundo* (1520; "Chronicle of the Emperor Clarimundo"), concerning the adventures of a fictitious progenitor of the kings of Portugal.

Among moralists and theologians, three ranked as masters of prose style: Frei Heitor Pinto, for his *Imagem da Vida Cristã* (part I, 1563, part II, 1572; "Image of the Christian Life"); Bishop Amador Arrais for 10 *Diálogos* (1589; "Dialogues") on religious and other topics; and Frei Tomé de Jesus for a mystic and devotional treatise, *Trabalhos de Jesus* (1602–09; "Works of Jesus"). The work of scientists included that of a cosmographer and mathematician, Pedro Nunes, and of a botanist, Garcia de Orta, whose *Colóquios dos Simples e Drogas* (1563; *Colloquies on the Simples and Drugs of India*) was the first Portuguese book to be printed in the East (at Goa in India).

THE 17TH CENTURY

From a literary and political point of view, the 17th century found Portugal in a state of decadence. Before the loss of independence to Spain in 1580, Spanish influence had introduced the Inquisition and, with it, censorship of books and the preparation of an Index of Forbidden Books. Between 1552 and 1555 the Jesuits gained control of higher education. The cult of classical Latin was already present, in the work of Camões and others, before the example of the Spanish poet Luis de Góngora y Argote was felt; but with the exhaustion of the national spirit that underlay political eclipse the influence of Góngora penetrated deeply. Its extent may be seen in the five volumes of the *Fénix Renascida* (1716–28; "Phoenix Reborn"), which anthologizes the poetry of the preceding century and reveals the futilities to which good talents could devote themselves. The trend survived the throwing-off of the Spanish yoke in 1640; Portuguese editions of Góngora continued to appear.

The influence of the Spanish poet Góngora

The foremost literary figure of the age was the encyclopaedic Francisco Manuel de Melo, a classic of Spanish and—with his series of historical episodes, *Epanáforas de Vária História Portuguesa* (1660; "Anaphoras of Diverse Portuguese History"), and dialogues on literary and social topics, *Apólogos Dialogais* (1721)—of Portuguese literature who strove, more successfully in prose than in verse, to free himself from subservience to Spanish form and style. Most lyricists of the period remained steeped in Gongorism. Epic poets continued active, but few of their productions were more than rhymed chronicles.

Frei Luís de Sousa, a monastic chronicler, won fame as a stylist with his *Vida do Arcebispo D. Frei Bartolomeu dos Mártires* (1619; "Life of Archbishop D. Frei Bartolomeu dos Mártires") and the *História de São Domingos* (three parts, 1623, 1662, 1678; "History of São Domingos"). A Jesuit, António Vieira, missionary and diplomatist, was highly regarded for his *Cartas* (1925–28; "Letters") and *Sermões* (1679–1748; "Sermons").

The popular theatre lived on obscurely with mostly anonymous plays that were never printed. Those that survived were mainly religious and showed the common Gongoristic abuse of metaphor and simile. All through the century most dramatists who aspired to be heard wrote in Spanish. The court after 1640 preferred Italian opera, French plays, and Spanish operettas—to the detriment of native drama and of acting.

THE 18TH CENTURY

The 18th century, in Portugal as in Spain, was predominantly prosaic, even in poetry. Yet signs gradually appeared of a literary revolution that developed eventually into the Romantic movement. Luís António Verney poured scorn on prevailing methods of education in *Verdadeiro Método de Estudar* (1746; "True Method of Studying"). Men of liberal ideas traveled to France and England, and to their example were largely due the reforms that invaded every branch of letters. Among the most influential were Alexandre de Gusmão, Francisco Xavier de Oliveira, António Ribeiro Sanches, José Correia da Serra, Avelar Brotero, and Francisco Manuel do Nascimento. New literary societies called *arcádias* cooperated in the task of reform. In 1720 King John V established the Academia Real da História Portuguesa, which counted among its members such men as Manuel Caetano de Sousa, author of a colossal *História Genealógica de Casa Real Portuguesa* (1735–49; "Genealogical History of the Portuguese Royal House"). The Academia Real das Ciências (Royal Academy of Sciences), founded in 1779, initiated research into the study of Portuguese literary history. In its ranks were found nearly all the scholars of note at the end of the century, such as the ecclesiastical historian Frei Manuel do Cenáculo; a scientist, António Ribeiro dos Santos; João Pedro Ribeiro, perhaps his country's first modern historian; and critics Francisco Alexandre Lobo and Frei Fortunato de São Boaventura.

Establishment of the Portuguese academies

In 1756 António Dinis da Cruz e Silva established the Arcádia Lusitana (or Ulissiponese), its first aim being the uprooting of Spanish influence. The bucolic verse of Dómingos dos Reis Quita signified a return to the native tradition of two centuries earlier. Sincerity and suffering spoke in the justly more famous *Marília de Dirceu* (1792), love lyrics in a pastoral setting, by Tomás António Gonzaga. In 1790, a Brazilian Nova Arcádia came into being, its two most distinguished members being the rival poets Manuel Maria Barbosa du Bocage and José Agostinho de Macedo.

Outside the Arcádias stood the "Dissidents," among whom were at least two writers of distinction. These were the satirist Nicolau Tolentino de Almeida, who painted the customs and follies of his day with devastating accuracy, and Francisco Manuel do Nascimento (pseudonym Filinto Elísio), who addressed himself perseveringly to purifying the language and to restoring the cult of the 16th-century poets.

Early in the century popular authors attempted a revival of the drama in Lisbon. The *Óperas Portuguesas* (1733–41) of Antônio José da Silva owe their name to the interspersing of the prose dialogue with arias, minuets, and *modinhas* (popular light songs).

The attempted revival of drama in Lisbon

THE 19TH CENTURY

Romanticism and after. The 19th century witnessed a general revival of Portuguese letters. The chief exponents of Romanticism were, in poetry and drama, João Baptista de Almeida Garrett and, in prose, Alexandre Herculano; both lived for some years in exile, the price of their political liberalism. Almeida Garrett read literature in English and French and introduced Portugal to nationalistic Romanticism through two epics, *Camões* (1825) and *Dona Branca* (1826). António Feliciano de Castilho, half-Romantic, half-Classicist, exercised much influence over a younger generation of poets: João de Lemos, Soares de Passos, Tomás Ribeiro (who won fame with the ardently patriotic *Dom Jaime*, 1862). In 1865 Antero de Quental, a student of German philosophy and poetry, and Teófilo Braga, disciple of Auguste Comte, led a revolt against the primacy of Castilho, much to the benefit of literature, and of poetry in particular. The *Campo de Flores* (1893; "Field of Flowers") of João de Deus contained some of the finest short poems in the language, marked by a spontaneous simplicity. Abílio Manuel Guerra Junqueiro, heir to Victor Hugo, was a would-be social revolutionary excessively prone to grandiloquence. In *Os Simples* (1892) he turned to the portrayal of peasant life, and this work constituted his finest poetry. Akin to him was António Duarte Gomes Leal, author of *Claridades do Sul* (1875; "Clarities of the South") and *O Anti-Cristo* (1884; "The Anti-Christ"), who could likewise achieve quiet sincerity when dealing with humble themes.

António Cândido Gonçalves Crespo stood out as the first of his country's Parnassians. By contrast Cesário Verde, considered to be the greatest poet of the century, addressed himself to the poetic essence of common realities. The *Só* (1892; "Alone") of António Nobre was intensely Portuguese in themes, mood, and rhythms; he and Teixeira de Pascoais developed a cult of *saudosismo* ("yearning," "nostalgia") that inspired a whole school of poets. French Symbolism found an enthusiastic adept in Eugénio de Castro.

Portuguese drama

Drama, the novel, and history. Almeida Garrett, seeking to reinvigorate drama, found he had to create alike theatre, plays, actors, and audience. In *Um Auto de Gil Vicente* (1838; "An *Auto* of Gil Vicente"), *O Alfageme de Santarém* (1841; "The Swordsmith of Santarém"), and especially in *Frei Luís de Sousa* (1843; *Brother Luiz de Sousa*), he proved himself to be, after Vicente, his country's most notable dramatist. João da Câmara was the outstanding dramatist of his day, and his works include *Afonso VI* (1890), *Rosa Enjeitada* (1901; "Rose Abandoned"), and *Os Velhos* (1893; "The Old Ones"). Herculano, returning from exile with an enthusiasm for Sir Walter Scott, launched the historical romance with *O Monasticon: Eurico, o presbitero* (1844; "The Monastic: Eurico, the Presbyter") and *Lendas e Narrativas* (1851; "Legends and Narratives"). Many followed suit, including Oliveira Marreca, Arnaldo da Gama, and Pinheiro Chagas, popular successes being *A Mocidade de D. João V* (1852; "The Youth of D. João V"), by Luís António Rebelo da Silva, and João de Andrade Corvo's *Um Ano na Côrte* (1850–51; "A Year in the Court"). This was the great age of the novel: Camilo Castelo Branco, Joaquim Gomes Coelho (better known as Júlio Dinis), and especially José Maria Eça de Queirós were names that would stand high in any country. The first was a master of the language and of dramatic, or melodramatic, plot; Júlio Dinis depicted country life, as in *As Pupilas do Senhor Reitor* (1867; "The Pupils of the Dean"); while Eça de Queirós introduced Realism with a powerful novel, *O Crime do Padre Amaro* (1876; *The Sin of Father Amaro*).

With his magnum opus, the *História de Portugal* (1846–53; "History of Portugal"), and the *História da Origem e Estabelecimento da Inquisição em Portugal* (1854–59; *History of the Origin and Establishment of the Inquisition in Portugal*), Herculano established himself as a leader of modern Spanish and Portuguese historians. Historiography flourished with the Visconde de Santarém, historian of the Cortes; José Simão da Luz Soriano (of constitutionalism); Rebelo da Silva (of the period of Spanish rule under the Philips); and José Maria Latino Coelho (of the dictatorship of Pombal). Henrique da Gama Barros and António da Costa Lôbo followed in the footsteps of Herculano. The works of Joaquim Pedro de Oliveira Martins gave proof of psychological imagination, a notable capacity for general ideas, and a gift of picturesque narration. He left in his numerous writings a vast portrait gallery of great figures of his country, particularly in the *Portugal Contemporâneo* (1881; "History of Contemporary Portugal").

THE 20TH CENTURY

The passage from monarchy to republic in Portugal in 1910 saw a revisionary urge in literature associated chiefly with Oporto and the Renascença Portuguesa ("Portuguese Renaissance"). Leonardo Coimbra was its philosopher, and António Sérgio its critic and historian. Its poets—Mário Beirão, Augusto Casimiro, and João de Barros—adopted the *saudosismo* of Teixeira de Pascoais as the key to the nation's recovery of greatness, though the inadequacy of this nostalgia was soon realized. The "Integralist" school reacted from 1913 onward in favour of the Roman Catholic monarchist tradition, led by António Sardinha, a historian and poet. Fernando Pessoa, author of *Mensagem* (1934; "Message"), was posthumously regarded as the most brilliant poet of his generation.

The Integralist school

Among novelists of the first half of the century, Aquilino Ribeiro was a prolific writer whose themes often were centred on his native Beira; his delight in life was combined with awareness of decay and death. Miguel Torga, a poet, storyteller, and author of autobiography, showed a radical individualism that took its strength from his peasant roots. The psychological novel, which had attained a sophisticated form with José Régio, also an outstanding dramatist and a religious poet, took new directions with Carlos de Oliveira. *Casa na Duna* (1943; "A House on a Sandhill"), his first novel, combines an acute perception of human motivation with social awareness, leading up to *Finisterra* (1978), his last book and a masterpiece. Vergílio Ferreira added a metaphysical dimension to the novel of social concern in *Alegria Breve* (1965; "Brief Joy") and explored the evanescent moods of the past and the idea of death in *Para Sempre* (1983; "Forever").

After the overthrow of the Salazar regime on April 25, 1974, poetry and fiction found a new freedom of expression. Jorge de Sena, a poet and novelist of outstanding quality, published *Sinais de Fogo* (1978; "Signs of Fire"), an impressive novel about the effects in Portugal of the Spanish Civil War. J. Cardoso Pires wrote his best novel, *Balada da Praia dos Cães* (1983; "Ballad of the Beach of the Dogs") based on the account of a political assassination. Women novelists have also brought a new voice to fiction. Agustina Bessa Luís, a most prolific writer, first came to notice in 1955. She extended the psychological insight evident in her drawing of fictional characters to enhance her portraits of historical figures, as in her work *Fanny Owen* (1979). Maria Velho da Costa, one of the authors of *Novas Cartas Portuguesas* (1971; *The Three Marias: New Portuguese Letters*)—a book that became a cause célèbre because its authors were put on trial by the state—explored with great subtlety the condition of women in a repressive society in *Lúcialima* (1983; "Lemon Verbena"). José Saramago came to the fore in the 1980s. His novels combine an acute observation of reality with flights of poetic fancy. In *Memorial do Convento* (1982; "Memorial of the Convent") the story of the building of a magnificent convent is also the allegory of human suffering in history, told in the form of an epic tale.

Women novelists

Notable work also came out in history and literary criticism, and through this Portugal's history and literature have become more adequately known.

BIBLIOGRAPHY

Histories: THEOPHILO BRAGA, *História de Literatura Portugueza*, 11 vol. (1896–1907); J. MENDES DOS REMÉDIOS, *História da Literatura Portuguesa*, 6th ed. (1930), with anthology; FIDELINO DE FIGUEIREDO, *História Literária de Portugal* (*Seculos XII–XX*), 3rd ed. (1966); ALBINO FORJAZ DE SAMPAIO (ed.), *História da Literatura Portuguesa Ilustrada*, 4 vol. (1928–42); GEORGES LE GENTIL, *La Littérature portugaise*, 2nd rev. ed. (1951); ANTONIO JOSÉ SARAIVA and OSCAR LOPES, *História da Literatura Portuguesa*, 12th rev. ed. (1982); JACINTO DO PRADO COELHO (ed.), *Dicionário das Literaturas Portuguesa, Galega e Brasileira* (1956–60); AUBREY F.G. BELL, *Portuguese Literature* (1922, reprinted 1970), a standard work, with extensive bibliography.

Critical studies: FIDELINO DE FIGUEIREDO, *Characteristics of Portuguese Literature* (1916; originally published in Portuguese, 1915), *Estudos de Litteratura*, 5 vol. (1917–51), and *A Epica Portuguesa no Seculo XVI*, rev. ed. (1950); AUBREY F.G. BELL, *Studies in Portuguese Literature* (1914, reprinted 1975); EDWARD GLASER, *Portuguese Studies* (1976); M. RODRIGUES LAPA, *Das Origens da Poesia Lírica em Portugal na Idade-Média* (1929), and *Lições de Literatura Portuguesa: Época Medieval*, 9th rev. ed. (1977); PIERRE LE GENTIL, *La Poésie lyrique espagnole et portugaise à la fin du moyen âge*, 2 vol. (1949–52, reprinted in 1 vol., 1981); FREDE JENSEN, *The Earliest Portuguese Lyrics* (1978); JOLE RUGGIERI, *Il canzoniere di Resende* (1931); MARCEL BATAILLON, *Études sur le Portugal au temps de l'humanisme*, 2nd ed. (1974); HERNANI CIDADE, *Tendências do Lirismo Contemporâneo do "Oaristos" às "Encruzilhadas de Deus,"* 2nd ed. (1939), with an anthology of modern poetry; *Lições de Cultura e Literatura Portuguesas*, 2 vol., 6th rev. ed. (1975); *O Conceito de Poesia Como Expressão da Cultura*, 2nd rev. ed. (1957), and *A Literatura Portuguesa e a Expansão Ultramarina*, 2 vol., 2nd rev. ed. (1963–64); H.V. LIVERMORE and W.J. ENTWISTLE (eds.), *Portugal and Brazil: An Introduction* (1953, reprinted 1963), which contains a literary bibliography of studies and of translations from Portuguese to English. Studies of individual authors include EDGAR PRESTAGE, *D. Francisco Manuel de Mello* (1922); AUBREY F.G. BELL, *Diogo do Couto* (1924), *Fernam Lopez* (1921), *Gaspar Corrêa* (1924), *Gil Vicente* (1921), and *Luis de Camões* (1923).

(W.C.A./N.J.L./L.de S.R.)

Postal Systems

In the modern world, the term "postal system" may be defined as the institution—almost invariably under the control of a government agency—that makes it possible for any person to send a letter, packet, or parcel to any addressee, in the same country or abroad, in the expectation that it will be conveyed according to certain established standards of regularity, speed, and security. The service is paid for in advance by the sender according to a relatively simple scale of fees based on weight and, in some countries, on speed of service required. Prepayment is ordinarily made by means of postage stamps, franking machine impression, or printed indication of postage paid; payment is not usually required of the addressee.

It may seem tedious to catalog at such length characteristics that are familiar to everyone. Their very familiarity, however, and the consequent unthinking acceptance of them make it important to emphasize that until recently postal systems lacked many of these features. Although the basic need for a system to exchange written communications has been felt by all human societies and has been met in many ways, the evolution of varied postal systems adopted by different societies through the centuries into the basically similar pattern of today's state monopoly service has been a long and difficult process.

Today the governments of many countries use their postal systems to provide a range of services that often have no direct connection with the traditional function of exchanging letters. To provide for the collection, transport, and delivery of letters throughout a country, it has been necessary to establish a network of post offices extending into the remotest areas. Such a network of offices, staffed by agents of the state, provides an efficient banking service in areas in which it would be uneconomic for a commercial or state bank to establish a branch office. Many governments also pay various social security benefits—such as pensions and family allowances—through vouchers that can be cashed at post offices. In some countries post offices also collect certain taxes, normally through the sale of licenses and revenue or tax stamps. In fact, a wide range of ancillary services is provided through the postal system. In some African and Asian countries, for instance, the postal system helps distribute antimalarial drugs. While the variety of such services clearly reflects the ever-increasing use made of postal systems throughout the modern world, they lie outside the scope of this article, which concentrates on the traditional postal or mail services.

For coverage of related topics in the *Macropædia* and *Micropædia*, see the *Propædia*, section 735, and the *Index*.

This article is divided into the following sections:

Postal operations and management 1067
History 1068
Message-relay systems of the ancient world
Growth of business correspondence in the Middle Ages
Growth of the post as a government monopoly
Rowland Hill's reforms
International postal reform: the Universal Postal Union
Development of airmail
Advanced communications technology
National postal systems 1070
United States
Great Britain
France
Germany
Italy
Former Soviet Union
China
India
Pakistan
Japan
Postal services in the developing countries 1074
The international system 1074
Postal technology 1075
Technological progress in postal transport
Automation of mail handling
Bibliography 1076

POSTAL OPERATIONS AND MANAGEMENT

The raw material of the postal services, always a single object that demands individual treatment, is something sent by one person (or entity) to another who may be anywhere in the world. Letters and parcels in all shapes and sizes are subject only to the limits of weight and dimension prescribed by postal legislation. Yet, if postal services are to be efficient and economical, these items must be mass-processed, as far as is possible. Thus, the basic function of postal organization is to convert the individual item as rapidly as possible into something that can be handled on a bulk basis, ensuring, however, that it finally regains its individual status.

The sorting process

The collection and sorting of individual items by the most economic method, concentrating together all items that are going to the same place or in the same direction, involves the use of local transport, usually operated by the postal services themselves, and sorting offices. The size of the sorting office depends on local requirements, but some are, in fact, large centres that handle several million items a day and employ thousands of personnel. Certain ancillary operations are involved, such as the canceling (by machine in all except the smallest offices) of stamps that have served to prepay the postal charge.

The next stage is to transport the grouped items to different destinations that may be intermediate sorting offices, perhaps a railway sorting car. Items are combined by further sorting with mail from other sources to permit grouping the mails for final destinations.

The third stage is the arrival of the mail at the sorting office of the final destination, where it is sorted systematically. The items finally recover their identity and are grouped for delivery to the individual address. In most countries, delivery is on a house-to-house basis, although boxes at a local post office are sometimes used.

All stages have to be planned and dovetailed to meet an overall standard of performance. The transport of mail between sorting offices is normally by services not necessarily directly operated by the post office—*i.e.*, by truck, bus, train, marine shipping, or air services. In some countries, however, the administration operates its own air service, usually by night, to supplement the public services. Post-office counter services (*i.e.*, the public offices that existed originally as depots where the public could buy stamps, make inquiries, and post and collect correspondence) have in many countries gone beyond the role of accessory to the postal service proper and pose their own problems of organization.

Postal services require extensive manpower. Many countries are striving to mechanize or even automate sorting, transport, and counter processes. Postal organization depends largely on geography; large agglomerations of population present the greatest problems. Furthermore, geography usually determines the routing of mails and the intermediate steps between the posting and destination offices. Traffic problems have forced postal managers, through the years, to develop their own approach to the operational aspects of postal organization. They developed

their own work-study methods and operational research techniques long before these terms were invented. Apart from postal operations as such, postal management is concerned with the efficient administration and deployment of large bodies of manpower, the organization of large transport fleets, many aspects of property management, and financial and economic problems, particularly in an increasing number of countries where the postal services are expected to depend on their postal revenue in order to meet all of their costs (including those related to capital expenditure). Computer technology is increasingly exploited as a management aid.

Problems of postal management

In this connection, many of the developed and developing countries have come to realize that postal services often operate within a commercial market where competition can be fierce and efficiency is the watchword. With the adoption of marketing and sales techniques, new services emphasizing speed, convenience, and reliability have been introduced. One such service is express mail, known under different service names according to the country (Express Mail in the United States, Datapost in Great Britain and Germany). At additional cost, this service, in which about half the UPU membership participates, provides expedited conveyance and individualized priority handling of correspondence and goods.

Express mail

HISTORY

Message-relay systems of the ancient world. Since good communications were clearly essential for governing the extensive empires of the ancient world, it is not surprising that among the earliest historical references to postal systems were those concerning Egypt about 2000 BC and China under the Chou dynasty 1,000 years later. It was probably in China that a posthouse relay system was first developed and was brought to a high state of development under the Mongol emperors. The great Persian Empire of Cyrus in the 6th century BC also employed relays of mounted messengers, served by posthouses. The system was favourably described by the Greek historians Herodotus and Xenophon. The admiration of the Greeks was natural since their political divisions inhibited the growth of a coherent postal system, although each city-state possessed its corps of messengers.

The development of Rome from a small city-state into a vast empire embracing most of the known world brought with it the necessity for reliable and speedy communications with the governors of distant provinces. This need was met by the *cursus publicus*, the most highly developed postal system of the ancient world. The relay stages of the *cursus publicus*, established at convenient intervals along the great roads of the empire, formed an integral part of its complex military and administrative system. The speed with which messengers were able to travel during the peak of the administration was not to be rivaled in Europe until the 19th century: it has been claimed that more than 170 miles (270 kilometres) could be covered in a day and a night. The maintenance of the *cursus publicus* required a high degree of organization; an inspectorial system existed to control its operation and prevent abuse for private ends.

The *cursus publicus*

The fall of the Roman Empire in the west during the 5th century did not completely destroy the *cursus publicus*. Its advantages were evident to the new barbarian rulers; some, such as Theodoric, king of the Ostrogoths, who ruled Italy from AD 493 to 526, are known to have maintained the essentials of the Roman postal system within their own domains. Even in the early 9th century, under the Carolingian Empire, the vestiges of the *cursus publicus* appear to have persisted, and posthouses were maintained. While the service did not follow a regularly organized pattern, it was at least reasonably frequent. The continued decay of the Roman roads, the increasing unwillingness of communities bordering the roads to support the expenses of the system, and the progressive political fragmentation of Europe, however, caused all traces of the Roman postal system to disappear.

The *cursus publicus* fared better in the Byzantine Empire because its provinces were eventually absorbed into the Islāmic Empire. The substitution of one centralized imperial regime for another meant that the *cursus publicus* could be incorporated into a similar Arabian postal system based in Baghdad.

The pre-Columbian civilizations of America, responding to the same needs as the imperial states of Asia and Europe, also evolved relay systems, limited to foot messengers. In the Inca Empire, posthouses were maintained at frequent intervals along the remarkable road network, and a like system probably served the Mayan civilization for more than 1,000 years.

Growth of business correspondence in the Middle Ages. The end of the reign of the last Carolingian king in 987 marked the beginning of several centuries of confusion in Europe, in which it is difficult to trace any postal system worthy of the title. Since the kings of the period were constantly struggling to assert their authority over their unruly feudal vassals, the strong central authority that sustained most postal systems was lacking. The uncertain political situation did not favour the creation of a regulated postal service, though it necessitated frequent contact between the kings and vassals and among the great princes. They, along with other powerful institutions—the municipalities, the religious orders, and the universities (notably in Paris)—started to maintain corps of messengers to serve their particular needs.

One of the more significant trends of the later Middle Ages was the development of international commerce and, with it, the growth of business correspondence. Many corporations or guilds established messenger systems to allow their members to maintain contacts with customers. Notable among these was the so-called Butcher Post (Metzger Post), which was able to combine the carrying of letters with the constant traveling that the trade required.

The mercantile corporations of Italy provided the most extensive and regular postal system of this period. Of particular importance were the links maintained from the mid-13th century between the great Italian commercial centres, such as Florence, Genoa, and Siena, and six important annual fairs held in the Champagne area of northern France. Two fixed dispatches were made to each of these fairs: the first to carry orders and commissions and the second to effect settlements. The service was carefully regulated. Conditions of acceptance, scales of payment, and timetables were laid down; the route was fixed, and hostels were maintained along the route. Since the Champagne fairs were attended by merchants from all over Europe, the postal system provided a valuable international link.

Early postal links between Italy and France

Italian business interests were also responsible for the only regular extra-European postal link of this period, between Venice and Constantinople. The extent and importance of Venetian business correspondence may be gauged from the fact that in 1320 the king of Persia accorded its couriers the right of free passage throughout his domains.

Russia shared in the general European trend toward the development of postal services in the 13th century. Horses and drivers for the transport of couriers were kept at regular staging posts to provide the so-called carriage express, which gradually developed into an organized system for the exchange of letters.

Growth of the post as a government monopoly. Institutional postal systems that developed during the later Middle Ages also conveyed letters between private persons, with or without official sanction and for a substantial fee in either case. Initially, such letters were relatively few. Outside the institutions with their own postal services, the number of literate people having interests that ranged beyond their own neighbourhoods was small.

In the late 15th century, however, the trend toward improved postal services was reinforced by Gutenberg's printing press (*c.* 1450) and the expansion of education. The growth of demand made letter carrying a profitable business, leading to the rise of private undertakings—the majority, like the Swiss Stumpelbotten, purely local in scope. Some, like the Paar family in Austria, developed postal organizations on a national scale. By far the most famous and extensive of such systems was that built up by the Thurn and Taxis family, who originally came from Bergamo near Milan, Italy. Under the patronage of the Habsburg emperors, they became the organizers of an

The Thurn and Taxis postal operation

extensive network of postal routes linking the imperial possessions. Their system developed throughout the 16th century until it covered most of Europe, using 20,000 couriers to operate a relay system that was speedy, efficient, and highly profitable.

Although the remnants of the Thurn and Taxis postal system survived in Germany up to 1867, it was essentially out of keeping—like the empire and the petty German states it served—with the main trend of development in Europe, the rise of nation-states with strong central governments. The first reflection of this trend in the postal sphere was the establishment of efficient national systems of relay posts under the control of the state. In France, Louis XI set up a Royal Postal Service in 1477 employing 230 mounted couriers. In England, a Master of the Posts was appointed by Henry VIII in 1516 to maintain a regular postal service along the main roads radiating from London. Neither of these systems was comprehensive, nor were they intended to serve the public. The security and regularity of the service along certain routes, however, inevitably resulted in an increasing amount of unofficial correspondence being carried. After initial attempts to prevent this practice in France, its fiscal advantages were realized, and the carrying of private mails was legalized about 1600. The basis of a real public service was not created until 1627, when fees and timetables were fixed and post offices established in the larger cities. In Britain, a separate public service was set up in 1635 by a royal proclamation "for the settling of the letter-office of England and Scotland." Thomas Witherings, a London merchant, was given the task of organizing regular services to run by day and night along the great post roads.

Legalizing private mails

In both countries, these state systems naturally began to develop into monopolies since such an evolution was seen by rulers as advantageous both to the security and to the revenues of the state. In England, the establishment of state posts along the principal roads was accompanied by the suppression, under the royal monopoly, of private and municipal posts, although "common carriers" were still permitted to convey letters on routes not covered by the royal system.

In 1672 France declared postal services to be a state monopoly under which operating rights were sold. Private undertakings that had established legal rights in this field were allowed to continue, but private messenger systems were eventually forced out of business by state competition or were bought out. In 1719 the University of Paris, the most important private competitor, gave up its last postal privileges in return for substantial compensation.

There was still opportunity for private enterprise to succeed by introducing services that were not at that time provided by the state systems. It was in this way that an important step in postal history, the establishment of local collection and delivery services in the great cities of London and Paris, was taken. London was the first city to benefit from an urban service when one William Dockwra set up his Penny Post in 1680. Striking features of the scheme were that letters were prepaid and stamped to indicate place of posting and the time they had been sent out for delivery. Deliveries were made almost hourly. Dockwra's scheme was so successful that he was prosecuted for infringing the state monopoly, and his service was closed down in November 1682, only to be reopened by the government. Not until 1759 was a similar local service introduced in Paris. It too was quickly absorbed by the state postal system; but its originator, Claude-Humbert Piarron de Chamousset, was paid compensation. Thus, the state monopolies expanded their scope, happily combining an improved service to the public with greater profitability.

The pace of postal progress in England during the later 18th century was accelerated by remarkable economic growth and a consequent demand for better mail services to the growing commercial and manufacturing centres. The most striking improvements came as the result of an extensive program of road building that began about 1765, paving the way for the era of the stagecoach. These were first used by the post office in 1784 and rapidly superseded the mounted postboys on the main routes. They began by averaging six or seven miles an hour, but continuing improvements in the roads and in the design of the vehicles pushed this up to 10 miles an hour in the 1830s. With the stagecoaches making possible a general reorganization of the entire system of mail circulation, letters could be delivered the morning after posting in towns more than 120 miles from London. A carefully regulated postal service—unprecedented for its standards of speed, frequency, and security—was evolved during this period.

The stagecoach

Despite the disruptive effects of the Revolutionary and Napoleonic wars, great progress was made throughout Europe in the late 18th and early 19th centuries in improving the speed and regularity of postal service and in providing internal delivery services for most of the larger cities. In the United States, too, postal services expanded at a remarkable rate: in 1789 only 75 post offices existed, but 40 years later there were more than 8,000.

Rowland Hill's reforms. The publication in 1837 of *Post Office Reform: Its Importance and Practicability,* by Rowland Hill (later Sir Rowland Hill), a British educator and tax reformer, is justly regarded as one of the most important milestones in postal progress. Based on an exhaustive study of the cost structure of postal operations, it demonstrated conclusively that conveyance charges were an insignificant factor in the total cost of handling a letter. The then current intricate charging scales based on distance were shown to be irrelevant: they inflated operating costs by requiring a host of clerks to apply them and to prepare complicated interoffice accounts. He also realized that another major item in the current cost structure—the collection of money payments on delivery—was easily avoidable. Hill's solution was a uniform rate of postage, regardless of distance, and prepayment of postage by means of adhesive stamps sold by the post office. Hill proposed a basic rate of one penny for each half ounce, calculating the "natural cost of distribution" to be slightly less than this. The cheapest current rate of postage was fourpence, and the average charge 6¼ pence (11.56 cents).

Foundation of modern postal systems

Not surprisingly, Hill's proposals rapidly gained strong support: popular agitation for the "penny post" overcame initial political disinterest, and the uniform rate and a system of prepayment by stamps were introduced in 1840. The originality of Hill's proposal for an adhesive postage stamp has been questioned but is irrelevant in considering the overall merits of his work. The significance of his reforms lies not only in the fact that they brought the post within the means of the mass of the people but also in the less obvious way in which they gave the postal system the technical capacity to deal with the vastly increased demand for postal service that ensued. The radical simplification of postal organization and methods characterizing Hill's reforms are the key to the speed and economy with which modern postal systems in many countries handle tens of millions of letters daily.

The chief features of Hill's system were gradually adopted in varying degrees by other countries throughout the world, first among which were Switzerland and Brazil in 1843.

The introduction of uniform cheap rates of postage for letters was accompanied by the establishment of even lower tariffs for newspapers (carried free in some countries) and for printed matter (*e.g.,* the British "Book Post" of 1848). These reduced rates were perhaps originally intended to favour the spread of education but quickly expanded, under the vigorous pressure of vested interests, to cover all sorts of commercial documents, advertising matter, magazines, etc. An inexpensive form of correspondence, the postcard, first introduced by Austria in 1869, was soon adopted by most other countries.

The general postal reforms of the mid-19th century ensured maximum benefit from the technological progress in transport in the great age of the railway and the steamship. These new modes of conveyance permitted a far speedier, more regular, and more reliable mail service, both internally and internationally. Railways in particular had a marked effect on the organization of postal work: instead of merely using trains to carry mailbags more speedily, postal administrations soon introduced the practice of sorting the letters in transit, using specially adapted railway cars. This greatly multiplied the advantages of railway

Effect of railways

conveyance. The first traveling, or railway, post offices ran in 1838 between Birmingham and Liverpool and London and Preston. By the end of the century, Britain, many continental European countries, the United States, and India had built up a complex network of such services, allowing the delivery of letters the day after mailing at distances three or four times as great as had been possible with the stagecoach, exceeding 400 miles in some cases.

International postal reform: the Universal Postal Union. The advent of the steamship and the railway had provided the opportunity for speedier international postal services, and the expansion of commerce ensured a growing demand for such facilities. Unfortunately, serious obstacles to the free exchange of international mails existed. Postal relations between states were the subject of bilateral postal treaties that had multiplied alarmingly during the 19th century. Most large European states were party to at least a dozen treaties by the 1860s. Such treaties necessitated the maintenance of detailed accounts between the countries concerned. Owing to the bewildering variety of currencies and units of weight and measurement then in use, the accounts attained a complexity described by a contemporary postmaster general of the United States as "almost beyond belief." Understandably, the users of the post suffered from this chaotic situation and from the high international postage rates that were its natural result.

The Paris Postal Conference

The first practical step toward reform did not come until May 1863, when the delegates of 15 European and American postal administrations met at the Paris Postal Conference, convening at the suggestion of the U.S. postmaster general. The conference established important general principles for the simplification of procedures, which were adopted as a model for subsequent bilateral treaties by the countries concerned.

The final step required the embodiment of these principles in a formal international treaty and the creation of an organization to administer them. An example was set by another conference at Paris two years later, which established the International Telegraph Union. Similar developments in the postal field were delayed by the advent of the American Civil War and the Franco-German War.

In 1868, however, a plan for a general postal union was put forward by the director of posts of the North German Confederation. Eventually, an international postal congress met, on Sept. 15, 1874, in Bern. It was attended by representatives of 22 states, all European except for Egypt and the United States. On October 9 a "Treaty concerning the Establishment of a General Postal Union" was signed. It was implemented on July 1, 1875, when the General Postal Union came into being. This title was changed in 1878 to the Universal Postal Union (UPU), and the basic treaty was renamed the Universal Postal Convention. The treaty provides a uniform framework of rules and procedures for the exchange of international mails. The union grew rapidly, increasing its membership to 55 within 10 years. By 1914, when China was admitted, it included almost all independent countries. The scope of the union's activities also expanded. In addition to its primary role, it gradually extended its functions to cover other international services provided by postal administrations, such as money orders (1878), parcel post (1885), postal checks (1920), cash on delivery (1947), and savings banks (1957). The UPU has been a specialized agency of the United Nations since 1948.

Activities of the UPU

Development of airmail. Balloon posts, apart from those organized during the sieges of Paris (1870) and Przemyśl (1915), for the most part only carried souvenir mail, owing to the balloons' uncontrollability. Airships overcame this problem but, again, did not establish themselves as a regular means of mail transport. It was only through the development of the airplane in the early decades of the 20th century that airmail truly came into its own. Certain experiments had been undertaken before World War I, such as airmail service between Hendon, on the northwestern outskirts of London, and Windsor in 1911 to mark the coronation of George V and flights between Paris and Bordeaux in 1913. Regular flights did not begin in the United States until 1918, and it was not until 1919, when the reliability of airplanes had considerably improved, that the first regular international service was introduced—between London and Paris. Other European links soon followed. On the long-distance continental and intercontinental routes airmail demonstrated its clear superiority over all surface transport. Technical factors delayed progress in opening up longer routes. The first American transcontinental airmail flight took place in 1920, but regular service did not begin until 1924. In 1926 a service between Egypt and Karāchi began and was linked to London in 1929. It was extended to Singapore by 1933 and to Australia in 1934. It was not until 1939 that a regular air service across the North Atlantic was launched with the takeoff of the *Yankee Clipper,* an American seaplane, on May 20.

While the growing availability of air flight did not affect basic postal organization as profoundly as the railways, its advantages of speed and operational reliability have been exploited in different ways since the 1920s. Prior to World War II, a number of European countries adopted the practice of forwarding letters to distant destinations at no extra cost to the sender (such as British mails sent to most parts of the British Empire). The consistently high costs of airmail curtailed this trend after the war. During the mid-1960s the UPU, in response to the continuing increase of aircraft capacity, adopted the policy of maximizing air conveyance of mail. In the mid-1970s, the concept of "surface air-lifted" (SAL) mails was developed in conjunction with the International Air Transport Association (IATA). This arrangement allows some mails to receive, for little or no surcharge, speedier transmission than by surface, but without the priority of fully surcharged mails. Use of SAL varies from country to country.

The aerogram

For individual correspondence, the most practical and inexpensive form of airmail remains the compact aerogram, which was introduced in Britain during World War II as a convenient way of writing to overseas military personnel. It consists of a sheet of lightweight paper suitably folded and gummed on all sides. Recognized by the UPU, the aerogram is available in most countries.

Advanced communications technology. Rapidly advancing computer and data transmission technologies of the late 20th century are being felt far more widely within the postal sector than were previous advances, such as improved roads, the railway, and the airplane. Although the latter enabled postal services to reform or enhance existing services, today's technologies go further by providing alternatives to the letter in the form of electronic messaging networks and electronic data-processing techniques to improve administrative efficiency.

NATIONAL POSTAL SYSTEMS

United States. Although the first official reference to overseas mail arrangements (concerning the receipt of overseas mail at Fairbanks' Tavern in Boston) dates to 1639, little real progress was made in building a postal system in colonial America until the appointment of Benjamin Franklin, formerly postmaster at Philadelphia, as deputy postmaster general for the American Colonies in 1753. Through diligent personal survey and inspection, he provided a more extensive, frequent, and speedier mail service, both within the Colonies and to England. Franklin built a sound foundation for the postal service in the United States, and, fittingly, he became its first postmaster general in 1775.

Postal service expanded rapidly after independence: annual revenue increased from $37,935 in 1790 to $1,707,000 by 1829, when the postmaster general first became a member of the Cabinet. The heavy cost of establishing a postal structure to keep pace with the remarkable economic progress of the country and the accelerating extension of its settled area caused expenditure to rise even faster than revenue. The trend toward annual postal deficits, which began in the 1820s, often exceeded an annual figure of $5,000,000 later in the 19th century.

By 1901, however, this expenditure had produced remarkable results. The accessibility, quality, and range of services provided had improved immeasurably. The number of post offices stood at a peak of 76,945. Postage rates had been considerably reduced with the gradual adoption of the principles of Rowland Hill: a single uniform

rate regardless of distance was adopted in 1863 (after an interim period with two rates since 1845), and postage stamps were introduced in 1847. Free collection services came with the provision of street letter boxes in 1858. A free delivery service was established in 1863, covering 49 cities and employing 440 letter carriers. By 1900 the service was provided at 796 offices by 15,322 carriers. The rural free delivery (RFD) service was introduced in 1896 and town delivery in 1912. These delivery services have greatly expanded their scope. The vast majority of mail is delivered by carriers, about one-tenth through post-office boxes, and only a small fraction at windows or counters.

Growth of services

The range of services available to the public has also grown steadily since the first supplementary postal service, registered mail, was introduced in 1855. The major milestones in this progress were postal money order service (1864); international money orders (1867); special delivery (1885); parcel post, with its accessory collect on delivery (COD) and insurances services (1913); and certified mail (1955), which provides proof of posting for items without intrinsic value. In 1911 a postal savings system was inaugurated, reaching a peak of more than 4,000,000 accounts in 1947. A decline to less than 1,000,000 depositors caused the service to be discontinued in 1966. Mail was formally divided into three classes in 1863, and a fourth was added in 1879. First-class, or letter, mail (called letter post in the United Kingdom) is the basis of the postal service monopoly and, as the class of mail most commonly used by the public, has generally had a simplified rate structure. The other classes were established according to mail content: second-class consists of newspapers and magazines, third-class encompasses other printed matter and merchandise weighing less than one pound, and fourth-class mail is either merchandise or printed matter that weighs one pound or more. The addition of these classes allowed the post office to adopt more complicated rate structures that would take into account factors affecting handling costs—such as the weight of the piece and the distance it would be conveyed. Second-class mail receives preferential rates because the dissemination of information through newspapers and other publications is considered to serve the public interest.

Division of mail into four classes

The post office has played a vital role as a pioneer and major user of all systems of transport as each was developed: the stagecoach, steamboat, canals, and railroads; the short-lived pony express; and airlines and motor vehicles. It also helped subsidize their development. A traveling post-office system, in which mail could be sorted in transit, was introduced experimentally in 1862, and it made railway mail service the dominant form of mail conveyance well into the 20th century. The gradual reduction of passenger train services during the 1930s led to the birth of a highway post-office service in 1941. Both of these services declined rapidly in the 1950s and 1960s. Railway post-office mileage was reduced from 96,400,000 in 1965 to 10,100,000 in 1969, and the number of highway post offices in operation during that period fell from 163 to none. Conversely, annual ton-miles of airmail flown grew from 188,103,000 in 1965 to more than 1,000,000,000 by the 1980s, indicating the significant trend toward air transportation of regular mail without surcharge.

The United States maintains the largest postal system in the world, handling almost half the world's volume of postal traffic. To deal with the problem of increasing deficits and to improve the overall management and efficiency of the post office, the U.S. Congress approved the Postal Reorganization Act of 1970, signed into law Aug. 12, 1970. The act transformed the Post Office Department into a government-owned corporation, called the United States Postal Service. Congress no longer retains power to fix postal tariffs (although changes may be vetoed) or to control employees' salaries, and political patronage has been virtually eliminated. Government subsidies continued on a declining basis until 1982, after which the U.S. Postal Service itself no longer received a direct subsidy from Congress. An indirect subsidy is still paid for certain mailers, however. These mailers, primarily nonprofit organizations or small publishers, pay lower rates than others, with Congress making up the difference in cost. The corporation has authority to raise capital to modernize its equipment and buildings. It is also subject to competition from private companies, a situation that in 1977 led to the introduction of Express Mail, which guaranteed overnight delivery.

Postal Reorganization Act of 1970

The availability of adequate funds for its mechanization and automation program has allowed the post office to benefit considerably from its sustained effort in research and development. More than half of all letter mail is handled by preparation and sorting machines, a trend greatly assisted by the ZIP (Zone Improvement Plan) Code program, which has come to be almost universally used.

Great Britain. The development of the British Post Office up to the reforms introduced by Rowland Hill has already been described. After 1840 the volume of postal traffic increased enormously and by 1870 had reached 10 times its prereform level. The growth was fostered by the introduction of new facilities, such as registration and postcards, and of preferential rates for books, printed papers, and samples. Financial services were also expanded: a savings bank was established in 1861, and postal orders were introduced in 1881 to supplement the money order service taken over from private interests in 1838. In 1883 a parcel post service was established.

British Post Office social services

The sweeping social reforms of the 20th century have given the post office an additional role as the chief payment agency for social security benefits, beginning with old-age pensions in 1908. This has been expanded to provide a variety of payments and also to collect large sums for state insurance schemes. The scale and range of financial transactions have been further boosted by the establishment in 1968 of the post office's banking arm, National Girobank, which provides an improved money transfer arrangement for the settlement of bills, as well as an account banking system and loan facilities.

Another important 20th-century trend has been the gradual recognition of the post office's role as a commercial enterprise rather than as a government revenue department. The process of achieving full commercial status took an important step forward in October 1969, when the post office became a public corporation. The British Telecommunications Act of 1981 divided the post office into two corporations, one for postal and banking operations and the other for telecommunications. This law also has provisions for the suspension of the post office's monopoly in certain categories of mail, allowing private companies to compete with it.

To maintain quality of service economically by removing the need to handle nonpriority mail at peak periods, a change to a two-tier system of letter classification was started in September 1968. The system abolished the complex preferential rate structure for printed papers and similar material—based on the contents of correspondence—and substituted the more relevant criterion of priority. The sender indicates the urgency of the item as being high (first-class) or lower (second-class) and pays on this basis, which simplifies the work of accepting the mail.

As in a number of other postal administrations, the sorting of mail has been gradually mechanized since the mid-1960s, with some 80 mechanized offices replacing more than 600 offices. The key to mechanization is an alphanumeric postal code that provides for sorting by machine at every stage of handling, including the carrier's delivery route. The coding equipment translates the postal code into a pattern of dots by means of which machines can sort mail at eight times the speed of manual sorting.

The post office transports mail by road, rail, and increasingly by air. More than half still travels by rail, using an extensive network of railway traveling post offices. To improve service and to enable next-day delivery for remote areas, letters are flown each night either by scheduled or chartered aircraft. New services based on communications and computer technology have been introduced. The Intelpost facsimile service operates nationally and with many other countries. An electronic mail system enables data for large mailings to be transmitted to local centres for enveloping and delivery.

France. The chief stages in the pre-Revolutionary development of postal services in France have already been

described. With the Revolution the system of farming, or selling the right to operate a postal service, was abolished. In the years following the 1789 Revolution, the postal and stagecoach services were reorganized into a Directorate of Posts, which became a national monopoly organization on June 16, 1801. After a brief return to the farming system, a general Directorate of Posts attached to the Ministry of Finance was created in 1804. During the 19th century, the service developed to keep pace with the Industrial Revolution, notably through improvements in administration and transportation. The first French postage stamp was issued on Jan. 1, 1849, thereby introducing the principle of prepayment as well as simplifying the rate structure. The scope and range of the postal administration's activities widened in the 19th century to include postal money orders (1817), registered letters (1829), parcel post (1881), the savings bank (1881), and postal checks (giro; 1918). During these years the post also took advantage of the newly developing methods of transport. Rail transport, for example, introduced in 1841, was so successful that by 1892 it had become an established means of conveyance; postal steamships served more than 100 ports; and bicycles and, later, motor vehicles were provided for postmen's collection rounds. In the early 1900s a full motor vehicle service was set up for the conveyance of mails in Paris.

French airmail service

The knack of turning technological innovation to good account was repeated with the airplane. The first airmail flights within France began in 1918, and in 1919 an irregular route was established between Avignon and Nice; by 1935 a network of routes linked the main French cities. During that period, the exploits of Antoine de Saint-Exupéry, Juan Mermoz, and other pilots of the Aéropostale airline, founded by Pierre Latécoère, served to establish an overseas network with French West Africa and Latin America.

World War II severely disrupted the postal service, and the German occupation forces exerted strict control over mail conveyed within a homeland that had been divided into virtually separate zones. The reinstatement of night airmail service in 1945 marked the beginning of a period of rapid reorganization for the postwar French postal service. Eventually the propeller-driven aircraft was replaced by relatively quiet turbojets, carrying 20 times the amount of airmail dispatched in 1948. Additionally, the postal service acquired its own custom-built rolling stock, which is used on the high-speed train run between Paris and Lyon. Daily mail traffic has grown by some two and a half times since the early postwar years, and, to cope with the increased volume, work began in the late 1970s on a network of mechanized sorting centres.

To speed up and reduce the cost of processing this huge quantity of mail traffic, the French postal service introduced a two-tier system in January 1969, thus enabling the customer to choose the priority of service desired by paying an appropriate charge. The postal service also entered the express mail market with the Postadex service, which operates with more than 50 other countries, and a national and international facsimile transmission service (Postéclair) has been established. An electronic mail printing and delivery service is being developed. The banking and savings service has also embarked upon a comprehensive program involving the microcomputerization of virtually every aspect of local office operation, as well as actively exploiting memory-card technology.

Germany. Postal organization in Germany remained on a relatively small scale until the latter half of the 19th century because of the numerous and fragmented sovereign states. The need for a more widely based postal system had been met, to a certain degree, since the 16th century, when the Thurn and Taxis postal service was begun. The fragmented political state of Germany, in fact, allowed the Thurn and Taxis organization to survive until 1867, when its last privileges were acquired by the postal service of the North German Confederation.

Establishment of the German postal service

The unification of Germany under Prussia during the second half of the 19th century, culminating in the establishment of the German Empire in January 1871, was followed by the creation of a German postal service in the same year, under a law establishing a state monopoly for conveying letters and newspapers. In 1924 the postal administration was accorded a considerable degree of financial autonomy, which allowed it to conduct business in a semicommercial manner, while still taking account of national economic and social factors in determining rate structure and pattern of service. The post office has developed a complete range of normal postal and financial services and an extensive postal passenger transport network and carries out various social security, revenue, and other agency functions.

Mail is transported chiefly by rail, but there is also an extensive complementary postal road network and an important night airmail service. The air service, inaugurated in 1961, carries letters and postcards without surcharge.

The present-day operation of the German post office has been affected by three major factors. First, there has been increased competition, which has brought about the adoption of customer-oriented business policies, along with appropriate reorganization. As a result, the post office entered the express mail market with overseas (Datapost) and inland services and introduced the surface air-lifted (SAL) parcel service to overseas destinations. Second, rapid technological development has taken place. This has been an agent of change in a number of ways: a new general messaging service, incorporating different forms of electronic mail and the traditional letter mail, is under development; entirely new services such as cash dispensing, interactive videotext, and money services based on memory cards have been introduced; and the post office has exploited modern technology to increase its administrative efficiency and ability to compete. The third factor is the extension of the postal system to serve former East Germany.

Introduction of new mail services

Italy. A long history of organized postal systems in Italy began with the *cursus publicus* of the Romans, but a modern national postal system was not established until 1862, when the long process of unifying the former small sovereign states into the new Kingdom of Italy had been completed. The new state postal service was given a monopoly for the collection, conveyance, and delivery of letters, printed papers, and newspapers, and a uniform tariff was established. The monopoly was surrendered for printed papers and newspapers in 1873 but was extended to cover parcels up to 20 kilograms in 1923. A registration service also began in 1862, and postal orders were introduced, extending on a national scale a service that had been made available to the public in the former Kingdom of Sardinia as early as 1845. The Post Office Savings Bank, set up in 1875, and the postal check service, founded in 1917, are other important aspects of the postal service. The post office also acts as an agency for the payment of such social security benefits as state pensions and various other grants.

Railways are the chief means of mail transport, although their use has declined since 1965, while road and particularly air transport have increased significantly, with the need for next-day service between major cities, such as Milan and Palermo. The Italian postal mechanization program moved forward significantly after the late 1970s. In the larger cities mechanized centres have been constructed in which electronic address-reading equipment processes certain mail for delivery.

Former Soviet Union. Following the October Revolution of 1917, postal services in the Soviet Union underwent important development, particularly in the Central Asian republics of Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, and Uzbekistan, where the number of post offices eventually increased to 30 to 40 times that of the 1913 figure. Before the dissolution of the Soviet Union in 1991, state enterprises and individual customers alike were served by a network of some 90,000 post offices, about three-fourths of which were located in rural areas that prior to 1917 had little or no service.

Airmail was especially important in the U.S.S.R. because of the vast distances involved. It accounted for about 60 percent of all postal traffic, which totaled some 54,500,000,000 items annually. Airmail continues to be central to the postal systems of such Soviet successor states as the Russian Federation.

Prompt delivery of central press publications had been

achieved by facsimile transmission of text, in some cases by satellite, for decentralized printing. These facilities contributed, in the early 1980s, to a significant increase in the volume of periodicals handled.

All types of mail were processed by a national network of large sorting centres. Work in public offices was being mechanized in various ways on a large scale, as exemplified by the national automation of postal order operations. Increasing use was made of computers at all levels of administration, particularly in mail handling and transport. Postal research efforts were centred on improving productivity, staff working conditions, and service to the public.

China. The first use of a postal system in China was under the Chou dynasty (*c.* 1111–255 BC). A reference by Confucius in the late 6th century demonstrates that it was already renowned for its efficiency: "The influence of the righteous travels faster than a royal edict by post-station service."

By the late 3rd century BC the postal network consisted of relays of couriers, who changed their mounts at staging posts about nine miles apart. This network was considerably enlarged following the opening of new territories in Central Asia under the Han dynasty (206 BC–AD 220), during which time contacts were made with the Romans, who, it was observed, maintained a postal system similar to that of the Han. Further improvements were made under the T'ang dynasty (618–907). During this period the number of staging posts increased, and correspondence could be conveyed by road or by river. Administered centrally by the secretary for transport under the Ministry of War, the postal service underwent periodic checks for quality of service. It was possible for urgent documents to be conveyed up to 93 miles in a day. The Sung dynasty (960–1279) instituted a parallel express service for military correspondence, its couriers regularly traveling as far as 124 miles in a day and even more in cases of extreme urgency. The postal network that had evolved under the T'ang and the Sung was to prove a valuable base for organizing the posts under later dynasties. In the late 13th century Marco Polo revealed to Europe the quality of the Chinese postal system under the Yüan, or Mongol, dynasty (1206–1368). Nothing comparable existed in Europe at the time.

Improvements under the Han dynasty

Until the end of the 14th century, the post was used purely for the conveyance of official documents. At the beginning of the 15th century, private post offices for the use of traders appeared, conveying private correspondence and arranging payment transfers. During the middle years of the Ch'ing dynasty (1644–1911/12), there were several thousand of these private post offices. In 1896 the Imperial Post was created and organized along European lines, and the old staging points that had functioned for more than 3,000 years were phased out. The new state system gradually absorbed the business of the private companies, although the last one did not close its doors until 1935. The last of the foreign post offices (maintained by Britain, the United States, France, Germany, and the Soviet Union) was withdrawn by the end of 1922.

The Imperial Post

When the Republic was proclaimed after the 1911 Revolution had overthrown the Ch'ing dynasty, the service was renamed the Chinese Post. In 1914 China joined the UPU. Development of services was extremely slow, however, because of internal strife and eight years of resistance to the Japanese invasion. Thus, on the eve of the founding of the People's Republic, there were only 4,868 postal establishments throughout the whole of China, of which 463 alone were in country areas. Postal transport was minimal by present-day standards.

One month after the proclamation of the People's Republic of China on Oct. 1, 1949, the Ministry of Posts and Telecommunications was established. On the mainland this administration reorganized the postal service, known thereafter as the People's Post of China. Much was accomplished in the years that followed. The mainland postal network grew to more than 3,000,000 miles of roads, railways, and air routes centred upon the capital, Peking. It reached into every corner of the land to serve remote towns and villages where prior to 1949 the postal service hardly had been known. Letter mail has increased fourfold since the late 1940s.

Postal research centres set up during the 1950s have actively exploited new technology in the design and manufacture of specialized equipment for the postal service. Notable examples include sorting machines using optical character or bar recognition (OCR or OBR) and computer- and microprocessor-controlled sorting machines.

India. The earliest reference to an official postal system in India was made in the 14th century by the Arab traveler and historian Ibn Baṭṭūṭah, remarking upon the organized official service of mounted couriers and runners. The system was brought to its height during the 16th century under the great Mughal emperor Akbar, with a network of 2,000 miles of post roads. Two centuries of political turmoil, without a strong central authority, destroyed this courier system. It was not until 1766 that an official post was reestablished to serve a new ruling power. It was made available for public use in 1774.

In 1837 the Imperial Post was established and granted a monopoly to provide efficient postal communications between the seat of government at Calcutta and the principal provincial towns. Within the provincial districts a complementary local service was maintained. In 1854 the basis of the modern Indian Post Office was established when these parallel systems were merged under a director general. A uniform postage rate was then introduced, a step of particular significance in so vast a country.

Establishment of the Indian Post Office

International postal relations developed rapidly after the establishment of a weekly steamer service between Bombay and England in 1867, with India becoming a member of the UPU by 1876. At the same time, the post office began to expand the range of services provided to the public: COD in 1877, an insured service in 1878, and money orders in 1880.

Internal communications were improved to keep pace with the progress made in these other fields. The "bullock train"—the Indian equivalent of the mail coach—gave way to the growing network of railways. A regular traveling post-office service was introduced in 1870. Although experiments with airmail conveyance started in 1911, a regular inland service did not begin until 1932. Rapid expansion followed in 1949. A complex night airmail network connects all major cities, carrying a growing percentage of the mail. Railways have retained much of the traffic as a result of speedier train services, although motor vehicles have become the dominant carrier of mail.

More traditional forms of transport—the foot runner, horse, mule, camel, bullock cart, and bicycle—still help distribute mail to many of India's villages. More than one-half of these now have daily delivery of letters, and almost all have at least weekly delivery.

Pakistan. Although Pakistan did not exist as an independent sovereign state until 1947, its postal history extends to the official postal systems established by Muslim emperors. Particularly noteworthy was the network of post relays at caravansaries (inns) established by the emperor Shēr Shāh of Sūr in the early 16th century. Under British rule in 1852, the province of Sind had the distinction of being the first region in the subcontinent to adopt a uniform letter postage rate.

On becoming independent, Pakistan was faced with a particularly difficult problem of postal communications because its eastern and western regions were separated by 1,000 miles of Indian territory. The growth of civil aviation services and maritime links overcame the initial difficulty. In 1952 an airmail service was introduced between East and West Pakistan in which letters and postcards were carried without payment of surcharge. This concessionary airmail system was extended to cover the internal mail service of both regions in 1959. The post office made use of the entire network of the national airline to connect 29 key points throughout the country. The service for printed paper and parcel mail was improved in 1962 by the introduction of direct sea-mail links between Karāchi and Chittagong.

Pakistan airmail service

The creation of Bangladesh out of East Pakistan in 1971 allowed Pakistan to improve its internal mail services, especially in rural areas, where three-fourths of the post offices are located. Postal facilities were greatly expanded throughout the country.

The post office is run on a commercial accounting basis but with strong emphasis on its function as a public utility. Postal traffic has grown rapidly. To improve its capacity to deal efficiently with the continuing growth, the postal administration has pursued a carefully considered and moderate policy of mechanization. At some of the busiest offices, machines designed to facilitate the acceptance of registered mail have been introduced. To speed up the handling of mail in sorting offices, electromechanical sorting machines have been installed at Karāchi and Lahore. All major post offices now use electrically operated stamp-canceling machines, as well as franking machines.

Japan. Although official and private systems of communications had existed in Japan from ancient times, it was not until 1870 that the creation of a comprehensive government-operated postal service was proposed. The idea was put forward by Hisoka Maejima, often called "the Father of the Post" in Japan. It was rapidly accepted by the government, which set up a service between Tokyo and Ōsaka on April 20, 1871, and extended it throughout the country in July 1872. In 1873 the postal service was proclaimed a monopoly and private courier systems were prohibited, a uniform postal tariff scale was adopted, and postage stamps and postcards were introduced. The first official overseas mail service was established in 1875 with the United States. International postal relations expanded rapidly, and in 1877 Japan became a member of the UPU.

Other landmarks include the introduction of a parcel post service in 1892, express delivery in 1911, airmail in 1929, and the achievement of a special accounting status on semicommercial lines in 1934.

A distinctive feature of the postal scene in Japan is the special New Year's Mail Service, introduced in 1900. Operating partly for the benefit of charities, this provides for the timely delivery of billions of New Year greetings.

Japan's postal code address system

The shortage of manual labour and the growth in postal business resulting from the country's rapid economic development have also led to the adoption of an extensive mechanization policy. To facilitate this process, a postal code address system was introduced in July 1968. Japanese-produced segregating and sorting equipment, including automatic postal code reader-sorters, has been installed at major post offices throughout the country.

POSTAL SERVICES IN THE DEVELOPING COUNTRIES

The establishment of efficient and comprehensive postal systems in the developing countries is important internationally as well as from the purely domestic viewpoint. Successful maintenance and progressive improvement of international postal service require the effective cooperation of all member countries of the UPU.

The internal need for a good postal service is sufficient in itself, however, to justify a high priority. A country-wide network of post offices provides government with many points of contact with its people for implementing administrative programs in such fields as social security, taxation, and public information. When its operation is properly developed, the post office may also become one of the principal employers in a country; it may help to promote economic growth through its need for buildings, vehicles, and equipment; and it has the potential to become a major user of transport services. The employment potential of a postal system is evident from the fact that the percentage of the working population engaged in providing postal services is generally several times higher in developed than in developing countries.

Importance to national economies

An efficient postal service, in addition to promoting national cohesion, provides an essential infrastructure for the expansion of industry and commerce. Postal money transfer and savings services are particularly valuable in developing economies, where banking facilities are limited. They may generate large resources that can be used for public investment.

Postal administrations in developed countries have long appreciated the importance of collaborating in the improvement of postal services throughout the world, and this participation in postal technical assistance is an aim embodied in the constitution of the Universal Postal Union. In addition to fostering bilateral assistance between members, the UPU itself has, since 1964, taken part as a specialized agency in the United Nations Development Programme (UNDP). UPU activities in this field are monitored and realigned as necessary at each postal congress.

In view of the rapid expansion of postal business in most developing countries, the training of staff is a most urgent need. The UPU's initial activity has, therefore, been largely devoted to this field. The needs of individual countries are evaluated by traveling postal experts attached to the International Bureau in Bern. In addition, expert aid is provided for the establishment of national training schools for postal workers and regional centres for middle and higher management staff, such as those set up in Abidjan, Ivory Coast, and in Bangkok and Damascus. Special instructor-training courses are conducted in Britain and France, as well as in multinational schools in Africa, Asia, and Latin America, where seminars are also held for higher grade officials. Specialist aid is also provided to evaluate operational and organizational needs and subsequently to enlist UNDP assistance in meeting them. The missions of experts are thus directed either to the postal services in general or to a certain sector of postal activity. Some larger projects may include overall postal organization as well as postal training; such projects have been undertaken in Ethiopia and Niger. Fellowships are awarded for overseas study of specialized subjects, such as international mails, philately, and, in particular, instructor training. Although the UNDP provides the major share of technical assistance funds, the UPU has increasingly supplemented these from its own resources to meet the increasing demand for assistance from its members. Moreover, a considerable amount of aid is given directly by developed countries within the framework of their own bilateral technical assistance arrangements.

Many developing countries are not able to provide even the minimum scale of postal facilities. To underline the pressing need for improvement and to provide a yardstick for future progress, the UPU has adopted, for the Second and Third United Nations Development Decades, certain key planning objectives, covering management, quality of service, promotion of its financial services, and improved public information. These are to be carried out to favour, in particular, the least developed countries.

UPU planning objectives

THE INTERNATIONAL SYSTEM

International mail is a key means of furthering economic, social, and cultural links between nations. The international postal system is in itself an outstanding example of worldwide organization and mutual trust. A postal administration relies completely upon the postal authorities of other countries to play their parts in ensuring that its foreign mails reach their destination.

International cooperation in this field has been greatly facilitated by the Universal Postal Union since 1875. It has built a comprehensive international organization, with a membership composed of numerous sovereign states and several dependent territories. The postal administrations that are not represented generally follow the rules of the UPU.

These fundamental rules of the international postal service are to be found in the Universal Postal Convention and General Regulations and have been little changed since adoption of the Bern Treaty. The first basic principle is that all member countries form "a single postal territory for the reciprocal exchange of correspondence." From it is derived the principle of freedom of transit: every member country guaranteeing to respect the inviolability of transit mails and to forward them by the most rapid transport used for its own mails.

Another important principle is that the charges for letter-post items are not shared. Since 1875 each country has retained the postage it collects on international mail. Although intermediate countries are paid for transit service, the country in which the mail is delivered receives no payment. This principle was adopted in order to minimize the need for complex international accounts and was justified on the supposition that a letter normally generates a reply. Certain developing countries, however, have found themselves at a considerable disadvantage under this rule,

due to an excessive imbalance between incoming and outgoing mail. To remedy this, the 1969 Congress of Tokyo provided for compensatory payments in such cases.

Provisions of the Universal Postal Convention

As a further measure of simplification, the convention prescribes international postal charges, as well as agreed tolerances, and specifies weight steps, limits of size, and conditions of acceptance for letter-post items. Disputes between postal administrations, which usually concern allocation of liability for the loss of registered or insured items, are to be settled by arbitration. The convention is completed by two other basic documents: the Final Protocol, which allows member countries to register certain general and specific reservations to the provisions of the convention; and the Detailed Regulations for implementing the convention. Apart from these obligatory documents, there are a number of optional agreements concerning services, such as parcel post and cash on delivery. The provision of a registration service is compulsory under the convention.

Mention should also be made of the constitutive acts of the union that prescribe its general aims, its organization, its financial structure, and the rules of membership, namely, the constitution and its general regulations.

This comprehensive framework of international regulations is regularly revised to take account of changing circumstances and technical advances. This is the chief function of the union's quinquennial congress. Between congresses, the continuity of the union's work is ensured by its elected Executive Council and its permanent office in Bern, the International Bureau. The bureau acts as a clearinghouse for the settlement of international accounts and for the exchange of information between members, especially notifications of important operational and organizational changes. Problems arising in the technical, operational, and economic fields are studied by another permanent organ, the Consultative Council for Postal Studies (CCPS). Regular contact is also maintained with other international bodies, such as the International Telecommunications Union and the International Standards Organisation.

The UPU Constitution authorizes member countries to establish restricted unions, a provision that enables regional groups such as the Arab, African, and Asian–Pacific postal unions to conclude agreements aimed at improving postal services between their members by such means as reduced rates of postage or the elimination of transit fees. These agreements are more easily achieved on a limited regional basis than on a worldwide scale, and the restricted unions have a valuable role in the task of the UPU, which is, basically, to improve international postal service by simplifying its organization and reducing its cost.

POSTAL TECHNOLOGY

Technological progress in postal transport. Postal administrations have been among the first to utilize new forms of transport. They have often applied considerable technical skill in maximizing the benefits to be derived from progress in this field, particularly in originating the traveling post-office concept and apparatus enabling express trains to pick up and discharge mails without slowing. They have also developed their own transport systems to combat traffic congestion in certain busy cities, such as the pneumatic tubes of Paris, New York, and other cities and the automatic underground railway, opened in 1927, that links London's chief mail centres to railway terminals.

Experiments with missilery

The advent of aerospace and telecommunications technology in the mid-20th century gave rise to research aimed at adapting this technology to postal systems. Experiments have been conducted using ballistic missiles to transport mail, but this remains a novelty because of costs and the problems of reusability and accuracy. Advances in computer and message transmission technologies are, however, being utilized by postal administrations.

Since 1980 public facsimile services have been available in a number of advanced postal administrations in various parts of the world. The United States, Great Britain, France, and Sweden were among the first countries to introduce tele-impression services, whereby bulk correspondence in electronic form is transmitted to regional postal printing centres for enveloping and delivery.

Automation of mail handling. Since the 1950s there has been a marked intensification of research and development efforts to apply technology to the handling of mails, especially in countries faced by manpower problems and higher labour costs. The wide variety of projects undertaken in many countries and the progress made have been summarized in CCPS studies.

Actual implementation has generally been slower than expected. There have been good reasons for this. Primarily, most postal administrations, being government agencies, are subject to strict control of their capital investment programs. Second, mail traffic patterns—with marked peaks of work—make economic utilization of machines difficult: the introduction of measures to counteract this problem takes considerable time. Similarly, the introduction of postal address codes and the standardization of sizes of envelopes and cards, which are prerequisites for mechanical handling, are relatively slow because of difficulties inherent in the change of procedures.

Materials-handling equipment. Postal systems continue to rely heavily on human labour for bulk materials handling and distribution, both at loading bays and between work processes within sorting centres. New mail centres, however, are normally built in the style of factories and include all appropriate materials-handling equipment.

Equipment used for loading and unloading sacks of mail, rigid containers, and loose parcels includes mobile belt conveyors, roller conveyors, forklift trucks, mobile and fixed cranes, and table lifts. Handling equipment within buildings includes chain conveyors; horizontal and rising belt conveyors of all types, for the transport of loose letters, packets, and trays of letters (notably used for continuous clearance of public posting boxes); tow conveyors, which allow wheeled containers to be hooked onto a fixed-path underfloor traction system; bucket or pan elevators; and chutes and other gravity devices.

Buffer-storage facilities

The use of a wide range of equipment is necessitated by the varied handling characteristics of different types of mail at particular stages. Buffer-storage facilities, in the form of ramps, hoppers, and moving belts, have to be incorporated to compensate for normal postal traffic fluctuations. The smooth distribution of traffic through the system is often monitored by closed-circuit television, which allows effective centralized control. Automatic regulation and recording, using a variety of sensing and counting devices linked to a computer, are the ideal. Modern systems-engineering techniques are thus able to ensure a carefully planned continuous mechanized mail flow with maximum productivity benefits.

Segregating machines. Mail collected from branch post offices and street mailboxes, although for the most part made up of ordinary letters and cards, also contains small parcels, newspapers, magazines, and large envelopes. These items, because of their size or shape, cannot be handled on machinery designed for the normal-sized letter and have to be segregated from the majority of standard "machinable" letters. Owing to its varied characteristics, most packet mail has to be manually stamped and sorted, although its movement between work processes may be fully mechanized. So-called packet sorting machines are, in fact, essentially conveyor systems for distributing manually sorted mail.

A commonly adopted type of segregator consists of a laterally inclined rotating drum, into the upper end of which a regulated flow of "mixed" mail is fed from a storage conveyor. Letters within a thickness standard, but of excessive length or breadth, are picked out by various simple mechanical devices installed on the conveyor belt that eventually delivers machinable letters to the storage stacks of the facer–canceler equipment.

Facing and canceling equipment. Facing is the process of aligning letters so that all will have the address side facing the canceler, with stamps in a uniform position. The process is normally combined with a separation of the mail into at least two streams, letter and printed-paper rate or first- and second-class, to allow priority handling for one of the streams.

Facer–canceler machines perform these processes by passing letters through sensing or stamp-detecting units,

which identify the presence or absence of a stamp on the side of the envelope facing them, and, when present, its position. Sensing units are also designed to separate mail in the priority class from nonpriority mail by identifying the stamp or commonly used combination of stamps representing the basic postage rate and manipulating selector gates accordingly. This identification is usually achieved by printing distinctive indexes on the stamps in normally invisible, phosphorescent or luminescent inks that are sensitive to ultraviolet radiation emitted by the sensing unit.

Coding and sorting machines. For manual sorting of letters, each operator normally uses a device with between 40 and 50 pigeonholes. This has been found by most administrations to be the optimum arrangement in view of the limited arm span and "memory" of the sorter. The development of various types of postal codes was aimed at making the sorting of a coded letter a mechanical process for the operator by dispensing with the need to memorize a sorting plan. To be totally effective these schemes need complete public cooperation, a requirement that has been difficult to achieve.

Mechanization by postal code

Postal administrations have responded to this dilemma by concentrating research on using an operator only to impress the postal code on each letter, employing phosphorescent or magnetic ink patterns that can be read by a sensing unit attached to a sorting machine. After the code has been impressed, the letter can be sorted at any subsequent stage by high-speed automatic machines, which are no longer utilized at the pace of a single operator and indeed can take the output of several operators. Furthermore, any second sortation required—even at an intermediate office or where the code includes the necessary information to letter carriers' routes at the delivery office—does not need further manual operations. Another potential advantage of this method is that letters may be directly encoded by the mail-processing machines used by large-volume mailers.

Optical character recognition. The ultimate aim in automated sorting has been to perfect a machine that can read some or all elements of the address on letters. Research in this field has been conducted in most of the industrial nations with sophisticated postal services. The immediate aims of these national research programs vary insofar as the type of character to be recognized is concerned: printed, typewritten, or addressing-machine characters; stylized handwritten scripts; and even ordinary handwriting. Some administrations require the machine to read a purely numeric code, others an alphanumeric code, and others the names of towns or regions. Several different techniques are used for the basic task of pattern matching in identifying the characters. For example, the observed character as a whole may be compared with matrices registered in the memory of the machine. Or the different traits of the character observed—vertical or horizontal strokes, curves, etc.—may be analyzed and their combination successively compared with a series of models registered by the computer.

Advanced machine sorting

An optical character reader (OCR) can be designed to either directly sort mail or mark it with a machine-readable code so that sorting at subsequent stages can be carried out by high-speed automatic machines. In 1965 the U.S. Postal Service began experimenting with an alphanumeric OCR. By the early 1980s the service had developed a machine capable of scanning up to three lines of an address, verifying the postal code, and imprinting the letter with a routing code.

Research in the United States subsequently has concentrated on various systems that print a machine-readable bar code to allow for high-speed automatic processing to individual carrier routes or blocks of addresses within carrier routes. In 1983 the U.S. Postal Service began deploying OCR's with this capability to major post offices throughout the country. The postal service regards this application of automation, combined with the use of ZIP+4 (a nine-digit postal code) by business mailers, as a major means of keeping postal costs under control as mail volumes expand.

Numerical speech translator. Another line of research being pursued in the United States is the development of equipment that translates five- and nine-digit ZIP codes and sorting-code numbers spoken by an operator into instructions for a sorting machine. Since this system obviates the need for a keyboard, it leaves the operator's hands free, making it particularly valuable in the operation of parcel- and sack-sorting machines. It also eliminates the need for keyboard training of operators. The testing of the equipment includes determination of the effects of regional speech variations, background noise, and operator speech fatigue.

BIBLIOGRAPHY. Postal services are treated in official publications, such as legislative texts, manuals, working instructions, postal guides, annual reports, and commission reports; much of this material is available to the public. Many postal administrations publish periodical reviews or bulletins informing employees and users of recent events and trends. ROWLAND HILL, *Post Office Reform: Its Importance and Practicability,* 4th ed. (1838), is a work by an expert whose reforms affected subsequent worldwide postal development; ROWLAND HILL and GEORGE BIRKBECK HILL, *The Life of Sir Rowland Hill and the History of Penny Postage,* 2 vol. (1880), provides a fuller account of Sir Rowland's innovations. Detailed studies of specific postal service aspects are A.D. SMITH, *The Development of Rates of Postage* (1917); A.D. LITTLE, *The Market for Postal Services* (1968); and MICHAEL E. CORBY, *The Postal Business 1969–79: A Study in Public Sector Management* (1979). International mail services are treated in GEORGE A. CODDING, JR., *The Universal Postal Union* (1964); and L. ZILLIACUS, *Mail for the World: From the Carrier to the Universal Postal Union* (1953), a popular treatment; both F. STAFF, *Transatlantic Mail* (1956, reissued 1980); and HOWARD ROBINSON, *Carrying British Mails Overseas* (1964), provide detailed treatment of the expansion of international mail conveyance.

General and historical literature includes HANS RACKOW (ed.), *Handwörterbuch des Postwesens,* 2nd ed. (1953), a small but comprehensive encyclopaedia; HMSO, *The Post Office: An Historical Summary* (1911); IVIE G.J. HAMILTON, *An Outline of Postal History and Practice with a History of the Post Office of India* (1910); HOWARD. ROBINSON, *The British Post Office* (1948, reissued 1970); and D. NORONA (ed.), *Cyclopedia of United States Postmarks and Postal History* (1933–35, reprinted 1975). Studies concerning the United States include PRESIDENT'S COMMISSION ON POSTAL ORGANIZATION, *Towards Postal Excellence,* 5 vol. (1968), also known as the Kappel Commission Report; CARL H. SCHEELE, *A Short History of the Mail Service* (1970); G. CULLINAN, *The Post Office Department* (1968); ROBERT J. MYERS, *The Coming Collapse of the Post Office* (1975), ALAN L. SORKIN, *The Economics of the Postal System: Alternatives and Reform* (1980); RICHARD J. MARGOLIS, *At the Crossroads: An Inquiry into Rural Post Offices and the Communities They Serve* (1980); JOHN T. TIERNEY, *Postal Reorganization: Managing the Public's Business* (1981); and *Evaluation of The United States Postal Service* (1982), a report by a panel of the National Academy of Public Administration. Somewhat dated but still useful are WILLIAM SMITH, *The History of the Post Office in British North America, 1639–1870* (1920, reprinted 1973); and WAYNE E. FULLER, *R.F.D.: The Changing Face of Rural America* (1964, reprinted 1966). The postal history of France is extensively described in EUGÈNE VAILLÉ, *Histoire générale des postes françaises,* 6 vol. (1947–53), and *Histoire des postes françaises,* 2 vol. (1946–47). Postal services in developing countries are treated in UNION POSTALE UNIVERSELLE, *Memorandum on the Role of the Post as a Factor in Economic and Social Development* (1969).

The origins, development, and procedures of the Universal Postal Union are treated in GEORGE A. CODDING, JR., *The Universal Postal Union* (1964); see also the bimonthly review *Union Postale,* which is devoted to a wide range of questions concerning the UPU and various general aspects of postal activities; and "Collection of Postal Studies," a series by the UPU's CONSULTATIVE COUNCIL FOR POSTAL STUDIES, comprising more than 150 reports on postal service modernization, mechanization, economics, transport, and working methods, available in French with much of the material also available in English.

(An.C.B./Ed.)

Prague

One of the finest cities of Europe, Prague (Czech: Praha), the capital of the Czech Republic and that nation's major economic and cultural centre, lies at the heart of the Continent. The city has a rich architectural heritage that reflects both the uncertain currents of history in Bohemia and an urban life extending back more than 1,000 years. The physical attractions and landmarks of the city are many. Among the finest is the Charles Bridge (Karlův most), which stands astride the Vltava River. The winding course of the Vltava, with its succession of bridges and changing vistas, contrasts with the ever-present backdrop of the great castle of Hradčany (Prague Castle), which dominates the left-bank region of the city from behind massive walls set high on a hill. The narrow streets and little taverns and restaurants of the older quarters contrast with the broad sweep of Wenceslas Square and modern parks and housing developments, while the great 18th-century Baroque palaces have their own elegance and splendour. Seen from the surrounding hills, the many church towers make up a unique perspective, giving Prague its description as the "city of a hundred spires." This architectural harmony was enhanced by post-1945 planning, which preserved the ancient core of the city as a major monument and carefully supervised all modern building.

Prague is famous for its cultural life. Wolfgang Amadeus Mozart lived there, and his *Prague Symphony* and *Don Giovanni* were first performed in the city. In addition, the lyric music of the great Czech composers Bedřich Smetana, Antonín Dvořák, and Leoš Janáček is commemorated each year in a spring music festival. The U kalicha ("At the Chalice") beer parlour, which is still popular with local residents and tourists alike, provided the setting for the humorously antiauthoritarian activities of Schweik, immortalized by the novelist Jaroslav Hašek in *The Good Soldier Schweik*. The writings of Franz Kafka, dwelling in a different way on the dilemmas and predicaments of modern life, also seem indissolubly linked with life in this city.

This article is divided into the following sections:

Physical and human geography 1077
The landscape 1077
The people 1077
The economy 1077
Industry and employment
Transportation
Administration and social conditions 1078
Government
Public services
Education
Health
Cultural life 1078
History 1079
The early period 1079
The foundation of the city
Medieval growth
The Reformation and the Thirty Years' War
Evolution of the modern city 1080
Bibliography 1080

Physical and human geography

THE LANDSCAPE

From its original small riverside settlements, Prague has spread over its hills, up river valleys, and along riverside terraces. The Prague metropolitan area covers 192 square miles (496 square kilometres).

The city's core, with its historic buildings, bridges, and museums, is a major centre of employment and traffic congestion. Around the core is a mixed zone of industrial and residential areas, containing about half the city's population and nearly half its jobs. Surrounding this area is the outer city development zone, and beyond this is yet another zone of development containing new industrial areas, parks and recreation areas, and sports facilities. Finally, there is a belt of agricultural land and open countryside, where farms and market gardening projects satisfy Prague's demand for food.

The lowest point in the city is 623 feet (190 metres) above sea level, and the highest point is 1,247 feet (380 metres) on White Mountain (Bílá hora). The climate of Prague is typically mid-continental, with temperatures there averaging 67° F (19.3° C) in July and 31° F (−0.6° C) in January.

THE PEOPLE

Prague has a homogeneous population. There is a small Slovak community, but the overwhelming majority of residents are Czechs. The city has a number of demographic peculiarities stemming mainly from the effects of World War II; there are more women than men, and a sizable proportion of the female population is past the age of fertility. The natural rate of population increase is very small. A tendency toward small families is a reflection of both difficulties in housing and increased participation by both parents in the workforce. Migration into the city has continued.

THE ECONOMY

Industry and employment. Though Prague is renowned for its cultural life and monuments, it has also played an important role in the economic life of what is now the Czech Republic since the early and intensive development in the 19th century of such industries as those producing textiles and machinery. Industry is the largest employer, followed by commerce, construction, education, culture, administration, and transport and communications. Nearly half the labour force is female; the proportion of women is almost one-half in manufacturing, but it is considerably higher in education and culture, in trade, and in the health field.

Women in the labour force

In manufacturing, the production of machinery occupies the majority of workers, followed, in about equal numbers, by the production of food, electronics, and chemicals. A large portion of the city's manufactured products are consumer goods.

As part of a decentralization plan for the city's growth, since the 1950s industrial districts and warehouses have been located or relocated on the outskirts of Prague. The aim is to provide increased job opportunities in the vicinity of new residential areas, thereby reducing the pressure on the city's central core.

Transportation. Much of the inner-city transportation is handled by bus, tram, and subway (metro) systems, which are inexpensive and subsidized. Despite the efforts to meet the demands of the growing population with an adequate public transportation system, the number of passenger cars and commercial vehicles has increased, resulting in plans for a major urban motorway system to include 10 radial arteries connecting Prague with the national road network.

Prague is one of the nation's major railway junctions, with three main stations and three freight transport circuits. The international airport at nearby Ruzyně was expanded and modernized in the 1960s to serve as a hub



Statue of Jan Hus in the Old Town Square.
Robert Harding Picture Library

at the centre of Europe. A new port has been built at the confluence of the Vltava and Berounka rivers. The passenger boats that ply the Vltava during the summer are a popular tourist attraction.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Prague is the seat of government of the entire Czech Republic. The city is also the administrative centre for the Central Bohemian Region, of which the conurbation occupies about one-third. The Prague Municipal National Committee and the Central Bohemian Regional National Committee coordinate town planning, environmental control, and other projects. Prague is divided into districts, each with its own district national committee, while the settlements included in Prague since the late 1940s have retained their own committees.

Public services. The standard municipal services—the supply of natural gas, electricity, and water and the treatment and disposal of sewage and refuse—were consolidated under state control after World War II and have been considerably modernized and expanded as part of overall urban planning. The high percentage of employed women has caused municipal authorities to turn attention toward the provision of nurseries for the children of working mothers. Other facilities include swimming pools, often run in conjunction with sports organizations. On the river the city provides mooring positions for pleasure boats.

Housing shortage

Like cities in other eastern European countries, Prague has difficulties with the supply and maintenance of housing. Much of the housing in the inner city consists of small apartments in need of renovation and modernization, while the rate of construction of apartments in the newer zones lags behind the need. Privately owned houses constitute less than 15 percent of all Prague's housing units. In response to the problem, new housing developments have been built in the peripheral areas. Referred to as "towns," they include North Town (Severní město), South Town (Jižní město), and Southwest Town (Jihozápadní město).

During the Communist era, all retail establishments—food and department stores and self-service establishments—were publicly owned and were part of the municipal system. There are numerous small restaurants and taverns, many of which—especially in the Malá Strana (Lesser Quarter)—have an intimate and historic atmosphere and offer fine views of the city and the river.

Education. There are several institutions of higher education in Prague, but by far the most famous is Charles University, founded in 1348 and the oldest in central Europe. The Academy of Arts and the Academy of Music (with a conservatory founded in 1811) are also important. The activity of the Czech Academy of Sciences (founded in 1952 as the Czechoslovak Academy of Sciences) is supplemented by many specialized institutions; the academy sponsors a number of international congresses. Higher education in the city benefits from a tradition that can count among its scholars and teachers the great 17th-century astronomers Tycho Brahe and Johannes Kepler and the noted modern physicist Albert Einstein, who taught in Prague in 1911–12.

Health. As the capital city, Prague contains some of the country's main health facilities. These include hospitals, specialized medical clinics, and outpatient clinics. The most noted facilities are those that specialize in plastic surgery, orthopedics, and urology.

CULTURAL LIFE

Prague has a renowned and active musical life, which reaches a high point each year in the internationally known spring music festival. The city's fine orchestras—the Prague Symphony and the Czech Philharmonic—have

won reputations abroad. Theatrical traditions are also strong, with more than 20 well-attended theatres in the city. There are also many museums and galleries, and a Palace of Culture was completed in 1981.

Architecture

Perhaps the greatest treasures of the city, however, are the 2,000 officially recognized architectural and artistic monuments, ranging in period from the Romanesque through the Gothic to the Baroque, Rococo, Classical, and Neoclassical. The interiors of the buildings, which often house major art collections, have been restored since 1945. The most notable Romanesque monument is probably the 10th-century Church of St. George, behind the north wall of Hradčany. To the west is its more massive successor, the basically Gothic St. Vitus' Cathedral, the twin spires of which dominate the city skyline. Other Gothic monuments include the Týn Church on Staroměstské Square; the elegant Powder Tower, marking the former city walls in what is now the busy Příkopy shopping area; the restored Bethlehem Chapel, where Jan Hus preached in the 15th century; and the St. Agnes Convent, built in 1234 and notable for its collection of 14th-century paintings. The Old-New Synagogue and the tumbling, crowded gravestones of the Old Jewish Cemetery—Europe's oldest—betoken the strong Jewish tradition in Prague life.

Baroque buildings are the city's greatest single artistic treasure, among them the splendid Valdštejn and Clam-Gallas palaces, St. Nicholas Church, and the Antonín Dvořák Museum. The geometric tiling of the Golz-Kinský Palace facade provides a distinctive glimpse of the Rococo style. Classical buildings include the Bedřich Smetana Museum on the riverside and the elegant Belvedere Palace (the former Royal Summer Palace). The National Museum and the National Theatre are the main Neoclassical buildings.

The beauty of the city is enhanced by its many parks and gardens, including a major cultural, entertainment, and sports centre in the park named for Julius Fučík (a resistance leader of World War II) and a large zoo in suburban Troja. Recreational facilities also include the vast Strahov sports complex—containing three stadiums, the largest of which, Spartakiáda Stadium, holds 250,000 spectators and is used for the mass gymnastic display known as the Spartakiáda—as well as numerous other sports and cultural centres, with emphasis on facilities for youth. The film studios at Barrandov, on the city outskirts, have produced a number of high-quality motion pictures, and there is a museum of modern sculpture at Zbraslav.

History

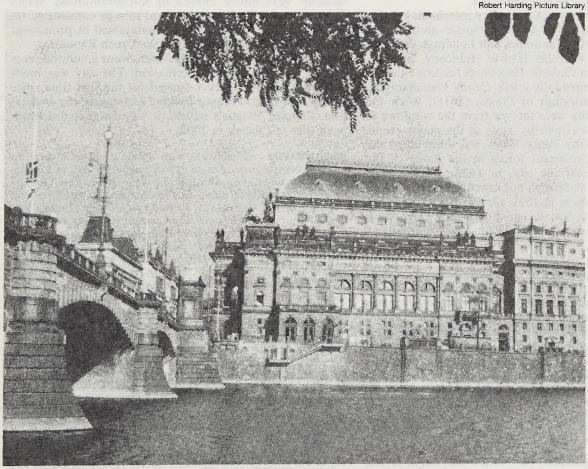
THE EARLY PERIOD

The foundation of the city. For thousands of years that portion of the Vltava's course where Prague was to rise was crossed by trade routes linking northern and southern Europe. The region is replete with Paleolithic relics, and Neolithic farmers inhabited the region from around 5000 to 2700 BC. Celts had settlements in the region from about 500 to 200 BC, including the fortified Závist, to the south of Prague. From the 4th to the 6th century AD, Slavs appeared on the Vltava banks, followed by the Avars.

The first settlement at what is now Prague has been traced to the second half of the 9th century. The oldest building was Vyšehrad (*hrad,* "castle"), set on a commanding right-bank hill. It was followed by what was to become Hradčany, set on an equally commanding left-bank site a little downstream. Legend (stirringly told in Smetana's opera *Libuše*) ascribes the foundation of Prague to a Princess Libuše and her husband, Přemysl, founder of the Přemyslid dynasty; legend notwithstanding, the Přemyslids, in power from about 800 to 1306, consolidated a political base centred on Prague that was to be the nucleus of the Bohemian state and that enabled the natural trade advantages of the city site to develop under defensive protection. The dynasty included St. Wenceslas (Václav), who was murdered by his brother Boleslav in about 939 and whose statue now looks down upon the square to which his name has been given; and Boleslav I, whose reign (*c.* 936–967) witnessed the consolidation of power against a German threat. The little community flourished, and in 965 the Jewish merchant and traveler Ibráhim ibn Ya'qūb was able to describe it as a "busy trading centre." In 973 the bishopric of Prague was founded.

Medieval growth. The economic expansion of the community was reflected in the topography of the city. A market centre on the right bank, opposite Hradčany, developed into the Old Town (Staré město), particularly after the construction of the first stone bridge, the Judith Bridge, over the river in 1170. By 1230 the Old Town had been given borough status and was defended by a system of walls and fortifications. On the opposite bank, under the

Robert Harding Picture Library



The National Theatre off the Vltava River and the 1st of May Bridge.

The role of Charles IV in planning the city

walls of Hradčany, the community known as Malá Strana (literally, "Small Side") was founded in 1257. Following the eclipse of the Přemyslids, the house of Luxembourg came to power when John of Luxembourg, son of the future emperor Henry VII, became king of Bohemia. His son, Charles IV, Bohemian king and Holy Roman emperor, had his capital at Prague from 1346 to 1378 and took considerable personal interest in the development of the city. In 1348 he founded Charles University, the first in central Europe, which was later to attract scholars and students from throughout the Continent. His reign also saw the growth of the planned New Town (Nově město) adjacent to the Old Town; construction of the Charles Bridge (1357, reconstructed in 1970) linking the Old Town and the Malá Strana; and the beginning (1344) of the great St. Vitus' Cathedral, which was not completed until 1929. Other buildings included the Carolinum (the central hall of the university), the town hall (destroyed in 1945), and several churches and monasteries in the New Town. The Jewish ghetto was also developed, and the bishopric was raised to an archbishopric in 1344.

By the 14th century Prague had become a major central European city, with the Czech money minted at nearby Kutná Hora serving as the hard currency of the entire region. Foreign merchants, notably Germans and Italians, became economically and politically powerful in uneasy alliance with the kings. The social order, however, became less stable because of the emergent guilds of craftsmen, themselves often torn by internal conflicts. The town paupers added a further volatile element.

The Reformation and the Thirty Years' War. Prague played a significant role in the Reformation. The sermons of Jan Hus, a scholar at the university, begun in 1402 at the now-restored Bethlehem Chapel and carrying forward the criticisms of the church developed by the English reformer John Wycliffe, endeared him to the common people but brought him into conflict with Rome; he was burned at the stake in the town of Constance (Konstanz, Ger.) in 1415. Popular uprisings in 1419, led by the Prague priest Jan Želivský, included the throwing of city councillors from the windows of the New Town Hall in the incident known as the first Defenestration of Prague. The next year Hussite peasant rebels, led by the great military leader Jan Žižka, joined forces with the Hussites of Prague to win a decisive victory over the Roman Catholic king (later emperor) Sigismund at nearby Vítkov Hill.

During the next 200 years, the wealthy merchants became ascendant once more, and the late Gothic architectural style flourished in many churches and buildings, reaching a peak in the fine Vladislav Hall of Hradčany. In 1526, however, the Roman Catholic Habsburgs became rulers of Bohemia and attempted to crush Czech Protestantism. The second Defenestration of Prague (1618), when the governors of Bohemia were thrown from the windows of the council room in Hradčany—one of the major events precipitating the Thirty Years' War—was followed by the decisive defeat of Protestant forces at the Battle of the White Mountain, near the city, in 1620. Twenty-seven Prague commoners and Czech noblemen were executed on the Staroměstské Square in 1621; the city ceased to be the capital of the empire, was occupied by Saxons (1631) and Swedes (1648), and went into a decline hastened by two outbreaks of plague.

EVOLUTION OF THE MODERN CITY

The return of more settled conditions in central Europe was marked by renewed economic growth, and Prague's population grew from 40,000 in 1705 to more than 80,000 by 1771. In 1784 the Old Town, the New Town, the Malá Strana, and the Hradčany complex were administratively united into one city. The merchants and the mostly German, Spanish, and Italian nobility who were active in and around Prague in this period had an enormous effect on both architecture and cultural life. Outstanding architects created magnificent palaces and gardens, and churches in the Prague version of the Baroque style sprang up throughout the city.

The onset of the Industrial Revolution had major effects in Prague. The first suburb (Karlín) was established in 1817, and in the next 20 years many factories sprang up, often in association with the coal mines and ironworks at Kladno and Králův Dvůr, not far away. The population exceeded 100,000 by 1837, and expansion continued after the city received its first railway eight years later. The rise of a working class and of strong nationalistic sentiments had a profound effect on the city; students, artisans, and workers took to the barricades against the ruling Austrians when revolution flared briefly in 1848. Within 20 years Czechs had won a majority on the City Council, and Czech cultural life was experiencing a renascence centred on Prague. The Neoclassical building of the National Museum and the National Theatre are only two examples of the building that took place in this period. By the 1890s the first electric streetcars (trams) were running in the city, urban services were being reorganized, and a replica of the Eiffel Tower overlooked the city from Petřín Hill.

Effects of the Industrial Revolution

In 1918 Prague became the capital of the newly independent Czechoslovak republic. By 1930 the population had reached 850,000. The city suffered a setback following the surrender of large parts of Bohemia and Moravia to Germany under the Munich Agreement of 1938. The citizens rose in revolt on May 5, 1945, and held the city until the Red Army arrived four days later. After World War II economic reconstruction began, careful planning was necessary to restore and preserve the historic monuments of the city centre. From the 1970s there was an increasing emphasis on the development of new satellite communities. The city continued to grow, although most of its population growth was attributable to annexation. The so-called Prague Spring of 1968, a short-lived excursion into liberalized social and governmental controls attempted by the government of Alexander Dubček, was terminated by Soviet military action in August of that year.

The democratic revolution

In November 1989, Prague became the cradle of the movement that swiftly ended four decades of communist rule in Czechoslovakia. An officially sanctioned march in the city, commemorating the death of a student at the hands of Nazis in 1939, resulted in police violence and public disorder. Indignation at the current regime kindled further unrest, and in the second half of November students, young intellectuals, and later older people, totaling some half a million, demonstrated in the streets of the capital. Subsequent pressure led to the resignation of the entire Communist Party leadership and the formation of a coalition government headed by noncommunists. When Czechoslovakia itself was dissolved into its constituent republics on Jan. 1, 1993, Prague maintained its prominent international status as capital of the Czech Republic.

Throughout the 1990s Prague underwent a cultural, economic, and political transformation. The city was modernized, the stock market opened for the first time since World War II, and Prague became a major tourist destination. Capping Prague's rebirth, it was designated a European City of Culture in 2000.

BIBLIOGRAPHY. Guidebooks with information on history include ALOIS SVOBODA, *Prague* (1965, reissued 1968), in English; and FRANTIŠEK KAFKA, *Baedeker's Prague*, trans. from German (1987). Photographic views of the city are presented in JIŘÍ DOLEŽAL and EVŽEN VESELÝ, *Památky staré Prahy* (1966), a survey of historical Prague, with brief commentary; EUGEN VASILIAK, *Nad Prahou: Prague Seen from Above* (1966), a book of aerial photographs with text in six languages; MIROSLAV KERECKÝ, *Prague in Colour* (1976; originally published in Czech, 1975); and RADOMIRA SEDLAKOVA, *Prague: An Architectural Guide*, trans. from Czech (1997). A study of the city from the perspective of urban sociology is provided by F.W. CARTER, "Prague and Sofia: An Analysis of Their Changing Internal City Structure," ch. 15 in R.A. FRENCH and F.E. IAN HAMILTON (eds.), *The Socialist City: Spatial Structure and Urban Policy* (1979), pp. 425–459. See also JIŘÍ HRŮZA and BLAHOMÍR BOROVIČKA, *Prague: A Socialist City* (1985). F.W. CARTER, "Kafka's Prague," ch. 2 in J.P. STERN (ed.), *The World of Franz Kafka* (1980), pp. 30–43, is an essay on Prague in Kafka's lifetime. Works describing the history of the city include COUNT (FRANZ) LÜTZOW, *The Story of Prague* (1902, reprinted 1971); JOSEF JANÁČEK (ed.), *Dějiny Prahy* (1964), covering the development and history of the city from the earliest times to 1960; JOSEPH WECHSBERG, *Prague: The Mystical City* (1971), essays on Prague's history; and JOSEF JANÁČEK, *Malé dějiny Prahy*, 3rd rev. ed. (1983), on history and architecture. (Ja.K./R.H.O./F.W.C.)

Property of
San Mateo
Public
Library

SAN MATEO PUBLIC LIBRARY
3 9047 05503367 6